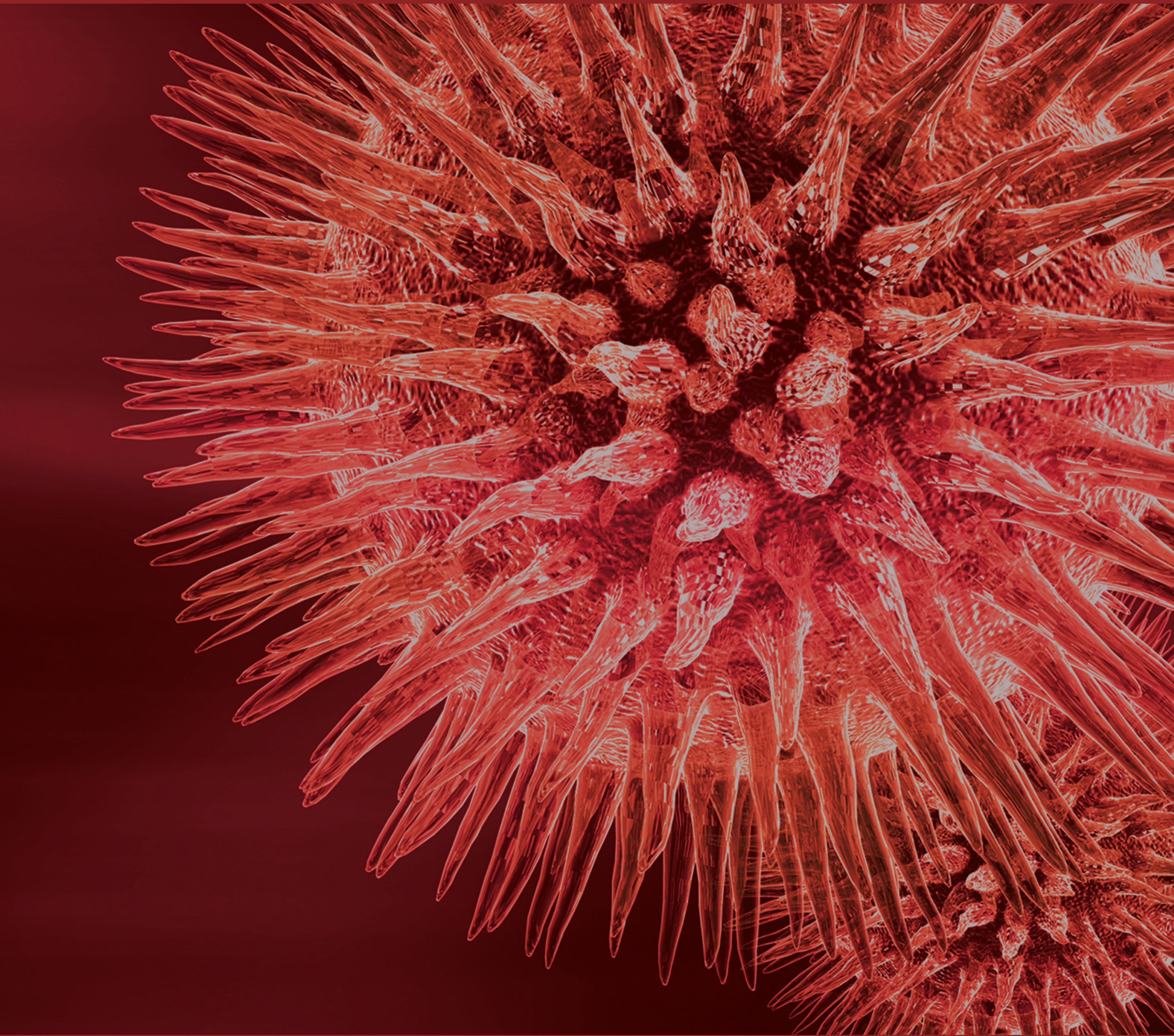
BioMed Research International

Current Advances in Molecular Phylogenetics

Guest Editors: Vassily Lyubetsky, William H. Piel, and Dietmar Quandt



Current Advances in Molecular Phylogenetics

BioMed Research International

Current Advances in Molecular Phylogenetics

Guest Editors: Vassily Lyubetsky, William H. Piel,
and Dietmar Quandt

Contents

Current Advances in Molecular Phylogenetics, Vassily Lyubetsky, William H. Piel, and Dietmar Quandt
Volume 2014, Article ID 596746, 2 pages

Phylogenetic Analysis of Entomoparasitic Nematodes, Potential Control Agents of Flea Populations in Natural Foci of Plague, E. I. Koshel, V. V. Aleshin, G. A. Eroshenko, and V. V. Kutyrev
Volume 2014, Article ID 135218, 26 pages

Reconciliation of Gene and Species Trees, L. Y. Rusin, E. V. Lyubetskaya, K. Y. Gorbunov, and V. A. Lyubetsky
Volume 2014, Article ID 642089, 22 pages

Differences in Brain Transcriptomes of Closely Related Baikal Coregonid Species, Oksana S. Bychenko, Lyubov V. Sukhanova, Tatyana L. Azhikina, Timofey A. Skvortsov, Tuyana V. Belomestnykh, and Eugene D. Sverdlov
Volume 2014, Article ID 857329, 11 pages

Changes in Bacterial Population of Gastrointestinal Tract of Weaned Pigs Fed with Different Additives, Mercè Roca, Miquel Nofrarías,Natàlia Majó, Ana María Pérez de Rozas, Joaquim SegalSegalés, Marisol Castillo, Susana María Martín-Orúe, Anna Espinal, Joan Pujols, and Ignacio Badiola
Volume 2014, Article ID 269402, 13 pages

The Variety of Vertebrate Mechanisms of Sex Determination, Antonina V. Trukhina, Natalia A. Lukina, Natalia D. Wackerow-Kouzova, and Alexander F. Smirnov
Volume 2013, Article ID 587460, 8 pages

Algorithms of Ancestral Gene Length Reconstruction, Alexander Bolshoy and Valery M. Kirzhner
Volume 2013, Article ID 472163, 10 pages

The Continuing Debate on Deep Molluscan Phylogeny: Evidence for Serialia (Mollusca, Monoplacophora + Polyplacophora), I. Stöger, J. D. Sigwart, Y. Kano, T. Knebelsberger, B. A. Marshall, E. Schwabe, and M. Schrödl
Volume 2013, Article ID 407072, 18 pages

Speciation in *Thaparocleidus* (Monogenea: Dactylogyridae) Parasitizing Asian Pangasiid Catfishes, Andrea Šimková, Celine Serbielle, Antoine Pariselle, Maarten P. M. Vanhove, and Serge Morand
Volume 2013, Article ID 353956, 14 pages

The Evolutionary Pattern and the Regulation of Stearoyl-CoA Desaturase Genes, Xiaoyun Wu, Xiaoju Zou, Qing Chang, Yuru Zhang, Yunhai Li, Linqiang Zhang, Jingfei Huang, and Bin Liang
Volume 2013, Article ID 856521, 12 pages

Phylogenetic Relationships of *Pseudorasbora*, *Pseudopungtungia*, and *Pungtungia* (Teleostei; Cypriniformes; Gobioninae) Inferred from Multiple Nuclear Gene Sequences, Keun-Yong Kim, Myeong-Hun Ko, Huanzhang Liu, Qiongying Tang, Xianglin Chen, Jun-Ichi Miyazaki, and In-Chul Bang
Volume 2013, Article ID 347242, 6 pages

Evolutionary Relations of Hexanchiformes Deep-Sea Sharks Elucidated by Whole Mitochondrial Genome Sequences, Keiko Tanaka, Takashi Shiina, Taketeru Tomita, Shingo Suzuki, Kazuyoshi Hosomichi, Kazumi Sano, Hiroyuki Doi, Azumi Kono, Tomoyoshi Komiyama, Hidetoshi Inoko, Jerzy K. Kulski, and Sho Tanaka
Volume 2013, Article ID 147064, 11 pages

Dynamic of Mutational Events in Variable Number Tandem Repeats of *Escherichia coli* O157:H7, A. V. Bustamante, A. M. Sanso, D. O. Segura, A. E. Parma, and P. M. A. Lucchesi
Volume 2013, Article ID 390354, 9 pages

Transcription Regulation of Plastid Genes Involved in Sulfate Transport in Viridiplantae, Vassily A. Lyubetsky, Alexander V. Seliverstov, and Oleg A. Zverkov
Volume 2013, Article ID 413450, 6 pages

Inferring Phylogenetic Networks from Gene Order Data, Alexey Anatolievich Morozov, Yuri Pavlovich Galachyants, and Yelena Valentinovna Likhoshway
Volume 2013, Article ID 503193, 7 pages

Comparative Analysis of Context-Dependent Mutagenesis Using Human and Mouse Models, Sofya A. Medvedeva, Alexander Y. Panchin, Andrey V. Alexeevski, Sergey A. Spirin, and Yuri V. Panchin
Volume 2013, Article ID 989410, 5 pages

miR-1279, miR-548j, miR-548m, and miR-548d-5p Binding Sites in CDSs of Paralogous and Orthologous *PTPN12*, *MSH6*, and *ZEB1* Genes, Anatoliy T. Ivashchenko, Assel S. Issabekova, and Olga A. Berillo
Volume 2013, Article ID 902467, 10 pages

miR156- and miR171-Binding Sites in the Protein-Coding Sequences of Several Plant Genes, Assyl Bari, Saltanat Orazova, and Anatoliy Ivashchenko
Volume 2013, Article ID 307145, 7 pages

Evaluating Phylogenetic Informativeness as a Predictor of Phylogenetic Signal for Metazoan, Fungal, and Mammalian Phylogenomic Data Sets, Francesc López-Giráldez, Andrew H. Moeller, and Jeffrey P. Townsend
Volume 2013, Article ID 621604, 14 pages

Origin and Status of Homologous Proteins of Biomineralization (Biosilicification) in the Taxonomy of Phylogenetic Domains, Igor E. Pamirsky and Kirill S. Golokhvast
Volume 2013, Article ID 397278, 7 pages

Molecular Identification and Ultrastructural and Phylogenetic Studies of Cyanobacteria from Association with the White Sea Hydroid *Dynamena pumila* (L., 1758), O. A. Koksharova, T. R. Kravzova, I. V. Lazebnaya, O. A. Gorelova, O. I. Baulina, O. E. Lazebny, T. A. Fedorenko, and E. S. Lobakova
Volume 2013 , Article ID 760681, 11 pages

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 596746, 2 pages
http://dx.doi.org/10.1155/2014/596746

Editorial

Current Advances in Molecular Phylogenetics

Vassily Lyubetsky,[1] William H. Piel,[2] and Dietmar Quandt[3]

[1] *Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow 127994, Russia*
[2] *Yale-NUS College & National University of Singapore, Singapore*
[3] *Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Correspondence should be addressed to Vassily Lyubetsky; lyubetsk@iitp.ru

Received 22 December 2013; Accepted 22 December 2013; Published 7 April 2014

Since its inception some 50 years ago, phylogenetics has permeated nearly every branch of biology. Initially developed to classify objects based on a set of cladistic rules, it has now become the central paradigm of evolutionary biology and a key framework for making sense of a wide range of disciplines [1], such as genomics [2], community ecology [3], epidemiology [4], conservation biology [5], and population dynamics [6], to name just a few. It is a testament to the power of phylogenetic methods that its application has expanded far beyond its original inception, now including the study of human culture, such as language and cultural memes [7].

Phylogenetic principles are used to reconstruct complex ancestral traits of morphological characters, genome structures and their properties, and evolutionary events (like gene duplications, losses, transfers, or chromosomal rearrangements). Phylogeny is also essential to infer gene and protein families, uncover complex population histories in epidemiological and other studies, and understand viral and cell genealogies in medicine and developmental biology. New concepts are developing that tackle various aspects of coevolution, including approaches to defining and algorithmically constructing complex evolutionary scenarios for genetic systems, their regulations, epigenetic and intrinsic factors, noncoding genome elements, sequence primary and secondary structures, the speciation process, and so forth.

The growth of phylogenetics is not just in breadth of disciplines, but also in the sheer volume of published phylogenetic results. Some twenty years ago, near-exponential growth in phylogenetic publications had already been noticed [8, 9], a growth that was probably attributable to the advent of powerful computers, PCR, and Sanger sequencing. An update on the assessment of phylogenetic growth (Figure 1) shows that not only is the growth in phylogenetic papers exponential, but more importantly the growth in the percentage of papers that report phylogenetic results is also exponential, indicating its increasing share in scientific research. Journals and databases have worked hard to keep pace with this growth, with the development of data repositories to archive and share data (e.g., TreeBASE, http://treebase.org/ and Dryad, http://datadryad.org/) that would otherwise be inefficient to distribute as supplementary addenda.

In the last ten years, the rate of growth of phylogenetic publications has waned somewhat (Figure 1), but with the recent advent of next-generation sequencing (NGS) we anticipate a new flood of phylogenetic results that is commensurate with this explosion of NGS data In addition to the phylogenetic results themselves, we also anticipate the need for new methodological papers to improve efficiencies in sequence assembly, multiple alignment, genome annotation, and pipelining of massive analyses.

Computational power is at risk of being outstripped because the volume of NGS data more than doubles each year, outpacing Moore's Law [10]. The limits of computational power portend the need for novel analytical approaches [11], among them "exact models" that avoid heuristics by finding mathematically provable global optima for a function, yet requiring low polynomial complexity, developing effective supertree and divide-and-conquer methods. Other perspective directions include modeling of coevolution as a system of stochastic processes, low-polynomial methods of simultaneous phylogeny and alignment construction, and applying mathematically proved methods to simulate test

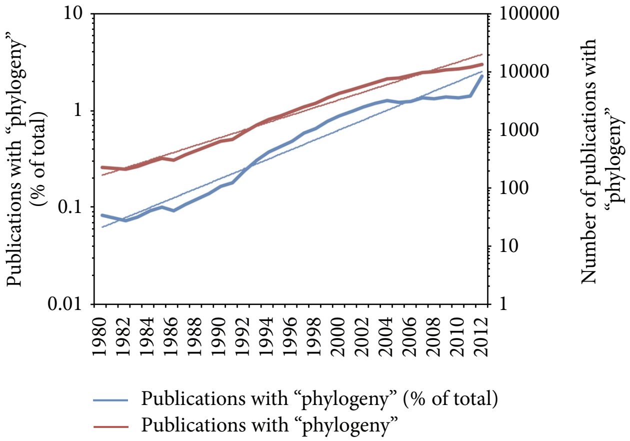


Figure 1: Growth of phylogenetic publications 1980–2012. Both the number of publications that involve "phylogeny" or "phylogenetic" terms and the proportion of publications appear to grow in a way that approximates exponential growth. Data were compiled from PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

datasets for benchmarking phylogenetic algorithms. These anticipated advances also need new publishing avenues for dissemination to the scientific community.

This special open-access issue, which we hope to be an annual occurrence with *Biomed Research International*, seeks to meet the anticipated demand for disseminating phylogenetic results and phylogenetic methods. The special issue covers a variety of topics in modern phylogenetics and its applications, from phylogenetic systematics to new methodological developments and reviews. Many authors of the special issue also contributed to the Moscow Conference on Molecular Phylogenetics (http://www.en.molphy.ru/), which is organized biannually by Moscow State University and the Institute for Information Transmission Problems of the Russian Academy of Sciences. The call for papers for the next issue "Molecular Phylogenetics 2014" is now open, and we believe that this series will serve as a platform to exchange ideas and publish research in this actively expanding interdisciplinary field.

Acknowledgments

We gratefully thank all scientific reviewers who dedicated their time to evaluate submissions in this special issue.

Vassily Lyubetsky
William H. Piel
Dietmar Quandt

References

[1] W. M. Fitch, "Uses for evolutionary trees," *Philosophical Transactions of The Royal Society of London B*, vol. 349, no. 1327, pp. 93–102, 1995.

[2] H. Ellegren, "Comparative genomics and the study of evolution by natural selection," *Molecular Ecology*, vol. 17, no. 21, pp. 4586–4596, 2008.

[3] C. O. Webb, D. D. Ackerly, M. A. McPeek, and M. J. Donoghue, "Phylogenies and community ecology," *Annual Review of Ecology and Systematics*, vol. 33, no. 1, pp. 475–505, 2002.

[4] S. C. Stearns, "Evolutionary medicine: its scope, interest and potential," *Proceedings of the Royal Society B*, vol. 279, pp. 4305–4321, 2012.

[5] K. A. Crandall, O. R. R. Bininda-Emonds, G. M. Mace, and R. K. Wayne, "Considering evolutionary processes in conservation biology," *Trends in Ecology and Evolution*, vol. 15, no. 7, pp. 290–295, 2000.

[6] P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee, Eds., *New Uses for New Phylogenies*, Oxford University Press, Oxford, UK, 1996.

[7] R. D. Gray, S. J. Greenhill, and R. M. Ross, "The pleasures and perils of darwinizing culture (with phylogenies)," *Biological Theory*, vol. 2, no. 4, pp. 360–375, 2007.

[8] M. J. Sanderson, B. G. Baldwin, G. Bharathan et al., "The growth of phylogenetic information and the need for a phylogenetic data base," *Systematic Biology*, vol. 42, no. 4, pp. 562–568, 1993.

[9] M. Pagel, "Inferring evolutionary processes from phylogenies," *Zoologica Scripta*, vol. 26, no. 4, pp. 331–348, 1997.

[10] D. Sheehan, "Next-generation genome sequencing makes non-model organisms increasingly accessible for proteomic studies: some implications for ecotoxicology," *Journal of Proteomics and Bioinformatics*, vol. 6, no. 1, Article ID 10000e21, 2013.

[11] C. X. Chan and M. A. Ragan, "Next-generation phylogenomics," *Biology Direct*, vol. 8, article 3, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 135218, 26 pages
http://dx.doi.org/10.1155/2014/135218

Research Article

Phylogenetic Analysis of Entomoparasitic Nematodes, Potential Control Agents of Flea Populations in Natural Foci of Plague

E. I. Koshel,[1] V. V. Aleshin,[2,3,4] G. A. Eroshenko,[1] and V. V. Kutyrev[1]

[1] *Russian Research Anti-Plague Institute "Microbe", Saratov 410005, Russia*
[2] *Belozersky Institute of Physical-Chemical Biology, Lomonosov Moscow State University, Moscow 119991, Russia*
[3] *Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia*
[4] *National Research Institute of Physiology, Biochemistry, and Nutrition of Farm Animals, Russian Academy of Agricultural Sciences, Kaluga Region, Borovsk 249013, Russia*

Correspondence should be addressed to E. I. Koshel; opossum39@mail.ru

Received 28 February 2013; Revised 5 June 2013; Accepted 22 October 2013; Published 3 April 2014

Academic Editor: Vassily Lyubetsky

Entomoparasitic nematodes are natural control agents for many insect pests, including fleas that transmit *Yersinia pestis*, a causative agent of plague, in the natural foci of this extremely dangerous zoonosis. We examined the flea samples from the Volga-Ural natural focus of plague for their infestation with nematodes. Among the six flea species feeding on different rodent hosts (*Citellus pygmaeus*, *Microtus socialis*, and *Allactaga major*), the rate of infestation varied from 0 to 21%. The propagation rate of parasitic nematodes in the haemocoel of infected fleas was very high; in some cases, we observed up to 1,000 juveniles per flea specimen. Our study of morphology, life cycle, and rDNA sequences of these parasites revealed that they belong to three distinct species differing in the host specificity. On SSU and LSU rRNA phylogenies, these species representing three genera (*Rubzovinema*, *Psyllotylenchus*, and *Spilotylenchus*), constitute a monophyletic group close to Allantonema and Parasitylenchus, the type genera of the families Allantonematidae and Parasitylenchidae (Nematoda: Tylenchida). We discuss the SSU-ITS1-5.8S-LSU rDNA phylogeny of the Tylenchida with a special emphasis on the suborder Hexatylina.

1. Introduction

More than 150 species of fleas feeding on different mammalian hosts, primarily rodents, are vectors of the bacterium *Yersinia pestis*, a causative agent of plague [1, 2]. In natural foci of plague, the dynamics of flea populations are among the main factors controlling the incidence of epizootics that pose a threat to humans inhabiting the areas [3–5]. Entomoparasitic nematodes of the order Tylenchida are known to control populations of various insect hosts [6–9]. The rate of tylenchid infestation in fleas reaches 50–60% in some cases [10, 11], when the nematodes cause castration and early death of the flea hosts [9, 12, 13].

Despite high importance of the Tylenchida as a nematode order harboring entomoparasites and notorious crop pests, their reliable phylogeny is still a challenge. Tylenchid nematodes differ widely in life cycle, parasitic strategies, and the host range that spans plants, fungi, and invertebrates. Phylogenies obtained from SSU and partial LSU rDNA data often disagree with classifications based on morphology and life cycle [14–21]. Phylogenetic resolution inside the order is far from being clear, which in many respects results from the insufficiency of data available to adequately describe its diversity. As for tylenchid parasites of fleas, only 31 species are described to date [9, 22–31], with no molecular vouchering. Here we present a study of parasitic nematodes isolated from fleas sampled from different rodent hosts in a natural focus of plague.

2. Materials and Methods

2.1. Collection of Samples. Samples were collected in 2012 (spring and autumn) and 2013 (spring) in the Volga-Ural natural focus of plague (Figure 1). The sampled rodents included sousliks (*Citellus pygmaeus*), mouse-like rodents (*Microtus socialis* and *Apodemus uralensis*), and jerboas (*Allactaga*

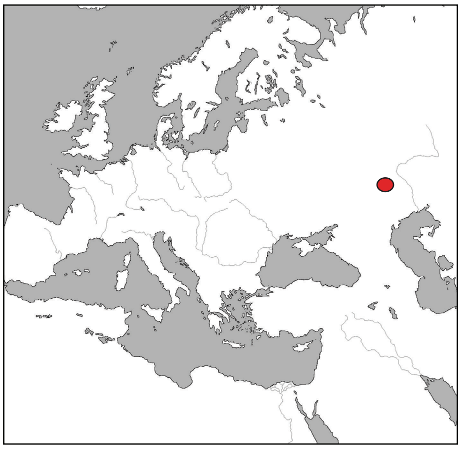


Figure 1: The sampling region on the map of Europe.

major). Three flea species (*Citellophilus tesquorum*, *Neopsylla setosa*, and *Frontopsylla semura*) were sampled on sousliks; two species (*Amphipsylla rossica* and *Ctenophthalmus secundus*) were on *M. socialis* voles; and one species (*Mesopsylla hebes*) was on jerboas. Fleas were examined for nematode infestation (Table 1). Examination and dissection of fleas were carried out using the dissecting microscope MBS-2 (LOMO, Russia). A half of parasitic nematodes sampled from each flea was preserved for subsequent DNA extraction, and another half was used for morphological analysis. Live fleas infected with nematodes were placed in glass flasks with river sand to obtain free-living forms. Insects were kept in a KBF 720 (E5.2) climate chamber (Binder, Germany) at 26°C and 80% humidity.

2.2. Morphological Analysis. Fixation and clarification of nematode preparations were performed using standard techniques described by De Grisse [32]. Material was mounted on slides in a drop of glycerin, bound by a paraffin circlet (http://pest.cabweb.org). Color staining of preparations was not performed. Morphometric analysis was conducted using the light microscope "Leica DM 1000" (Leica, Germany) with an eyepiece micrometer. Pictures of nematodes were taken with the microscope "DFC 425" (Leica, Germany). Published data on morphometrics [23, 25, 26] were used for comparison.

2.3. DNA Extraction, PCR, and Sequencing. DNA samples were extracted with a Diatom DNA Prep (IsoGen Lab, Russia). rDNA fragments were amplified using an Encyclo PCR kit (Evrogen, Russia) and primers given in Table 2. The amplified rDNA fragments were sequenced using an Applied Biosystems 3500xL DNA analyzer. Sequence reads were assembled with the CAP contig assembly program [33] and proofread with the BioEdit software [34]. For three isolates, almost complete sequences of 18S and 28S rRNA and complete sequences of 5.8 rRNA, internal transcribed spacers ITS1 and ITS2 were assembled. The sequences were submitted to GenBank under accession nos. KF155281–KF155283. For the rest of isolates, partial (750–800 bp) sequences of 18S and 28S rRNA genes were submitted to GenBank under accession nos. KF373731–KF373740.

2.4. Phylogenetic Analysis. The newly obtained rDNA sequences of tylenchid parasites of fleas were aligned with a selected set of other tylenchid sequences obtained from the GenBank. The main selection criterion was to sample representatives of all clades that occur in published SSU and LSU rDNA phylogenies of the Tylenchida [16–21, 39]. Apart from the D2-D3 LSU rDNA expansion segment commonly used in previous studies, we included all LSU rDNA sequence data available for the Tylenchida, with the exception of *Basiria* sp. SAN-2005 (accession nos. DQ145619, DQ145667) that in our preliminary analyses (data not shown) demonstrated a disputable affinity to the Tylenchida. For the species *Anguina tritici*, *Globodera pallida*, *Heterodera glycines*, *Pratylenchus vulnus*, and *Radopholus similes* the nearly complete rDNA sequences were assembled with appropriate cDNA fragments identified with BLAST [40]. Partial LSU rDNA sequence of *Ditylenchus dipsaci* was combined with the soil environmental clone NTS_28S_061A_2_b4 (accession no. KC558346), as the clone sequence appeared to represent a close tylenchid relative of *D. dipsaci.* Chimeric sequences were also created in some cases when closely related partial rDNA sequences were found in the database. All sequences and their accession numbers are listed in Table 3. Cephalobidae and Chambersiellidae were chosen as the outgroup. Alignments were constructed with the MUSCLE program [41] and refined manually using the MEGA 5.0 software package [42]. Three alignments were generated: (1) SSU rDNA, (2) D3 region of LSU rDNA, and (3) concatenated rDNA data including SSU, LSU, 5.8S rDNA, and highly conserved regions of ITS1. After discarding ambiguously aligned positions, the alignments length was 1,723, 592, and 4,930 positions, respectively. Bayesian reconstruction of phylogeny was done with the PhyloBayes software, version 3.2 [43] under the GTR + CAT + DP model [44]. Eight independent runs were performed with 4,000,000 cycles each; the first 3,000,000 cycles were discarded. A consensus tree with Bayesian posterior probabilities was constructed for the remained tree sample. Bayesian reconstruction was also performed using the MrBayes software [45] under the GTR + G8 + I model [46] in two independent runs, each with four Markov chains. The chains were run for 5,000,000 generations, with trees sampling every 1,000th generation. The consensus posterior probabilities were calculated after discarding the first 3,000,000 generations. Partitioning "by genes" was used for the concatenated alignment with all parameters unlinked, except for the topology and branch lengths. In addition, node support was estimated with maximum likelihood bootstrap as implemented in the RAxML software, version 7.2.6 [47], under the GTR + G + I model with 1,000 bootstrap replicates. Alternative topologies were tested using the approximately unbiased (AU) [48] and Kishino and Hasegawa [49] tests implemented in the CONSEL software [50] and the expected likelihood weight test [51] implemented in the TREE-PUZZLE software [52]. TREEVIEW

Table 1: Number of fleas studied and the percentage of fleas infected with nematodes.

Time of sampling	Host rodent species	Flea species	Number of collected fleas	Number of infected fleas	Percentage of infected fleas
April 2012	*Citellus pygmaeus*	*Citellophilus tesquorum*	41	7	17.1%
		Neopsylla setosa	73	5	6.8%
		Frontopsylla semura	54	7	13%
October 2012	*Microtus socialis*	*Amphipsylla rossica*	135	9	6.7%
		Ctenophthalmus secundus	88	1	1.1%
April 2013	*Citellus pygmaeus*	*Citellophilus tesquorum*	34	0	0
		Neopsylla setosa	271	22	8.1%
		Frontopsylla semura	19	4	21%
	Microtus socialis and *Apodemus uralensis*	*Amphipsylla rossica*	6	0	0
		Ctenophthalmus secundus	52	0	0
	Allactaga major	*Mesopsylla hebes*	34	2	5.9%

Table 2: Nucleotide sequences of primers used in this study.

Primer	Sequence	Orientation	References
Nik22	tmycygrttgatyctgyc	F	This study
A	gtatctggttgatcctgccagt	F	[35]
Q5nemCh	gccgcgaayggctcattayaac	F	This study
G18SU	gcttgtctcaaagattaagcc	F	[36]
Ves18-d9	gtcgtaacaaggtatccgtaggtgaac	F	This study
R18Tyl1	ggtccaagaatttcacctctc	R	[36]
B	gtaggtgaacctgcagaaggatca	R	[35]
Q39nem	gaaaccttgttacgacttttrcbygg	R	This study
58d1	rcatcgatgaagaacgywg	F	[37]
58r nem	gcwgcgttcttcatcgacyc	R	This study
28d3	gtcttgaaacacggaccaagg	F	[37]
28d6	ggtyagtcgrtcctrag	F	[37]
D2A	acaagtaccgtgagggaaagttg	F	[38]
28r4	gctatcctgagggaaacttcgg	R	[37]
28r2nem	cggtacttgttcgctatcg	R	This study
28r7	agccaatccttwtcccgaagttac	R	[37]
28r12	ttctgacttagaggcgttcag	R	[37]
D3B	tcggaaggaaccagctacta	R	[38]

[53] was used as the tree viewer and editor, and site-wise log-likelihoods were computed with TREE-PUZZLE under the GTR + G8 + I model with substitution matrix parameters estimated by MrBayes.

3. Results

3.1. Infestation of Fleas with Nematodes. The infestation rate is shown in Table 1 (in total, 807 flea specimens were studied). Among the six flea species studied, the population size and the percentage of infected fleas varied depending on the season. Three flea species sampled on sousliks (*Citellophilus tesquorum*, *Neopsylla setosa,* and *Frontopsylla semura*) exhibited a stable population density. In the two species, *N. setosa* and *F. semura*, the infestation rate was moderate to high in the spring seasons of 2012 and 2013. In *C. tesquorum*, no infected fleas were detected in spring 2013, whereas in spring 2012 the fleas were highly infested (17.1%). The vole flea *Amphipsylla rossica* was abundant and moderately infested in autumn, whereas being less abundant in spring, which may explain the absence of infected fleas in the spring sample. Another vole flea, *Ctenophthalmus secundus*, exhibited a consistently high population density and low infestation rate in both spring and autumn samples.

Adult parasitic females and their progeny were found in the haemocoel of infected fleas. In the infected fleas *C. tesquorum*, *A. rossica*, *C. secundus*, and *Mesopsylla hebes*, only one generation of parasitic females was observed. Their amount in a flea specimen is determined by the number of free-living infective females that penetrate into the flea larva. We observed 1 to 2 or 1 to 4 adult parasitic females per flea specimen in spring and autumn, respectively. An additional parthenogenetic generation of parasitic females was found in some fleas of *N. setosa* and *F. semura*, where

Table 3: List of OTUs and accession numbers of sequences.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Chambersiellidae∗						
Fescia grossa	KC242218	—	DQ145636 DQ145684	87.1/—	[54] [55]	Chambersiellidae
Geraldius sp. SAN-2010a	—	—	GU062821	17.8/—	[56]	Chambersiellidae
Cephalobidae						
Acrobeloides maximus	EU196016	JX026706	EU195987	94.8/—	[57] [58] [57]	Cephalobidae
Cephalobus cubaensis	AF202161	AF202161	EU253570	89.8/—	[59] [57]	
Panagrolobus sp. SN-2010	—	—	HM439771	51.9/—	[60]	
Cephalobidae Gen. sp. MHMH-2008	FJ040406	—	—		Holterman et al., 2008, unpublished.	
Zeldia punctata	—	DQ146426	EU195988	96.6/—	[61] [57]	
Zeldia sp.	AY284675	—	—			
Aphelenchidae						
Aphelenchus avenae	JQ348399	AF119048	—	96.9/—	[62] [63]	Aphelenchidae
Aphelenchus sp.	—	—	DQ145664 DQ145714		[55]	
Paraphelenchus acontioides	—	—	HQ218322	45.5/—	[64]	
Paraphelenchus sp.	AY284642	—	—		[18]	
Hexatylina + “Anguinata (part)”: Iotonchioidae						
Allantonema mirable	—	—	JX291132	10.6/85.8	[39]	Allantonematidae
Bradynema listronoti	DQ915805		DQ915804	45.6/96.8	[65]	
Bradynema rigidum			DQ328730	10.4/86.3	[20]	
Contortylenchus sp.	—	—	DQ328731	—/85.4	[20]	
Deladenus durus	JQ957898	—	—	34.0/—	[66]	Neotylenchidae
Deladenus proximus	JF304744	JF304744	—	35.2/—	[67]	
Deladenus siricidicola isolate 354	AY633447		AY633444	45.8/98.1	[68]	
Deladenus siricidicola isolate 466	FJ004890	FJ004890	—	41.7/—	[69]	
Deladenus siricidicola isolate 1093	FJ004889	FJ004889	—	42.0/—	[69]	
Fergusobia camaldulensae	AY589294	—	AY589346	45.7/98.0	[68]	
Fergusobia sp. 444	EF011667	—	EF011675	45.7/97.3	[68]	
Fergusobia sp. SBG	FJ393270	—	FJ386996	45.7/98.3	[70]	
cf. *Gymnotylenchus* sp. TSH-2005	AY912040	—	—	12.9/—	Powers et al., unpublished.	

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Howardula aoronymphium	AY589304	AY589304	AY589395	49.7/96.1	[68]	Allantonematidae
Howardula dominicki	AF519234	AF519234	—	37.4/—	[71]	
Howardula neocosmis	AF519226	AF519226	—	38.2/—	[71]	
Howardula phyllotretae	JX291137	—	DQ328728	41.9/86.1	[39] [20]	
Howardula sp. CD353	—	—	JX291131	—/93.9	[39]	
Howardula sp. SP-A	AF519232	AF519232	—	37.7/—	[71]	
Howardula sp. SP-F	AF519222	AF519222	—	38.2/—	[71]	
Howardula sp. SP-MA	AF519233	AF519233	—	38.1/—	[71]	
Howardula sp. SP-PS	AF519231	AF519231	—	38.1/—	[71]	
Parasitylenchus bifurcatus	KC875397	—		44.0/85.3	[72]	Parasitylenchidae
Parasitylenchus sp.	—	—	DQ328729		[20]	
Psyllotylenchus sp. ex *Frontopsylla semura*	KF373734	—	KF373739	27.1/93.7	This study	
Psyllotylenchus sp. ex *Neopsylla setosa*	KF373733	—	KF373738	27.1/93.7	This study	
Rubzovinema sp. ex *Amphipsylla rossica*	KF155281	KF155281	KF155281	90.0/100.0	This study	Neotylenchidae
Rubzovinema sp. ex *Ctenophthalmus cecundus*	KF155282	KF155282	KF155282	89.8/100.0	This study	
Rubzovinema sp. ex *Citellophilus tesquorum*	KF155283	KF155283	KF155283	93.2/100.0	This study	
Rubzovinema sp. ex *Frontopsylla semura*	KF373732	—	KF373737	27.1/93.7	This study	
Rubzovinema sp. ex *Neopsylla setosa*	KF373731	—	KF373736	27.1/93.7	This study	
Skarbilovinema laumondi	—	—	JX291136	10.9/91.0	[39]	Iotonchioidea
Skarbilovinema lyoni	JX291138	—	DQ328733	41.8/86.3	[39] [20]	
Spilotylenchus sp. ex *Mesopsylla hebes*	KF373735	—	KF373740	27.1/93.4	This study	Parasitylenchidae
cf. *Sychnotylenchus* sp. CSP1-09	DQ080531	—	—	12.9/—	Powers et al., unpublished.	Sychnotylenchidae
Wachekitylenchus bovieni	—	—	DQ328732	—/85.9	[20]	Parasitylenchidae
Unidentified Allantonematidae HaMW	JQ941710	—	—	18.5/—	Rhule, unpublished.	Allantonematidae
Unidentified Allantonematidae NK2011_2	AB663183	—	—	12.0/—	[73]	
Unidentified Allantonematidae NK2011_3	AB663184	—	—	12.0/—	[73]	
Unidentified nematode 804U-025	EU880149	—	—	12.0/—	[74]	

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Unidentified nematode CD289	—	—	JX291133	—/84.1	[39]	
Unidentified nematode RGD591T12	AB455970	—	—	12.0/—	[73]	
Unidentified nematode WY2009_BAR-1	—	—	FJ661075	—/96.3	[75]	
Unidentified parasite ex *Chrysobothris affinis*	—	—	DQ202658	—/51.0	Hunt et al., unpublished.	
Hexatylina + "Anguinata (part)": Sphaerularioidea						
Deladenus sp. PDL-2005	AJ966481	—	—	35.0/—	[16]	Neotylenchidae
cf. *Helionema* sp. MHMH-2008	EU669913	—	—	34.0/—	[19]	Parasitylenchidae (genera dubia in Hexatylina)
cf. *Hexatylus* sp. Westplace	AY912050	—	—	12.9/—	Powers et al., unpublished.	Neotylenchidae
Nothotylenchus acris	AY593914	—	—	34.0/—	[76]	Anguinidae
Sphaerularia bombi	AB250212	—	DQ328726	56.7/100.0	Takahashi, unpublished. [20]	Sphaerulariidae
Sphaerularia vespae	AB300595	AB300595	AB300596	54.7/100.0	[77]	
Unidentified nematode 801L-022	EU880129	—	—	12.1/—	[74]	
Anguinata						
Anguina tritici	AY593913	JF826515	HO058555 DQ328723	57.6/92.9	Holterman et al., unpublished. Rao and Rao, unpublished. Rao et al., unpublished. [20]	Anguinidae
Ditylenchus adasi	EU669909	—	—	34.6/—	[19]	
Ditylenchus angustus	AJ966483	—	—	34.6/—	[16]	
Ditylenchus destructor		JX162205		50.0/99.5	[78]	
Ditylenchus dipsaci	AY593911	AY593911	JF327759	60.9/100.0	[76] Zhao 2011, unpublished.	
clone NTS_28S_061A_2_b4			KC558346		[79]	
Ditylenchus drepanocercus	JQ429768	JQ429774	JQ429772	48.7/89.3	[80]	
Ditylenchus halictus	AY589297			52.8/97.3	[68]	
Ficotylus congestae	EU018049			45.6/97.5	[81]	
Halenchus fucicola	EU669912	—	—	34.6/—	[19]	
Pseudhalenchus minutus	AY284638			34.6/—	[19]	
Unidentified entomoparasitic nematode SAS-2006 "*Neotylenchus*" sp.	—	—	DQ328725	—/85.6	[20]	
"Tylenchina": Tylenchidae						
Aglenchus agricola	FJ969113	—	—	46.0/—	van Megen et al., unpublished.	Tylenchidae
Aglenchus sp.	—	—	JQ004996		[82]	

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Coslenchus costatus	AY284581	—	—	45.5/—	[18]	
Coslenchus sp.	—	—	JQ005007		[82]	
Filenchus annulatus	JQ814880	—	JQ005017	46.4/—	[82]	
Tylenchus davainei	AY284588	—	—	33.9/—	[18]	
"Tylenchina": Tylodoridae						
Eutylenchus excretorius	EU915487	EU915500	EU915490	35.8/—	[83]	Atylenchidae
Cephalenchus hexalineatus	AY284594	—	—	44.1/—	[18]	Tylodoridae
"Tylenchina": Boleodoridae						
Basiria gracilis	EU130839	—	DQ328717	44.6/—	[84] [20]	Tylenchidae
Basiria sp. 3 TJP-2012	—	—	JQ004998	12.0/—	[82]	
Boleodorus thylactus	AY993976	—	—	46.7/—	[16]	
Boleodorus sp.	—	—	JQ005001		[18]	
Neopsilenchus magnidens	AY284585	—	—	45.6/—	[18]	
Neopsilenchus sp. 3 TJP-2012	—	—	JQ005020		[82]	
Neopsilenchus sp. 1 TJP-2012	—	—	JQ005018	11.9/—	[82]	
"Hoplolaimina": Merliniidae						
Nagelus leptus	—	—	DQ328715	45.2/—	[20]	Telotylenchidae
Nagelus obscurus	EU306350	—	—		[17]	
Pratylenchoides ritteri	AJ966497	—	JX261964	48.7/—	[16] [85]	Pratylenchidae
Psilenchus cf. *hilarulus*	AY284593	—	EU915489	44.1/—	[18] [83]	Psilenchidae
Scutylenchus quadrifer	AY284599	—	—	41.5/—	[18]	Telotylenchidae
Scutylenchus sp.	—	JQ069956	—		[86]	
"Tylenchina": Ecphyadophoridae						
Ecphyadophora sp. JH-2004	AY593917	—	—	33.7/—	[76]	Ecphyadophoridae
"Ditylenchus" brevicauda	AY284635	—	—	33.9/—	[18]	Anguinidae
Malenchus andrassyi	AY284587	—	—	32.3/—	[18]	Tylenchidae
Ottolenchus discrepans	AY284590	—	—	33.7/—	[18]	
Criconematina						
Hemicriconemoides gaddi	—	KC520471	KC520470	55.6/—	[87]	Criconematidae
Hemicriconemoides pseudobrachyurus	AY284622	—	—		[18]	
Hemicycliophora lutosa	—	GQ406237	GQ406240	53.2/—	[88]	Hemicycliophoridae
Hemicycliophora thienemanni	AY284628	—	—		[18]	

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Meloidoderita kirjanovae	—	DQ768427	DQ768428	50.8/—	[89]	
Sphaeronema alni	FJ969127	—	—		van Megen, unpublished.	Sphaeronematidae
Meloidoderita sp.	GU253916	GU253917	JQ771954	50.8/—	[90] Cudejkova and Cermak, unpublished.	
Tylenchulus semipenetrans	AJ966511	FJ588909	FJ969710	57.5/—	[16] [91] [92]	Tylenchulidae
“Hoplolaimina”: Belonolaimidae						
Belonolaimus longicaudatus	AY633449	DQ672366	GQ896548	55.8/—	[68] [93] [94]	Belonolaimidae
Ibipora lolii	JQ771535	—	—	30.9/—	[95]	
“Hoplolaimina”: Hoplolaimidae						
Carphodorus sp.	JQ771538	—	JQ771550	41.3/—	[95]	

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Globodera pallida	EU855119	EU85511	BM415342 BM415248 CV577211 CV577977 CV579301E U85511	93.6/—	Nowaczyk et al., unpublished. Opperman, unpublished [96].	Heteroderidae
Heterodera glycines	AF216579 BI704127 BI748392 CA940548 CB379240 CB379263 CB379850 CB380242 CB825296 CB825409 CB825970 CB935610 CK348871 CK348904 CK349175 CK352112	AF216579	AF133304 AF216579 BI704144 BI704144 BI749520 CA940190 CA940212 CA940243 CA940406 CA940424 CA940429 CA940589 CB238697 CB279977 CB299455 CB373844 CB373981 CB379125 CB379140 CB379219 CB379312 CB379439 CB379505 CB379696 CB379707 CB379996 CB380091 CB380241 CB824788 CB824878 CB825995 CB934877 CB934931 CB934950 CB934954 CK348525 CO036619 HM560850 JN684906	98.3/—	[97] [96]. [98] Yan and Davis, unpublished. [99] Ye et al., unpublished. Wei et al., unpublished.	
Morulaimus sp.	JQ771540	—	—	31.5/—	[95]	Belonolaimidae

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Radopholus similis	AJ966502 AY912509 EF384224 EY190988 EY191076 EY191697 EY191883 EY192786 EY192788 EY193123 EY193253 EY194340 EY194464 EY194646 EY195472 FJ040398	AY912509 EF384224	EU555409 EY189839 EY190550 EY190620 EY190961 EY191066 EY191073 EY191135 EY191160 EY191173 EY191237 EY192021 EY192028 EY192080 EY192091 EY192247 EY192381 EY192472 EY192501 EY192526 EY192892 EY192907 EY193005 EY193037 EY193249 EY193314 EY193798 EY193897 EY193971 EY194395 EY194454 EY194530 EY195146 EY195204 EY195406	97.5/—	[16] [100] Long et al., unpublished. [101] Holterman et al., unpublished. [102] [100] Zhao unpublished. [86]	Pratylenchidae

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
			EY195408 EY195580 EY195889 EY195943 GQ281471 JN091962 JQ782249			
Rotylenchulus reniformis	JX406356	FJ374686	HM131884 FJ906072	59.4/—	[103] Rahman et al., unpublished. [104]	Rotylenchulidae
“Hoplolaimina”: Pratylenchidae						
Dolichodorus sp. WY-2006	DQ912918	—	—	33.9/—	[105]	Dolichodoridae
Hirschmanniella loofi	EU306353	EU620472	EU620469	51.6/—	[17] [106]	Pratylenchidae
Macrotrophurus arbusticola	AY284595	—		33.9/—	[18]	Telotylenchidae
Meloidogyne arenaria	U42342	U42342	U42342 AF023855 AF023856	99.2/—	Georgi and Abbott, unpublished.	Meloidogynidae
Meloidogyne artiellia	AF248477	AF248477	AF248477	99.2/—	[107]	

Table 3: Continued.

Name	18S rRNA	ITS1-5.8S rRNA	28S rRNA	%, SSU-ITS1-5.8S-LSU/D3	Reference	Family by [8]
Nacobbus aberrans	AJ966494	DQ017473	U47557	49.0/—	[16] [108] [109]	Pratylenchidae
Pratylenchus vulnus	EU669955	JQ966892	BQ580554 CV198923 CV198995 CV199233 CV199349 CV199490 CV200136 CV200423 CV200464 CV200467 CV200471 CV200530 CV200687 CV200896 CV201004 CV201135 EL887566 EL887705 EL888035 EL888060 EL888174 EL888269 EL888739 EL888778 EL889241 EL889472 EL889797 EL889934 EL889934 EL889977 EL889994 EL890380 EL890701 JQ003993 JQ003994 JX047008	100.0/—	[19] [110] [96] [96] [111] Zhao, unpublished. [112]	
Tylenchorhynchus dubius	EU306352	—	DQ328707	53.2/—	[17] [20] [113]	Telotylenchidae
Tylenchorhynchus zeae	—	EF519711	—			

*Clades of the tree, marked by boldface.

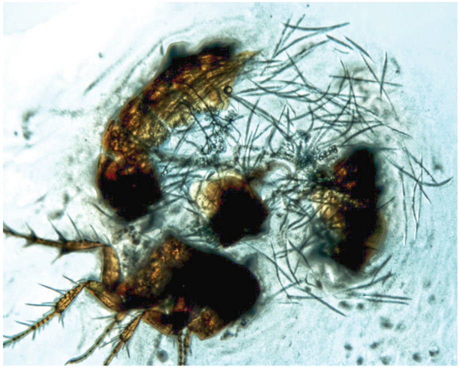


Figure 2: Numerous juveniles of *Rubzovinema* sp. extracted from the dissected body of a *Citellophilus tesquorum* flea.

up to 16 specimens per flea were observed. As in other entomoparasitic nematodes, the propagation rate depends on the host age. Thus, in young fleas up to 10 juveniles was found per flea specimen, whereas up to 1,000 juveniles of different stages were contained in some old fleas (Figure 2). After the 2nd molt the number of juveniles is maximal, and 3rd stage juveniles massively migrate to the rectal section of the flea intestine for exit to the environment. In some cases, the observed infestation level was so high that nematodes penetrated distal segments of the flea legs, from where they have no way to the environment.

3.2. Morphological Analysis of Entomoparasitic Stages in Nematode Isolates and Their Taxonomic Identification. Analysis of morphology of entomoparasitic stages suggests that the studied nematode isolates from three distinct groups. A single generation of parasitic females was observed in the first two groups and an additional parthenogenetic generation—in the third group. According to morphometric data on adult parasitic females (Tables 4–6), the first two groups belong to the genera *Rubzovinema* or *Spilotylenchus* and the third group to the genus *Psyllotylenchus*. Photographs of parasitic females of *Rubzovinema* sp., *Spilotylenchus* sp., and *Psyllotylenchus* sp. are depicted in Figure 3. Figure 4 shows their distribution among flea samples studied.

According to morphometric evidence, parasitic females and juveniles of the genera *Rubzovinema* and *Spilotylenchus* are very similar. However, in the first two groups of isolates we found characters bearing discriminative and identificational value. In particular, the oesophageal glands in juveniles III of the first group are poorly developed. This is a distinctive feature of the genus *Rubzovinema*, where males and females have shortened oesophageal glands located close to the nerve ring. In the second group of isolates, oesophageal glands are well developed and elongated, which is characteristic of the genus *Spilotylenchus*. In the first group, the stylet possesses a heavily sclerotized distal spear with a length of approximately half the total stylet length and has a stem with a weaker sclerotization and widening to the base. This stylet structure is characteristic of the genus *Rubzovinema*, and stylet length (18.5 (14–22) μm) is in accordance with morphometrics given in the description of this genus [26]. In the genus *Spilotylenchus,* the stylet varies in shape but always possesses a shortened conical distal spear. In the second group of isolates, the stylet structure was similar to that of *Spilotylenchus*. Also, the vulval lips of the first group are more protruded than in *Spilotylenchus*. Other features, including the morphometrics, vary widely in both genera, which hampers taxonomic identification. Nevertheless, based on distinctive traits, we identified the first and second group of isolates as *Rubzovinema* sp. and *Spilotylenchus* sp., respectively.

In the genus *Rubzovinema*, the single species described to date is *Rubzovinema ceratophylla* [26]. This species is known to parasitize exclusively the flea *Citellophilus tesquorum* that feeds on sousliks. The specimens of *Rubzovinema* studied in this work were isolated from five flea species, *C. tesquorum*, *Neopsylla setosa*, *Frontopsylla semura*, *Amphipsylla rossica*, and *Ctenophthalmus secundus*, of which the latter two were sampled on mouse-like rodents. Also, the parasitic females of *Rubzovinema* sp. differed from *R. ceratophylla* by morphology; they have a shorter tail and more protruded vulval lips. A morphometric comparison of *Rubzovinema* sp. and *R. ceratophylla* is given in Table 4.

The parasitic females of *Spilotylenchus* sp. were isolated from the flea *Mesopsylla hebes* associated with jerboas. The females were not identified to the species level because of a small number of available specimens and the lack of a free-living stage. A morphometric comparison of *Spilotylenchus* sp. and the morphologically closest species *Spilotylenchus maisonabei* [23] is given in Table 5.

In the genus *Psyllotylenchus*, descriptions of most species are fragmentary and incomplete, which precluded the species identification of the *Psyllotylenchus* isolates from the fleas *N. setosa* and *F. semura* feeding on sousliks. A morphometric comparison of *Psyllotylenchus* sp. and the type species of this genus, *Psyllotylenchus viviparous* [25], is given in Table 6.

The 18S and 28S rDNA sequences of *Rubzovinema* sp. specimens from *A. rossica* and *C. secundus* were 100% identical, which indicates that the isolates belong to the same species. The sequences of *Rubzovinema* sp. ex *C. tesquorum*, *Rubzovinema* sp. ex *N. setosa*, and *Rubzovinema* sp. ex *F. semura* diverged from one another and from the gene sequences of *Rubzovinema* sp. ex *A. rossica* and *Rubzovinema* sp. ex *C. secundus* by 0.4–0.7%, which corresponds to the levels of intraspecific variation [14, 114–119]. The 18S and 28S rDNA sequences of *Psyllotylenchus* sp. ex *N. setosa* and *Psyllotylenchus* sp. ex *F. semura* were 100% identical, indicating that they belong to the same species. The 18S and 28S rDNA sequences of *Rubzovinema* sp. and *Psyllotylenchus* sp. diverge by 1.2% and 1.9%, respectively. Those of *Spilotylenchus* sp. ex *M. hebes* were found to be more divergent. The degree of divergence of the 18S rDNA sequence of *Spilotylenchus* sp. ex *M. hebes* from those of either *Rubzovinema* sp. or *Psyllotylenchus* sp. was 2.4%; the D3 expansion segment of 28S rDNA diverged by 13.1% and 12.0%, respectively. The observed divergence rate of rDNA sequences agrees well with published evidence on entomoparasitic nematodes [14, 114–118]. Thus, intraspecific divergence of 18S rDNA in *Deladenus siricidicola* is 1% [120], of D2 and D3 expansion segments in the phytoparasite *Bursaphelenchus xylophilus* is from 0% to 0.6%, and the interspecific variation between the

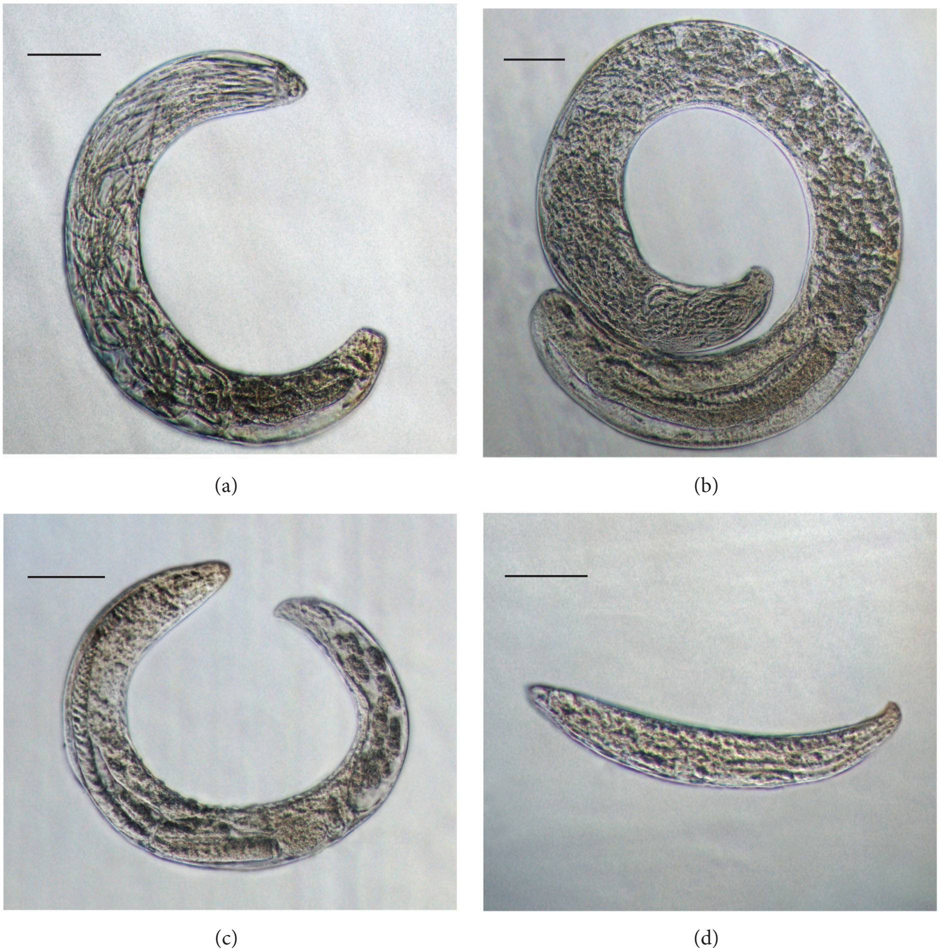


Figure 3: Parasitic females of the studied nematode species. (a) *Rubzovinema* sp., heterogeneous female; (b) *Spilotylenchus* sp., heterogeneous female; (c) *Psillotylenchus* sp., heterogeneous female of the first generation; (d) (c): *Psillotylenchus* sp., parthenogenetic female of the second generation. Scale bar—200 μm.

Table 4: Comparison of morphometrics in parasitic females of *Rubzovinema* sp. and *Rubzovinema ceratophylla*.

Character	*Rubzovinema* sp. (this study)	*Rubzovinema ceratophylla* [26]
N	29	27
L	1278,6 (840–1570)	1265,1 (810–1840)
D	120,8 (85–145)	137,3 (62–200)
A	11,19 (7,9–16,1)	9,51 (6,4–16,8)
C	65,4 (31,4–100)	44,10 (10–86,4)
V%	96,4 (93,1–97,9)	95,44 (92–98,9)
Total length of stylet (St)	18,5 (14–22)	19,5 (18–21)
Length of distal edge of stylet	7,2 (5–8,7)	—
Distance between anterior end and excretory pore (Ex)	20,7 (10–31)	—
Distance between anterior end and nerve ring	61,2 (50–74,5)	
Total length of tail (Cd)	21,9 (10–42)	26,35 (14–47,5)
Distance between vulva and tail end	46,1 (23–75)	—
Distance between vulva and anus (V–A)	26,9 (13–40)	—

All measurements are in μm and in the form mean (range).

Table 5: Comparison of morphometrics of parasitic females in *Spilotylenchus* sp. and *Spilotylenchus maisonabei*.

Characters	*Spilotylenchus* sp. (this study)	*Spilotylenchus maisonabei* [23]
N	2	6
L	1,600–1,840	1,244 (1,200–1,320)
D	155–160	125 (107–160)
A	10.3–11.5	10.3 (7.5–12)
C	167.3–177.8	84.4 (64.5–121)
V%	97.4–97.7	96.2 (95.8–96.5)
Total length of stylet (St)	9.5–9.8	9-10
Distance between anterior end and excretory pore	1.5–15.5	23.3 (20–28)
Distance between anterior end and nerve ring	—	52–54
Total length of tail (Cd)	9–11	15.4 (10–19)
Distance between vulva and tail end	41.5–43	47 (42–52)
Distance between vulva and anus (V–A)	32-33	—

All measurements are in μm and in the form mean (range).

Table 6: Comparison of morphometrics of parasitic females in *Psyllotylenchus* sp. and *Psyllotylenchus viviparous*.

Character	*Psyllotylenchus* sp. (this study)		*Psyllotylenchus viviparous* [25]	
	Gamogenetic	Parthenogenetic	Gamogenetic	Parthenogenetic
N	3	7	8	10
L	1,016.7 (900–1,100)	446 (420–500)	1,000 (840–1,480)	500 (360–840)
D	81.3 (79–84)	70 (60–80)	77 (62–115)	60 (54–100)
A	12.5 (11.1–13.3)	6.25 (5.6–7)	—	—
C	64.3 (60–68.2)	40.15 (37.1–43.5)	—	—
V%	95.1 (95–95.4)	93.3 (90–95.3)	—	—
Total length of stylet (St)	17.5 (17–18,5)	5.25 (4–6)	17 (15–20)	7 (5–8)
Length of the distal edge of stylet	8.6 (8-9)	—	—	—
Distance between anterior end and excretory pore	26.5 (25–31.5)	17.5 (15–19.5)	23 (13–33)	22 (14–46)
Distance between anterior end and nerve ring	—	51.7 (50–55)	—	—
Total length of tail (Cd)	15.8 (15–17)	11.1 (10.5–11.5)	25 (17–35)	9 (1–17)
Distance between vulva and tail end	48 (45–51)	30.5 (19.7–55)	56 (37–71)	52 (40–104)
Distance between vulva and anus (V–A)	30.8 (29–31.5)	13.5 (11.7–21.6)	—	—

All measurements are in μm and in the form mean (range).

phytoparasites *B. xylophilus* and *Bursaphelenchus mucronatus* is from 1.7% to 3.7%. The spacers ITS1 and ITS2 are generally more diverged; the intra- and interspecific variation for these species is from 0 to 3.1% and 11.2 to 13.4%, respectively [121–123].

Molecular vouchering is proved to efficiently complement morphological species identification in nematodes [73, 122, 124–128]. Combining the rDNA and morphological data confirms the species identity within each of the three studied groups of isolates.

3.3. Phylogenetic Analysis. In phylogenetic analyses of rDNA we used a dataset with extensive species and gene sampling (SSU-ITS1-5.8S-LSU) compared to earlier published tylenchid phylogenies, most of which were based on SSU rDNA or D2-D3 expansion segments [17, 19–21, 39, 129]. The SSU-ITS1-5.8S-LSU rDNA tree topology (Figure 5) is highly similar to other published phylogenies of tylenchids. In this tree, tylenchomorphs are represented by the sister groups Aphelenchidae and Tylenchida. Most of the tylenchid clades occur in published trees but often contradict classifications based on morphology, as it was also noted by other authors [17, 19–21, 39, 129]. The three robust major branches in the SSU-ITS1-5.8S-LSU rDNA tree (Bayesian posterior probabilities of 0.99–1.0) are (1) the clade includes representatives of the suborders Hoplolaimina, Criconematina, and Tylenchina (excluding Anguinoidea); (2) the majority of classic Anguinata; (3) the suborder Hexatylina. The studied parasites of fleas form a monophyletic group (bootstrap support of 100%) within the Hexatylina.

The nonredundant rDNA data on the Hexatylina in GenBank mostly represents the D2-D3 expansion segments of LSU rDNA. To maximize species sampling of the Hexatylina, we chose the D3 expansion segment as the molecular marker. The phylogenetic tree with the Anguinoidea as an outgroup is shown in Figure 6. In this tree, the suborder Hexatylina consists of two well-supported clades, in accordance with previously published D2-D3 rDNA phylogenies [19, 20, 39]. The clade of the studied flea parasites is placed within the

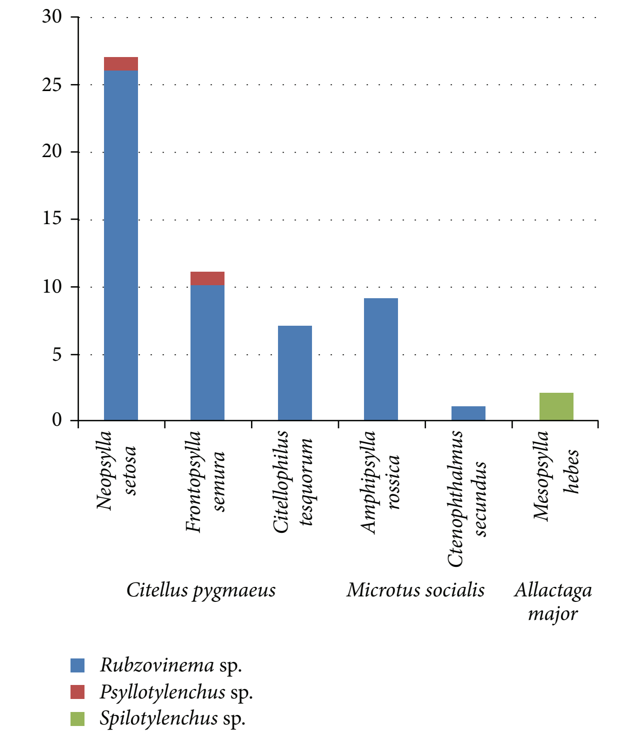


Figure 4: Distribution of the studied nematode species among the flea species studied, whose rodent hosts are given below. The vertical axis shows the numbers of infected fleas.

largest branch of the Hexatylina, similarly to the result of the concatenated rDNA analysis.

The three alternative relationships between the three major branches of Tylenchida (Figure 5) are not discriminated by the AU and Kishino and Hasegawa tests, and only the basal position of the Hexatylina is rejected by the expected-likelihood weights test (Table 7). All three tests do not discriminate between the alternative placement of the flea parasites as closest to the *Allantonema*, *Parasitylenchus*, or *Deladenus* branches; however, its positioning outside this grouping is not rejected only by a less conservative Shimodaira-Hasegawa test [50].

4. Discussion

4.1. Ribosomal DNA Phylogeny of the Tylenchida and Relationships within the Suborder Hexatylina. Phylogenetic analyses of SSU [16, 17, 19, 39] and D2-D3 [20, 39] rDNA data using various methods and species sampling generally agree on the monophyly of most tylenchid clades and contradict classic morphology based classifications. In the SSU-ITS1-5.8S-LSU tree (Figure 5), the monophyletic Tylenchida consists of three major robust clades. The first clade diverges into six groups: (1) the "Tylenchidae (part 2)" (by [17]), (2) the Tylodoridae (represented by the two genera, *Cephalenchus* and *Eutylenchus* [83]), (3) Boleodorinae + "Tylenchidae (part 1)" (by [Bert]), (4) the Merliniidae [130], (5) Criconematina + Sphaeronematidae + selected Tylenchina, and (6) Belonolaimidae + "Hoplolaimina." The Merliniidae group corresponds to Clade C in [19] and includes partially the polyphyletic "Telotylenchinae" [131], "Pratylenchidae", and "Hoplolaimina" (*Psilenchus* cf. *hilarulus*). Group (5) corresponds to Clade 12A in [129], where Sphaeronematidae (*Sphaeronema* and *Meloidoderita*) were earlier shown to be closely related to Criconematina [20, 89], and selected Ecphyadophoridae + *Ottolenchus* + *Malenchus* were found to represent a monophyletic clade within the paraphyletic Tylenchina likely to be related to the Criconematina [18, 82]. Group (6) corresponds to Clade VII in [20], Clade 12B in [129], and Clade A + Clade B in [19]. Belonolaimidae (the genera *Belonolaimus* and *Ibipora*) tend to occupy the basal position. Clade A in [19] contains a "long branch" of the burrowing nematode *Radopholus similes* ("Pratylenchidae") in sister position to the Hoplolaimidae [17, 19]. This nematode occupies a similar position relative to the Hoplolaimidae in the SSU-ITS1-5.8S-LSU tree, and we consider this unlikely to be an LBA artefact. Similarly to [95], *Carphodorus* and *Morulaimus* that belong to the classic Belonolaimidae comprise the basal branch of Clade A *sensu* [19]. The clade corresponding to Clade B in [19] contains Meloidogynidae, Dolichodoridae, paraphyletic Pratylenchidae, and a part of Telotylenchidae.

The second major clade of the Tylenchida includes representatives of the classic infraorder Anguinata, with a well-supported monophyletic origin, except for a few species. They belong outside the second clade and may initially have been wrongly identified.

The third major clade includes representatives of the classic suborder Hexatylina and consists of two groups. The smaller one unites the three species of *Sphaerularia*, *Helionema* sp., cf. *Hexatylus* sp., *Deladenus* sp. PDL-2005, and *Nothotylenchus acris* (Anguinata: Nothotylenchidae). It is further referred to as the Sphaerularioidea according to the type genus. The larger group contains the clade of studied flea parasites and members of the superfamilies Iotonchioidea (*Skarbilovinema* spp., *Parasitylenchus* spp., and *Wachekitylenchus bovieni*) and Sphaerularioidea (*Allantonema mirable*, *Bradynema* spp., *Howardula* spp., and *Contortylenchus* sp. (fam. Allantonematidae); *Deladenus durus*, *Deladenus proximus*, *Deladenus siricidicola*, *Fergusobia* spp., and *Gymnotylenchus* sp. (fam. Neotylenchidae)). One species of the Anguinata, *Sychnotylenchus* sp., also joins the larger group. Our study renders the genera *Howardula* and *Deladenus* paraphyletic, as was earlier shown in [19, 39, 71, 119].

The genus *Howardula* is paraphyletic in published rDNA and mitochondrial COI phylogenies [71]. Such characters of *Howardula* as the degeneration of oesophagus, tail shape, and the absence of stylet in males seem to have evolved independently by convergence. The paraphyletic genus *Deladenus* is more closely related to either ancestral forms of the Hexatylina or forms typical to the Anguinata. The infraorder Anguinata includes soil-dwelling nematodes, mostly mycetophagous or parasitizing various parts of plants. However, an unidentified entomoparasitic nematode was also grouped within the Anguinoidea [39]. The life cycle of *Deladenus* spp. is an irregular alternation of free-living and entomoparasitic forms. The nematode *D. siricidicola* is able of producing an unlimited number of free-living generations in the absence of the host larvae of siricid

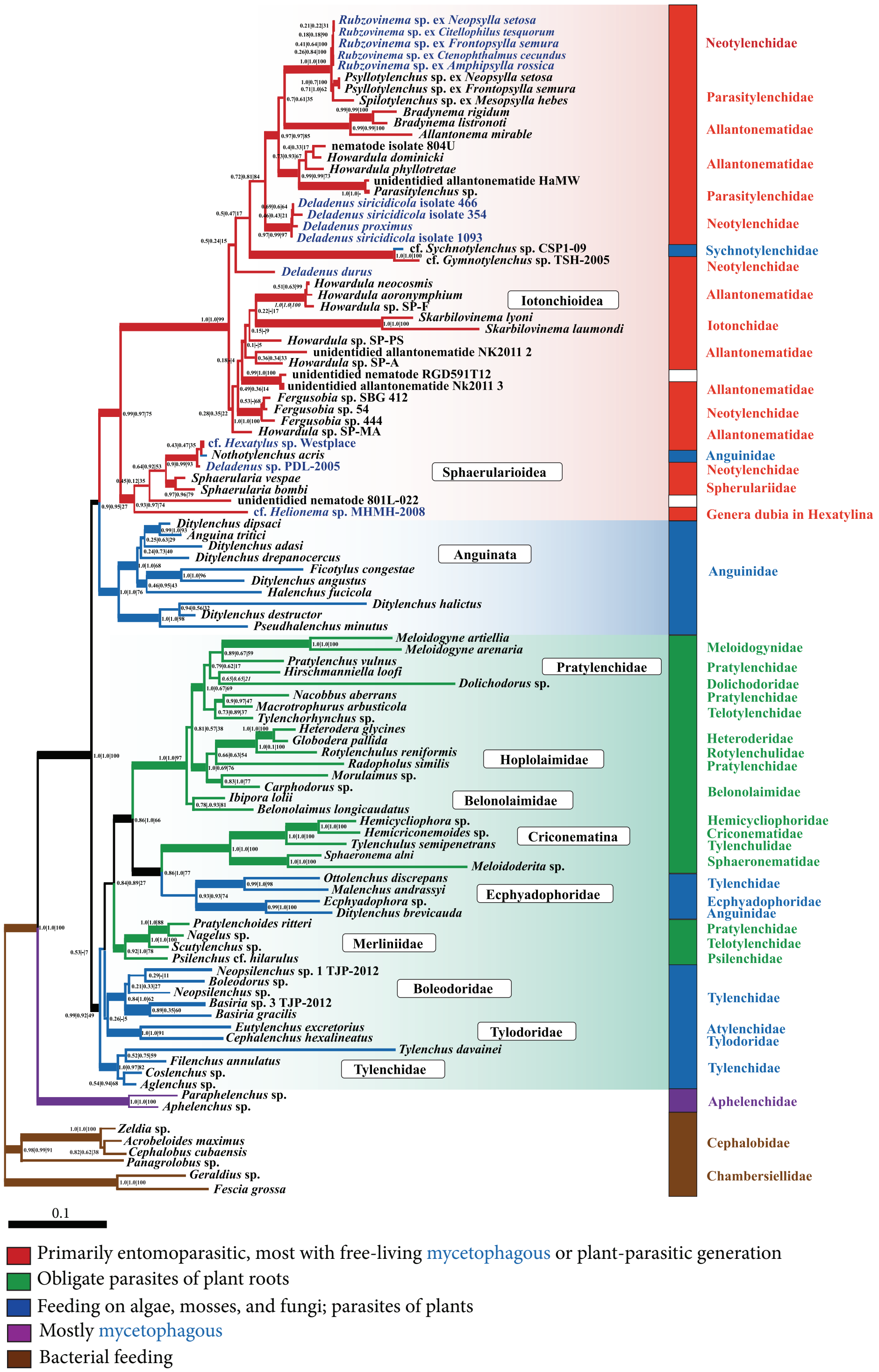


Figure 5: Phylogenetic tree of Tylenchida, inferred from SSU-ITS1-5.8S-LSU rDNA sequences. Topology was inferred using the PhyloBayes software (maxdiff = 0.36). Node support values are shown as follows: the first two values are Bayesian posterior probability assessed using the PhyloBayes and MrBayes software, respectively, and the third is bootstrap support assessed by the ML method. Thick lines lead to the nodes, in which at least one support value of posterior probability is 0.95 and higher. Names of clades (framed) are mainly given by type genera included in them (with the exception of Iotonchioidea). Formal taxonomic position (family by [8]) is shown on the right to the color bar. Colors indicate the ecologies (see the legend). Names of the species of Hexatylina that have a mycetophagous stage in their life cycle are shown in blue. The three robust major branches of Tylenchida are marked by gradient.

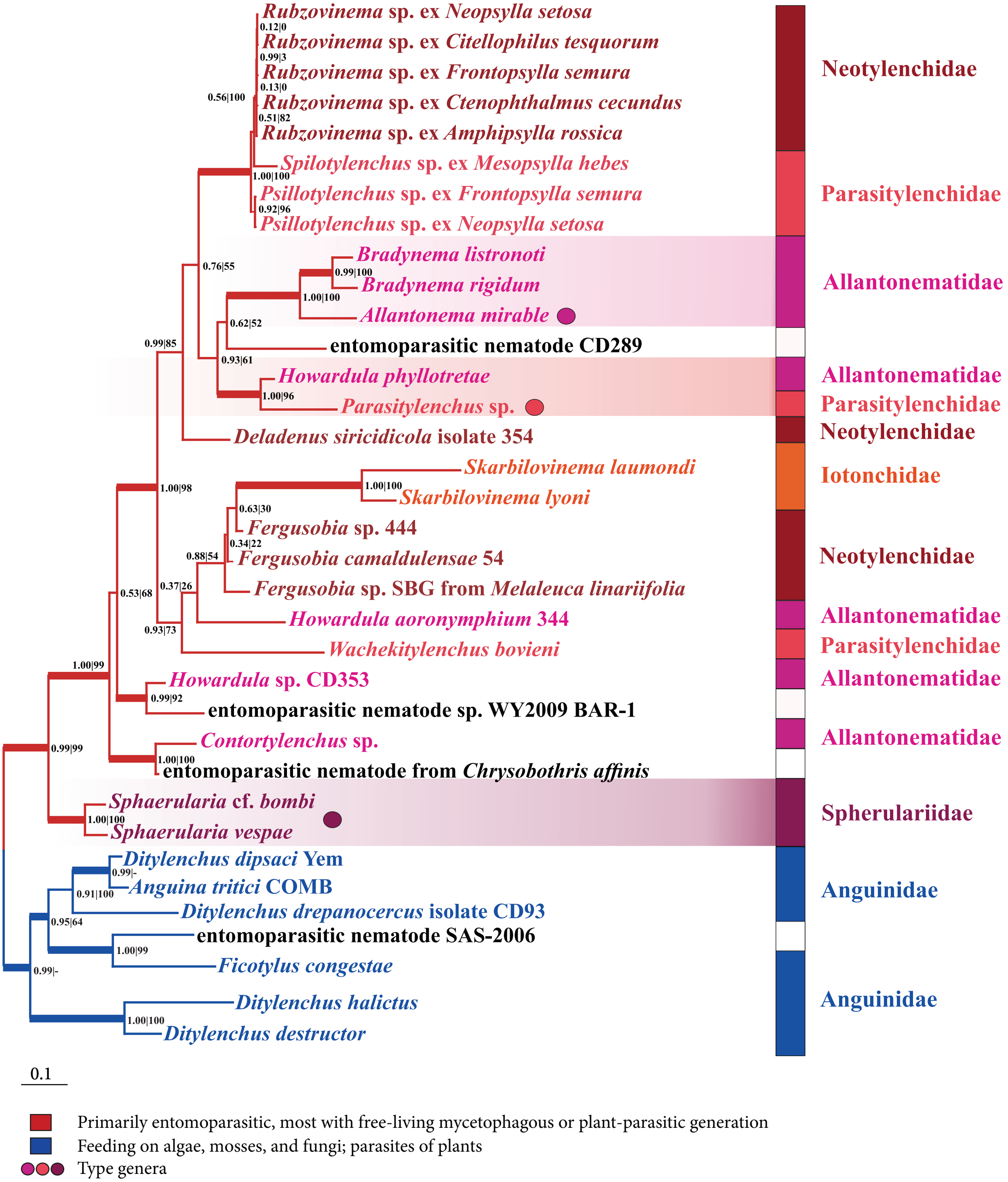


Figure 6: Phylogenetic tree of Hexatylina, inferred from D3 expansion segment of LSU rDNA. Topology was inferred using the PhyloBayes software. Node support values are shown as follows: Bayesian posterior probability/bootstrap support assessed by the ML method. Thick lines indicate the nodes supported at the level of 0.95 and higher. Color of lines indicates the ecologies (see the legend). Names of species were shown in different colors indicating their taxonomic position. Three families that include their type genera (shown as circles) are marked by gradient.

Table 7: Results of tree topology tests for alternative hypotheses on (1) the initial divergence of Tylenchida (Figure 4) and on (2) the relationships within the monophyletic branch that includes the studied group of nematodes parasitizing fleas (designated by asterisk).

Topology	Rank	obs	au	np	bp	pp	kh	sh	c-ELW
1									
(((H,An),T),o)	1	–1.8	0.787	0.415	0.402	0.804	0.663	0.969	0.4197
((An,(H,T)),o)	2	4.1	0.326	0.198	0.205	0.013	0.254	0.623	0.1848
((H,(An,T)),o)	3	6.9	0.061	0.013	0.014	0.001	0.101	0.492	0.0186
2									
((((∗,Al),P),Ds),o)	1	–1.8	0.787	0.415	0.402	0.804	0.663	0.969	0.4197
((((∗,P),Al),Ds),o)	2	1.8	0.495	0.242	0.247	0.130	0.337	0.813	0.2249
(((∗,(Al,P)),Ds),o)	3	2.7	0.371	0.110	0.105	0.052	0.243	0.824	0.1209
((∗,((Al,P),Ds)),o)	6	15.7	0.063	0.024	0.025	$1e-007$	0.053	0.153	0.0272
(((∗,Ds),(Al,P)),o)	7	18.3	0.013	0.002	0.002	$9e-009$	0.020	0.096	0.0028

Al: Allantonematidae, An: Anguinata, Ds: *Deladenus siricidicola—D. proximus* group, H: Hexatylina, P: Parasitylenchidae, T: Tylenchina, o: outgroup.

pine-killing wood wasps [132]. Like in Anguinata, the free-living forms of *Deladenus* spp. are fungal feeding. Such characters of *Deladenus* asthe mycetophagy, enlargement of subventral glands in entomoparasitic females versus their reduction in free-living forms, the hypertrophy of dorsal glands, and stylet reduction in free-living forms seem to be symplesiomorphic. Resemblance with the Anguinata is also typical of other mycetophagous free-living forms: *Hexatylus* (Neotylenchidae), *Rubzovinema* (Neotylenchidae), *Prothallonema* (Sphaerularioidae) *Helionema* (Hexatylina *dubia*), and Paurodontidae. For the latter, the entomoparasitic stage is expected but has never been observed. The relationship between the Hexatylina and Anguinata was earlier hypothesized based on morphology [7, 8, 130, 133, 134]. On rDNA phylogenies of tylenchids, the monophyly of the Hexatylina + Anguinata is either supported [19] or not rejected [20]. In the SSU-ITS1-5.8S-LSUrDNA tree obtained in this study, the monophyly of the Hexatylina + Anguinata has the Bayesian posterior probability of 0.91, but the maximum-likelihood bootstrap support is low; the AU and Kishino and Hasegawa tests did not discriminate between alternative hypotheses.

According to our SSU-ITS1-5.8S-LSU rDNA phylogeny (Figure 5), the major robust branches of the Tylenchida are incongruent with morphology-based classifications suggesting three rather than four suborders (the rank is adopted from morphological systems of tylenchids). Among them, the Hexatylina and Anguinata (both are monophyletic) are likely to be sister groups. The third emerged suborder includes representatives of three classic suborders: Tylenchina, Hoplolaimina, and Criconematina, among which only the latter does not contradict morphology-based classifications.

Considering ecological traits coded in Figure 5, the mycetophagy and/or facultative ectophytoparasitism are likely to be ancestral in the Tylenchida. Sedentary phytoparasites (root-knot species of *Meloidogyne*, the false root-knot genus *Nacobbus*, and cyst-forming *Heterodera* and *Globodera*) and other obligate endoparasites of plants evolved several times from free-living or facultative sedentary forms, as it was previously hypothesized in accordance with the concept of evolutionary trend to endoparasitism in phytonematodes [135]. Similarly, obligate endoparasites of insects from the Hexatylina are likely to have evolved from mycetophagous forms, with some species retaining the ancestral mycetophageous stage in the life cycle (e.g., species of the paraphyletic genus *Deladenus* and flea nematodes of the genus *Rubzovinema*). An interesting specific case in the Hexatylina is the genus *Fergusobia* that includes plant parasites associated with insects [68, 70], which may have transited to plant parasitism via entomoparasitism [39].

4.2. Ribosomal DNA Phylogeny of the Flea Nematodes and Their Classification. The nematodes of fleas do not group with the families known as their relatives in morphology-based systems, as these families do not form monophyletic groups in the tree. However, they do group with both type genera of the families Parasitylenchidae and Allantonematidae (*Parasitylenchus* and *Allantonema*, resp.). This grouping is preceded by a successive divergence of *Deladenus durus* and *Deladenus siricidicola* (Figure 5). As mentioned above, the pronounced free-living form in *Deladenus* seems to be ancestral to this group.

Only 31 tylenchid species that parasitize in fleas have been described to date. They differ by morphology, life cycle, and the host specificity, and belong to the five genera: *Spilotylenchus* (8 species), *Psyllotylenchus* (20 species), *Incurvinema* (1 species) *Kurochkinitylenchus* (1 species), and *Rubzovinema* (1 species). According to the classification of Siddiqi [8], the genera *Spilotylenchus* and *Psyllotylenchus* belong to the family Parasitylenchidae, whereas the genus *Rubzovinema* is a member of the Neotylenchidae. The two families represent two superfamilies, Iotonchioidea and Sphaerularioidea, respectively. All rDNA phylogenies published to date suggest that these superfamilies are paraphyletic [19, 20, 39], which is also inferred in our study with an extensive gene and taxon sampling.

A high degree of rDNA similarity in the three studied species suggests a closer relationship of these species than that assumed by the accepted system of classification. Earlier, Slobodyanyuk proposed to unite all known flea parasites into one family, the Spilotylenchidae. Its four subfamilies, Spilotylenchinae, Rubzovinematinae, Psyllotylenchinae, and Kurochkinitylenchinae, are discriminated based on the life cycle features [28]. In Spilotylenchinae and Rubzovinematinae, the entomoparasitic stage is represented by parasitic females of one heterosexual generation. In Psyllotylenchinae, in addition to the heterosexual generation, a parthenogenetic generation occurs in the flea haemocoel. In Kurochkinitylenchinae, two heterosexual generations exist in the haemocoel: the first generation produces parasitic females and the second generation produces both females and males [28]. Siddiqi also considered the unification of all flea tylenchids into one family but observed the need for further evidence in support [8].

Our results strongly suggest the inclusion of the three genera, *Rubzovinema*, *Psyllotylenchus*, and *Spilotylenchus*, in one family, the Spilotylenchidae [28]. The ribosomal DNA genetic distance within the family Spilotylenchidae is much smaller than that of certain tylenchid genera, for example, *Meloidogyne* (Figure 4) or *Pratylenchus* [19, 84].

4.3. Host Specificity of Flea Nematodes. The majority of tylenchid nematodes are monoxenous or oligoxenous; in particular, flea parasites were thought to be strictly host specific. Earlier papers suggested the lack of strict host specificity in *Psyllotylenchus pawlowskyi* and *Psyllotylenchus viviparous* [13, 25]. However, later these species were found to be heterogeneous and sustained revision [9, 27–29]. *Spilotylenchus pawlowskyi* and *Spilotylenchus caspius* were referred to as single-host parasites of the flea *Coptopsylla lamellifer* [27, 136]. *Kurochkinitylenchus laevicepsi* and *Spilotylenchus ivashkini* also share the same flea host, *Nosopsyllus laeviceps* [28, 29]. Before our study, the genus *Rubzovinema* was known to contain a single species, *Rubzovinema ceratophylla*, which parasitizes exclusively the flea *Citellophilus tesquorum*.

We found that at least two out of the three studied species are not single-host parasites. *Psyllotylenchus* sp. was shown to parasitize two flea species feeding on sousliks, *Frontopsylla semura* and *Neopsylla setosa*. *Rubzovinema* sp. was found on five flea species feeding on different rodent hosts: *C. tesquorum*, *F. semura*, *N. setosa* (all sampled from sousliks), *Ctenophtalamus secundus*, and *Amphipsylla rossica* (all sampled from voles). *A. rossica, F. semura*, and *C. tesquorum* belong to different families of the superfamily Ceratophylloidea (Leptopsyllidae and Ceratophyllidae), whereas *C. secundus* and *N. setosa* belong to the superfamily Hystrichopsylloidea. Unlike the host-specific *R. ceratophylla*, the studied *Rubzovinema* sp. parasitizes taxonomically distant fleas feeding on different rodents. Thus, the common opinion that flea nematodes are strictly host specific should be revisited.

As the two species of *Rubzovinema* demonstrate, even closely related parasites may exhibit different host range size. Among other known examples are the entomoparasitic nematodes of the genus *Howardula* parasitizing various beetles and flies [71, 137, 138], many phytonematodes [8], sibling species of parasitoid flies [128], and herbivorous insects [139]. The host range of parasites is an indicator of their evolutionary strategy in the ecosystem. Multihost parasites can be considered ecological generalists, in contrast to specialists that coevolve with a particular host. Generalists and specialists play different roles in the ecosystem [140], where they keep in balance, taking advantages and disadvantages of the two strategies. The advantages of generalization are yet to be explained by evolutionary biologists, whereas advantages of specialization are obvious, and it is generally accepted that evolution favors specialism [141, 142]. In the flea parasites, this trend is demonstrated by a greater species diversity of ecological specialists, the genera *Spilotylenchus* and *Psyllotylenchus*.

Nevertheless, the generalist *Rubzovinema* sp. was most abundant in the studied samples, which indicates that extending the host range may be evolutionarily successful. Besides the need to combat the immune response of several hosts, which is a requirement to widen the hosts range [143], the free-living stage of *Rubzovinema* sp. is to adapt to diverse microbioclimatic conditions of complex environments of rodent habitats. Multihost parasites pay a cost of adapting to alternative conditions [141, 144] compensated by stable survival of the species. Considering the spatial and temporal dynamics of flea populations feeding on a particular rodent host (one or two flea species usually dominate over a sampling season), multihost nematode parasites gain an advantage of their relative independence of population waves of either flea hosts or their rodent hosts. A higher infestation rate observed for *Rubzovinema* sp., compared to the two other studied species, may be an indicator of a greater ecological plasticity of this multihost parasite.

4.4. Entomoparasitic Nematodes in Natural Foci of Plague. In natural foci of plague, the epizootic dynamics are influenced by numerous climatic and biotic factors. The spatial and temporal population dynamics of the plague agent, *Y. pestis*, affect the population dynamics of the flea vectors and their mammalian hosts. Members of the transmission route of the plague agent also closely interact with other living organisms. For example, parasites of fleas that in turn feed on rodents are hyperparasites that play the role of high-level control agents on the ecosystem level, the role that entomoparasitic nematodes share with the bacterial plague agent. High-level control agents render the epidemiological state of a natural focus of disease less predictable. On the one hand, a lower density of the flea vector population reduces the plaque transmission rate; on the other, its growth causes an exponential decay of the host rodent population [145] below its epidemiological threshold, above which there is a threat of spillover of plague infection into human population [145]. Hypothetically, nematode-induced decrease of flea population is able to increase the number of rodents above the threshold and thus trigger an epidemic. The dual effect of high-level control agents is well exemplified by cases, when during plague episodes the extermination of rodents by humans causes the return of infection through stimulating the migration of fleas, the plaque vectors [5].

The studied entomoparasitic nematodes possess high potential as control agents of the flea vectors of plague owing to their high propagation rate within the flea host (Figure 2) and high infestation level (up to 21% observed in this study and from 50 to 60%, as estimated by other authors

[10, 11]). One of the studied nematode species, *Rubzovinema* sp., is a multihost parasite. Host-specific parasites reach the optimal level of pathogenicity by maintaining the trade-off between pathogenicity and transmissibility. Adding of a new host to a multihost system makes the model more complicated [141]. The multihost parasite *Rubzovinema* sp. is expected to exhibit different levels of pathogenicity with respect to different flea hosts which, in turn, play different roles in the transmission of plague. Epizootics cause sporadic mortality in local populations of all members involved in the interaction with the plague agent, and their survival is contingent on migrations within a metapopulation. It is the case when the Cope's law [139, 146] governs the extinction of specialists on a shorter time scale rather than a geological period, and evolution may favor the ecological generalists, such as *Rubzovinema* sp.

Some authors surmised the involvement of entomoparasitic nematodes in the transmission of the plague agent [4], as it was observed that biofilms of *Yersinia pestis* adhere to cuticle receptors of *Caenorhabditis elegans* [147–149]. In this perspective, nematodes parasitizing fleas in natural foci of plague take on greater importance, as they may provide for the transmission route that does not include a mammal [4]. Further studies will clarify the role of flea nematodes in the transmission of plague infection.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank G. S. Mirzaeva for help with PCR amplification and rDNA fragment analysis, N. V. Popov and the staff of the Laboratory of Epizootics Monitoring for advice and assistance in collecting and processing rodent samples, and, particularly, A. N. Porshakov for help in identification of flea specimens. They also thank O. V. Slobodyanyuk for helpful discussions of results, S. E. Spiridonov for advice on cultivation of entomoparasitic nematodes, S. A. Subbotin for valuable comments on the earlier version of the paper, and E. Yu. Talanova and L. Yu. Rusin for discussions of its final version, and E. A. Musatkina for assistance with the manuscript preparation. They are grateful to the Supercomputer Center of Moscow State University (http://parallel.ru/cluster) and the Bioportal of the University of Oslo (http://www.bioportal.uio.no) for providing computing resources.

References

[1] A. W. Bacot and C. J. Martin, "Observations on the mechanism of the transmission of plague by fleas," *The Journal of Hygiene*, vol. 13, supplement 3, pp. 423–439, 1914.

[2] V. V. Kutyrev, G. A. Eroshenko, N. V. Popov, N. A. Vidyaeva, and N. P. Konnov, "Molecular mechanisms of interaction of the plague agent with invertebrates," *Molecular Genetics, Microbiology and Virology*, vol. 24, no. 4, pp. 7–12, 2009.

[3] E. N. Pavlovsky, *Natural Focality of Transmissive Diseases in Connection with Landscape Epidemiology of Zooanthroponoses*, Moscow, Russia, 1964.

[4] N. V. Popov, E. I. Koshel, G. A. Eroshenko, and V. V. Kutyrev, "Formation of modern concepts on the mechanism of plague enzooty," *Problems of Particularly Dangerous Infections*, vol. 3, no. 109, pp. 5–8, 2011.

[5] M. J. Keeling and C. A. Gilligan, "Bubonic plague: a metapopulation model of a zoonosis," *Proceedings of the Royal Society B*, vol. 267, no. 1458, pp. 2219–2230, 2000.

[6] R. A. Bedding, R. J. Akhurst, and H. K. Kaya, *Nematodes and the Biological Control of Insect Pests*, CSIRO Press, Melbourne, Australia, 1993.

[7] M. R. Siddiqi, *Tylenchida Parasites of Plants and Insects*, vol. 645 of *Commonwealth Agricultural Bureaux*, Farnham Royal, Slough, UK, 1986.

[8] M. R. Siddiqi, *Tylenchida: Parasites of Plants and Insects*, CABI, Wallingford, UK, 2nd edition, 2000.

[9] I. A. Rubtsov, *Parasites and Enemies of Fleas*, Nauka, Leningrad, Russia, 1981.

[10] Y. A. Morozov, "About infestation with fleas great gerbils different ages," in *Proceedings of the Conference Anti-Plague Facilities in Central Asia and Kazakhstan*, pp. 337–338, Alma-Ata, 1974.

[11] Y. A. Morozov, "Effect of infestation of nematode on reproduction of fleas gerbils in Muyunkum," in *Proceedings of the Conference Anti-Plague Facilities in Central Asia and Kazakhstan*, pp. 338–340, Alma-Ata, 1974.

[12] J. Deunff, *Parasites de Siphonaptères. Étude de la systématique, de la biologie et du pouvoir pathogène des Tylenchides (Nematodea) dans une perspective de lutte biologique [Ph.D. thesis]*, Etat Sc. Pham, Rennes, France, 1984.

[13] Y. V. Kurochhkin, "The nematode Heterotylenchus pawlowskyi sp. n., castrating flea-vectors of plague," *Doklady Akademyi Nauk SSSR*, vol. 135, pp. 1281–1284, 1960.

[14] M. L. Blaxter, P. De Ley, J. R. Garey et al., "A molecular evolutionary framework for the phylum Nematoda," *Nature*, vol. 392, no. 6671, pp. 71–75, 1998.

[15] P. De Ley and M. L. Blaxter, "Systematic Position and Phylogeny," in *The Biology of Nematodes*, D. L. Lee, Ed., Taylor & Francis, London, UK, 2002.

[16] B. H. M. Meldal, N. J. Debenham, P. De Ley et al., "An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa," *Molecular Phylogenetics and Evolution*, vol. 42, no. 3, pp. 622–636, 2007.

[17] W. Bert, F. Leliaert, A. R. Vierstraete, J. R. Vanfleteren, and G. Borgonie, "Molecular phylogeny of the Tylenchina and evolution of the female gonoduct (Nematoda: Rhabditida)," *Molecular Phylogenetics and Evolution*, vol. 48, no. 2, pp. 728–744, 2008.

[18] M. Holterman, A. Van Der Wurff, S. Van Den Elsen et al., "Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1792–1800, 2006.

[19] M. Holterman, G. Karssen, S. Van Den Elsen, H. Van Megen, J. Bakker, and J. Helder, "Small subunit rDNA-based phylogeny of the tylenchida sheds light on relationships among some high-impact plant-parasitic nematodes and the evolution of plant feeding," *Phytopathology*, vol. 99, no. 3, pp. 227–235, 2009.

[20] S. A. Subbotin, D. Sturhan, V. N. Chizhov, N. Vovlas, and J. G. Baldwin, "Phylogenetic analysis of Tylenchida Thorne, 1949 as inferred from D2 and D3 expansion fragments of the 28S rRNA gene sequences," *Nematology*, vol. 8, no. 3, pp. 455–474, 2006.

[21] S. A. Subbotin, D. Sturhan, B. J. Adams et al., "Molecular phylogeny of the order Tylenchida: analysis of nuclear ribosomal RNA genes," *Journal of Nematology*, vol. 38, no. 2, pp. 296–296, 2006.

[22] H. Launay, J. Deunff, and O. Bain, "*Spilotylenchus arthuri*, gen. n., sp. n. (Nematodea, Tylenchida: Allantonematidae), parasite of *Spilopsyllus cuniculi* (Dale, 1878) (Siphonaptera: Pulicidae)," *Annales de Parasitologie Humaine et Comparee*, vol. 58, no. 2, pp. 141–150, 1983.

[23] H. Launay and J. Deunff, "*Spilotylenchus maisonabei* n. sp. (Nematoda: Allantonematidae) parasite of the European rabbit flea *Spilopsyllus cuniculi* (Dale, 1878) (Siphonaptera: Pulicidae)," *Annales de Parasitologie Humaine et Comparee*, vol. 13, pp. 293–296, 1990.

[24] C. Laumond and J. C. Beaucournu, "*Neoparasitylenchus megabothridis* n. sp. (Tylenchida: Allantonematidae) parasite of *Megabothris tubidus* (Siphonaptera: Ceratophyllidae); observations on fleas tylenchides in the S.W. of Europa (author's transl)," *Annales de Parasitologie Humaine et Comparee*, vol. 53, no. 3, pp. 291–302, 1978.

[25] G. O. Poinar Jr. and B. C. Nelson, "*Psyllotylenchus viviparus*, n. gen., n. sp. (Nematodea: Tylenchida: Allantonematidae) parasitizing fleas (Siphonaptera) in California," *Journal of Medical Entomology*, vol. 10, no. 4, pp. 349–354, 1973.

[26] V. O. Slobodyanyuk, "Validation of genus Rubzovinema gen.n. (Sphaerularioidea) and redescription of species *Rubzovinema ceratophylla* comb.n. parasite of fleas *Citellophillus tesquorum*," *Zoological Journal*, vol. 70, no. 9, pp. 33–43, 1991.

[27] O. V. Slobodyanyuk, "Revision of the species *Psyllotylenchus pawlowskyi* (Kurochkin, 1960) Poinar & Nelson, 1973. I. Redescription of Spilotylenchus pawlowskyi (sensu stricto) comb. n. (Tylenchida: Allantonematidae)," *Russian Journal of Nematology*, vol. 5, no. 2, pp. 103–112, 1997.

[28] O. V. Slobodyanyuk, "Revision of the species *Psyllotylenchus pawlowskyi* (Kurochkin, 1960) Poinar & Nelson, 1973. II. Description of *Kurochkinitylenchus laevicepsi* gen. n., sp. n. and Spilotylenchidae fam. n," *Russian Journal of Nematology*, vol. 7, no. 1, pp. 1–18, 1999.

[29] O. V. Slobodyanyuk, "Revision of the species *Psyllotylenchus pawlowskyi* (Kurochkin, 1960) Poinar & Nelson, 1973. III. Description of *Spilotylenchus ivashkini* sp. n," *Russian Journal of Nematology*, vol. 8, no. 1, pp. 45–55, 2000.

[30] J. Deunff and H. Launay, "*Psyllotylenchus chabaudi*, n. sp. (Nematodea, Tylenchida: Allantonematidae), a parasite of *Nosopsyllus fasciatus* (Bosc) (Siphonaptera: Ceratophyllidae)," *Annales de Parasitologie Humaine et Comparee*, vol. 59, no. 3, pp. 263–270, 1984.

[31] J. Deunff, H. Launay, and J. C. Beaucournu, "*Incurvinema helicoides* n. gen., n. sp. Nematodea, Tylenchida: allantonematidae parasite de *Rhadinopsylla pentacantha* (Rothschild, 1897) (Siphonaptera : Hystrichopsyllidae)," *Annales de Parasitologie Humaine Et Comparée*, vol. 60, pp. 739–746, 1985.

[32] A. T. De Grisse, "Redescriptions ou modifications de quelques techniques utilisées dans l'étude des nématodes phytoparasites," *Mededelingen Rijksfakulteit Landbouwwetenschappen Gent*, vol. 34, pp. 351–369, 1969.

[33] X. Huang, "A contig assembly program based on sensitive detection of fragment overlaps," *Genomics*, vol. 14, no. 1, pp. 18–25, 1992.

[34] T. A. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Symposium Series*, vol. 41, pp. 95–98, 1999.

[35] L. Medlin, H. J. Elwood, S. Stickel, and M. L. Sogin, "The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions," *Gene*, vol. 71, no. 2, pp. 491–499, 1988.

[36] V. N. Chizhov, O. A. Chumakova, S. A. Subbotin, and J. G. Baldwin, "Morphological and molecular characterization of foliar nematodes of the genus *Aphelenchoides:* A. fragariae and *A. ritzemabosi* (Nematoda: Aphelenchoididae) from the Main Botanical Garden of the Russian Academy of Sciences, Moscow," *Russian Journal of Nematology*, vol. 14, no. 2, pp. 179–184, 2006.

[37] G. Van der Auwera, S. Chapelle, and R. De Wächter, "Structure of the large ribosomal subunit RNA of *Phytophthora megasperma*, and phylogeny of the oomycetes," *FEBS Letters*, vol. 338, no. 2, pp. 133–136, 1994.

[38] S. A. Subbotin, N. Vovlas, R. Crozzoli et al., "Phylogeny of Criconematina Siddiqi, 1980 (Nematoda: Tylenchida) based on morphology and D2-D3 expansion segments of the 28S rDNA gene sequences with application of a secondary structure model," *Nematology*, vol. 7, pp. 927–944, 2005.

[39] V. N. Chizhov, N. N. Butorina, and S. A. Subbotin, "Entomoparasitic nematodes of the genus Skarbilovinema: *S. laumondi* and *S. lyoni* (Nematoda: Tylenchida), parasites of the flies of the family Syrphidae (Diptera), with phylogeny of the suborder Hexatylina," *Russian Journal of Nematology*, vol. 20, no. 2, pp. 141–155, 2012.

[40] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[41] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[42] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, Oxford, UK, 2000.

[43] N. Lartillot, T. Lepage, and S. Blanquart, "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating," *Bioinformatics*, vol. 25, no. 17, pp. 2286–2288, 2009.

[44] N. Lartillot and H. Philippe, "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1095–1109, 2004.

[45] J. P. Huelsenbeck and F. Ronquist, "MRBAYES: bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[46] C. Lanave, G. Preparata, C. Saccone, and G. Serio, "A new method for calculating evolutionary substitution rates," *Journal of Molecular Evolution*, vol. 20, no. 1, pp. 86–93, 1984.

[47] A. Stamatakis, "Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, pp. 2688–2690, 2006.

[48] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection," *Systematic Biology*, vol. 51, no. 3, pp. 492–508, 2002.

[49] H. Kishino and M. Hasegawa, “Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoide,” *Journal of Molecular Evolution*, vol. 29, no. 2, pp. 170–179, 1989.

[50] H. Shimodaira and M. Hasegawa, “CONSEL: for assessing the confidence of phylogenetic tree selection,” *Bioinformatics*, vol. 17, no. 12, pp. 1246–1247, 2002.

[51] K. Strimmer and A. Rambaut, “Inferring confidence sets of possibly misspecified gene trees,” *Proceedings of the Royal Society B*, vol. 269, no. 1487, pp. 137–142, 2002.

[52] H. A. Schmidt, K. Strimmer, M. Vingron, and A. Von Haeseler, “TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing,” *Bioinformatics*, vol. 18, no. 3, pp. 502–504, 2002.

[53] R. D. M. Page, “TreeView: an application to display phylogenetic trees on personal computers,” *Computer Applications in the Biosciences*, vol. 12, no. 4, pp. 357–358, 1996.

[54] O. Holovachov, S. Bostrom, I. T. De Ley et al., “Morphology, molecular characterisation and systematic position of the genus *Cynura* Cobb, 1920 (Nematoda: Plectida),” *Nematology*, vol. 15, pp. 611–627, 2013.

[55] S. A. Nadler, P. De Ley, M. Mundo-Ocampo et al., “Phylogeny of Cephalobina (Nematoda): molecular evidence for recurrent evolution of probolae and incongruence with traditional classifications,” *Molecular Phylogenetics and Evolution*, vol. 40, no. 3, pp. 696–711, 2006.

[56] O. Holovachov, S. Bostrom, S. A. Nadler, and P. De Ley, “Systematics and phylogenetic position of the genus *Tricirronema* Siddiqi, 1993 (Cephalobomorpha),” *Journal of Nematode Morphology and Systematics*, vol. 12, no. 2, 2009.

[57] K. Kiontke, A. Barrière, I. Kolotuev et al., “Trends, stasis, and drift in the evolution of nematode vulva development,” *Current Biology*, vol. 17, no. 22, pp. 1925–1937, 2007.

[58] R. Campos-Herrera, F. E. El-Borai, and L. W. Duncan, “Wide interguild relationships among entomopathogenic and free-living nematodes in soil as measured by real time qPCR,” *Journal of Invertebrate Pathology*, vol. 111, no. 2, pp. 126–135, 2012.

[59] M.-A. Félix, P. De Ley, R. J. Sommer et al., “Evolution of vulva development in the Cephalobina (Nematoda),” *Developmental Biology*, vol. 221, no. 1, pp. 68–86, 2000.

[60] S. Boström, O. Holovachov, and S. A. Nadler, “Description of *Scottnema lindsayae* Timm, 1971 (Rhabditida: Cephalobidae) from Taylor Valley, Antarctica and its phylogenetic relationship,” *Polar Biology*, vol. 34, no. 1, pp. 1–12, 2011.

[61] J. W. Waceke, D. J. Bumbarger, M. Mundo-Ocampo, S. A. Subbotin, and J. G. Baldwin, “*Zeldia spannata* n. sp. (Nematoda: Cephalobidae) from the Mojave Desert, California,” *Journal of Nematode Morphology and Systematics*, vol. 8, no. 1, pp. 57–67, 2005.

[62] S. Kumari, “*Aphelenchus avenae* (Nematoda: Aphelenchidae) under the rhizosphere of *Brassica napus*,” *Helminthologia*, vol. 49, no. 1, pp. 57–59, 2012.

[63] H. Iwahori, K. Tsuda, N. Kanzaki, K. Izui, and K. Futai, “PCR-RFLP and sequencing analysis of ribosomal DNA of *Bursaphelenchus* nematodes related to pine wilt disease,” *Fundamental and Applied Nematology*, vol. 21, no. 6, pp. 655–666, 1998.

[64] L. K. Carta, A. M. Skantar, Z. A. Handoo, and M. A. Baynes, “Supplemental description of *Paraphelenchus acontioides* (Tylenchida: Aphelenchidae, Paraphelenchinae), with ribosomal DNA trees and a morphometric compendium of female *Paraphelenchus*,” *Nematology*, vol. 13, no. 8, pp. 887–899, 2011.

[65] Y. Zeng, R. M. Giblin-Davis, Y. Weimin, G. Bélair, G. Boivin, and K. W. Thomas, “*Bradynema listronoti* n. sp. (Nematoda: Allantonematidae), a parasite of the carrot weevil *Listronotus oregonensis* (Coleoptera: Curculionidae) in Quebec, Canada,” *Nematology*, vol. 9, no. 5, pp. 609–623, 2007.

[66] K. Rybarczyk-Mydłowska, P. Mooyman, H. van Megen, and Setal, “Small subunit ribosomal DNA-based phylogenetic analysis of foliar nematodes (Aphelenchoides spp.) and their quantitative detection in complex DNA backgrounds,” *Phytopathology*, vol. 102, no. 12, pp. 1153–1160, 2012.

[67] Q. Yu, P. de Groot, I. Leal, C. Davis, W. Ye, and B. Foord, “First report and characterization of *Deladenus proximus* (Nematoda: Neotylenchidae) associated with *Sirex nigricornis* (Hymenoptera: Siricidae) in Canada,” *International Journal of Nematology*, vol. 21, no. 2, pp. 139–146, 2012.

[68] W. Ye, R. M. Giblin-Davis, K. A. Davies et al., “Molecular phylogenetics and the evolution of host plant associations in the nematode genus *Fergusobia* (Tylenchida: Fergusobiinae),” *Molecular Phylogenetics and Evolution*, vol. 45, no. 1, pp. 123–141, 2007.

[69] I. Leal, B. Foord, C. Davis, P. de Groot, X. O. Mlonyeni, and B. Slippers, “Distinguishing isolates of *Deladenus siricidicola* ,a biological control agent of *Sirex noctilio*, from North America and the Southern Hemisphere using PCR- RFLP,” *Canadian Journal of Forest Research*, vol. 42, pp. 1173–1177, 2012.

[70] K. A. Davies, W. Ye, R. M. Giblin-Davis, G. S. Taylor, S. Scheffer, and W. K. Thomas, “The nematode genus *Fergusobia* (Nematoda: Neotylenchidae): molecular hylogeny, descriptions of clades and associated galls, host plants and Fergusonina fly larvae,” *Zootaxa*, no. 2633, pp. 1–66, 2010.

[71] S. J. Perlman, G. S. Spicer, D. Dewayne Shoemaker, and J. Jaenike, “Associations between mycophagous *Drosophila* and their *Howardula* nematode parasites: a worldwide phylogenetic shuffle,” *Molecular Ecology*, vol. 12, no. 1, pp. 237–249, 2003.

[72] C. L. Raak-van den Berg, P. S. van Wielink, P. W. de Jong et al., “Invasive alien species under attack: natural enemies of *Harmonia axyridis* in the Netherlands,” *BioControl*. In press.

[73] N. Kanzaki, R. M. Giblin-Davis, R. H. Scheffrahn et al., “Reverse taxonomy for elucidating diversity of insect-associated nematodes: a case study with termites,” *PLoS ONE*, vol. 7, no. 8, Article ID e43865, 2012.

[74] T. O. Powers, D. A. Neher, P. Mullin et al., “Tropical nematode diversity: vertical stratification of nematode communities in a Costa Rican humid lowland rainforest,” *Molecular Ecology*, vol. 18, no. 5, pp. 985–996, 2009.

[75] C. Hazir, R. M. Giblin-Davis, N. Keskin et al., “Diversity and distribution of nematodes associated with wild bees in Turkey,” *Nematology*, vol. 12, no. 1, pp. 65–80, 2010.

[76] M. Holterman, S. van den Elsen, H. van Megen et al., “Evolutionary relationships between members of the phylum Nematoda based on small subunit ribosomal DNA sequences,” 2004.

[77] N. Kanzaki, H. Kosaka, K. Sayama, J.-I. Takahashi, and S. Makino, “*Sphaerularia vespae* sp. nov. (Nematoda, Tylenchomorpha, Sphaerularioidea), an endoparasite of a common Japanese hornet, *Vespa simillima* Smith (Insecta, Hymenoptera, Vespidae),” *Zoological Science*, vol. 24, no. 11, pp. 1134–1142, 2007.

[78] Q. Yu, “*Ditylenchus destructor* Thorne, 1945 (Tylenchida: Anguinidae) in Canada,” *Journal of Nematology*, vol. 44, no. 4, pp. 500–500, 2012.

[79] B. Steven, L. V. Gallegos-Graves, C. Yeager, J. Belnap, and C. R. Kuske, "Common and distinguishing features of the bacterial and fungal communities in biological soil crusts and shrub root zone soils," *Soil Biology and Biochemistry*, vol. 69, pp. 302–312, 2014.

[80] R. D. L. Oliveira, A. M. Santin, D. J. Seni et al., "*Ditylenchus gallaeformans* sp. n. (Tylenchida: Anguinidae)—a neotropical gall forming nematode with biocontrol potential against weedy Malastomataceae," *Nematology*, vol. 15, no. 2, pp. 179–196, 2013.

[81] K. A. Davies, W. Ye, R. M. Giblin-Davis, and K. W. Thomas, "*Ficotylus congestae* gen. n., sp. n. (Anguinata), from ficus congesta (moraceae) sycones in Australia," *Nematology*, vol. 11, no. 1, pp. 63–75, 2009.

[82] M. R. Atighi, E. Pourajam, T. J. Pereira et al., "Redescription of *Filenchus annulatus* (Siddiqui & Khan, 1983) Siddiqi, 1986 based on specimens from Iran with contributions to the molecular phylogeny of the Tylenchida," *Nematology*, vol. 15, no. 2, pp. 129–141, 2013.

[83] J. E. Palomares-Rius, S. A. Subbotin, G. Liébanas, B. B. Landa, and P. Castillo, "*Eutylenchus excretorius* Ebsary & Eveleigh, 1981 Nematoda: Tylodorinae) from Spain with approaches to olecular phylogeny of related genera," *Nematology*, vol. 11, no. 3, pp. 343–354, 2009.

[84] S. A. Subbotin, E. J. Ragsdale, T. Mullens, P. A. Roberts, M. Mundo-Ocampo, and J. G. Baldwin, "A phylogenetic framework for root lesion nematodes of the genus *Pratylenchus* (Nematoda): evidence from 18S and D2-D3 expansion segments of 28S ribosomal RNA genes and morphological characters," *Molecular Phylogenetics and Evolution*, vol. 48, no. 2, pp. 491–505, 2008.

[85] Z. Majd Taheri, Z. Tanha Maafi, S. A. Subbotin, E. Pourjam, and A. Eskandari, "Molecular and phylogenetic studies on Pratylenchidae from Iran with additional data on *Pratylenchus delattrei*, *Pratylenchoides alkani* and two unknown species of *Hirschmanniella* and *Pratylenchus*," *Nematology*, vol. 15, pp. 633–651, 2013.

[86] C. Xu, H. Xie, C. Zhao, S. Zhang, and X. Su, "Review of the genus *Scutylenchus Jairajpuri*, 1971 (Nematoda: Tylenchida), with description of *Scutylenchus dongtingensis* n. sp. from rhizosphere soil of grass in China," *Zootaxa*, vol. 3437, pp. 32–42.

[87] Y. Yang and S. Zhang, "Identification of *Hemicriconemoides gaddi* from the Rhizosphere of Longan," *Chinese Journal of Tropical Crops*, vol. 5, pp. 935–941, 2013.

[88] E. Van den Berg, S. A. Subbotin, and L. R. Tiedt, "Morphological and molecular characterisation of *Hemicycliophora lutosa* Loof & Heyns, 1969 and *H. typica* de Man, 1921 from South Africa (Nematoda: Hemicycliophoridae)," *Nematology*, vol. 12, no. 2, pp. 303–308, 2010.

[89] N. Vovlas, B. B. Landa, G. Liébanas, Z. A. Handoo, S. A. Subbotin, and P. Castillo, "Characterization of the cystoid nematode *Meloidoderita kirjanovae* (Nemata: Sphaeronematidae) from Southern Italy," *Journal of Nematology*, vol. 38, no. 3, pp. 376–382, 2006.

[90] J. E. Palomares-Rius, N. Vovlas, S. A. Subbotin et al., "Molecular and morphological characterisation of *Sphaeronema alni* Turkina & Chizhov, 1986 (Nematoda: Sphaeronematidae) from Spain compared with a topotype population from Russia," *Nematology*, vol. 12, no. 4, pp. 649–659, 2010.

[91] G. Liu, J. Chen, S. Xiao, D. Pan, and S. Zhang, "Intraspecific variability of *Tylenchulus semipenetrans* populations on citrus and Chinese fir," *Scientia Agricultura Sinica*, vol. 9, 2011.

[92] B. Y. Park, S. N. Park, J. K. Lee, and C. H. Bae, "Morphometric and genetic variability among *Tylenchulus semipenetrans* populations from citrus growing area in Korea," *Plant Pathology Journal*, vol. 25, no. 3, pp. 236–240, 2009.

[93] U. Gozel, B. J. Adams, K. B. Nguyen, R. N. Inserra, R. M. Giblin-Davis, and L. W. Duncan, "A phylogeny of *Belonolaimus* populations in Florida inferred from DNA sequences," *Nematropica*, vol. 36, no. 2, pp. 155–171, 2006.

[94] Z. A. Handoo, A. M. Skantar, and P. Mulrooney, "First report of the sting nematode *Belonolaimus longicaudatus* on soybean in Delaware," *Plant Disease*, vol. 94, no. 1, p. 133, 2010.

[95] G. R. Stirling, A. M. Stirling, R. M. Giblin-Davis et al., "Distribution of southern sting nematode, *Ibipora lolii* (Nematoda: Belonolaimidae), on turfgrass in Australia and its taxonomic relationship to other belonolaimids," *Nematology*, vol. 15, pp. 401–415, 2013.

[96] J. Parkinson, M. Mitreva, N. Hall, M. Blaxter, and J. P. McCarter, "400000 nematode ESTs on the Net," *Trends in Parasitology*, vol. 19, no. 7, pp. 283–286, 2003.

[97] D. D. Sui, N. Atibalentja, G. R. Noel, and L. L. Domier, "Genetic diversity of the rDNA locus among 27 populations and 8 races of *Heterodera glycines* from China, Japan, and the United States, as revealed by PCR-RFLP," *Journal of Nematology*. In press.

[98] B. Gao, R. Allen, T. Maier, E. L. Davis, T. J. Baum, and R. S. Hussey, "Identification of putative parasitism genes expressed in the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*," *Molecular Plant-Microbe Interactions*, vol. 14, no. 10, pp. 1247–1254, 2001.

[99] D. P. Puthoff, M. L. Ehrenfried, B. T. Vinyard, and M. L. Tucker, "GeneChip profiling of transcriptional responses to soybean cyst nematode, *Heterodera glycines*, colonization of soybean roots," *Journal of Experimental Botany*, vol. 58, no. 12, pp. 3407–3418, 2007.

[100] G. E. Múnera Uribe, W. Bert, A. R. Vierstraete, E. de la Peña, M. Moens, and W. Decraemer, "Burrowing nematodes from Colombia and their relationship with Radopholus similis populations, R. arabocoffeae and R. duriophilus," *Nematology*, vol. 12, pp. 619–629, 2010.

[101] J. Jacob, M. Mitreva, B. Vanholme, and G. Gheysen, "Exploring the transcriptome of the burrowing nematode *Radopholus similis*," *Molecular Genetics and Genomics*, vol. 280, no. 1, pp. 1–17, 2008.

[102] J. Li, D. Peng, and W. Huang, "Phylogentic analysis of *Radopholus similis* from D2 and D3 fragments of the 28S rRNA gene sequences," *Journal of Huazhong Agricultural University*, vol. 5, 2008.

[103] S. T. Nyaku, V. R. Sripathi, R. V. Kantety, Y. Q. Gu, K. Lawrence, and G. C. Sharma, "Characterization of the two intra-individual sequence variants in the 18S rRNA gene in the plant parasitic nematode, *Rotylenchulus reniformis*," *PLoS ONE*, vol. 8, no. 4, Article ID E60891, 2013.

[104] Y. Zhan, A. Matafeo, H. Shi, and J. Zheng, "Morphological and molecular characterization and host range of *Rotylenchulus reniformis* population occurring in Hangzhou, Zhejiang, China," *Acta Phytopathologica Sinica*, vol. 41, no. 1, pp. 37–43, 2011.

[105] Y. Zeng, R. M. Giblin-Davis, and W. Ye, "Two new species of *Schistonchus* (Nematoda: Aphelenchoididae) associated with *Ficus hispida* in China," *Nematology*, vol. 9, no. 2, pp. 169–187, 2007.

[106] E. Van Den Berg, S. A. Subbotin, Z. A. Handoo, and L. R. Tiedt, "*Hirschmanniella kwazuna* sp. n. from South Africa with notes

on a new record of *H. spinicaudata* (Schuurmans Stekhoven, 1944) Luc & goodey, 1964 (nematoda: Pratylenchidae) and on the molecular phylogeny of *Hirschmanniella* Luc & goodey, 1964," *Nematology*, vol. 11, no. 4, pp. 523–540, 2009.

[107] C. D. Giorgi, P. Veronico, F. D. Luca, A. Natilla, C. Lanave, and G. Pesole, "Structural and evolutionary analysis of the ribosomal genes of the parasitic nematode *Meloidogyne artiellia* suggests its ancient origin," *Molecular and Biochemical Parasitology*, vol. 124, no. 1-2, pp. 91–94, 2002.

[108] G. Anthoine and D. Mugniéry, "Variability of the ITS rDNA and identification of *Nacobbus aberrans* (Thorne, 1935) Thorne & Allen, 1944 (Nematoda: Pratylenchidae) by rDNA amplification," *Nematology*, vol. 7, no. 4, pp. 503–516, 2005.

[109] L. Al-Banna, V. Williamson, and S. L. Gardner, "Phylogenetic analysis of nematodes of the genus *Pratylenchus* using nuclear 26S rDNA," *Molecular Phylogenetics and Evolution*, vol. 7, no. 1, pp. 94–102, 1997.

[110] X. Li and J. Zheng, "Identification of four *Pratylenchus* species based on morphology and PCR-RFLP of rDNA-ITS," *Acta Phytopathologica Sinica*, vol. 43, no. 4, pp. 444–448, 2013.

[111] M. T. Britton, C. A. Leslie, G. H. McGranahan, and A. M. Dandekar, "Functional genomic analysis of walnut-nematode interactions," Walnut Research Reports 2009, California Walnut Board, 2010.

[112] J. C. Wang, G. M. Huang, Y. D. Wei et al., "Phylogenetic analysis of *Pratylenchus* (Nematoda: Pratylenchidae based on ribosomal internal transcribed spacers (ITS) and D2/D3 expansion segments of 28S rRNA gene," *Acta Zootaxonomica Sinica*, vol. 4, pp. 687–693, 2012.

[113] D. Y. Chen, H. F. Ni, and T. T. Tsay, "Identification of a new recorded stunt nematode *Tylenchorhynchus zeae* (Nematoda: Belonolaimidae) in Taiwan," *Plant Pathology Bulletin*, vol. 16, no. 2, pp. 79–86, 2007.

[114] J. Liu, R. E. Berry, and A. F. Moldenke, "Phylogenetic relationships of entomopathogenic nematodes (Heterorhabditidae and Steinernematidae) inferred from partial 18S rRNA gene sequences," *Journal of Invertebrate Pathology*, vol. 69, no. 3, pp. 246–252, 1997.

[115] M. V. Blanco, P. Lax, J. C. Rondan Dueñas, C. N. Gardenal, and M. E. Douc, "Morphological and molecular characterization of the entomoparasitic nematode *Hammerschmidtiella diesingi* (Nematoda, Oxyurida, Thelastomatidae)," *Acta Parasitologica*, vol. 57, no. 3, pp. 302–310, 2012.

[116] S. A. Nadler, E. Bolotin, and S. P. Stock, "Phylogenetic relationships of *Steinernema* Travassos, 1927 (Nematoda: Cephalobina: Steinernematidae) based on nuclear, mitochondrial and morphological data," *Systematic Parasitology*, vol. 63, no. 3, pp. 161–181, 2006.

[117] O. Douda, M. Zouhar, E. Nováková, J. Mazáková, and P. Ryšánek, "Variability of D2/D3 segment sequences of several populations and pathotypes of potato cyst nematodes (*Globodera rostochiensis*, *Globodera pallida*)," *Plant Protection Science*, vol. 46, no. 4, pp. 171–180, 2010.

[118] S. P. Stock, J. F. Campbell, and S. A. Nadler, "Phylogeny of *Steinernema travassos*, 1927 (Cephalobina: Steinernematidae) inferred from ribosomal DNA sequences and morphological characters," *Journal of Parasitology*, vol. 87, no. 4, pp. 877–889, 2001.

[119] E. E. Morris, R. M. Kepler, S. J. Long, D. W. Williams, and A. E. Hajek, "Phylogenetic analysis of *Deladenus* nematodes parasitizing northeastern North American *Sirex* species," *Journal of Invertebrate Pathology*, vol. 113, pp. 177–183, 2013.

[120] Q. Yu, P. de Groot, I. Leal, C. Davis, W. Ye, and B. Foord, "Characterization of *Deladenus siricidicola* (Tylenchida: Neotylenchidae) ssociated with Sirex noctilio (Hymenoptera: Siricidae) in Canada," *Intemational Joumal of Nematology*, vol. 19, no. 1, pp. 23–32, 2009.

[121] K. Zhang, H. Liu, J. Sun et al., "Molecular phytogeny of geographical isolates of *Bursaphelenchus xylophilus*: implications on the origin and spread of this species in China and worldwide," *Journal of Nematology*, vol. 40, no. 2, pp. 127–137, 2008.

[122] R. B. Gasser and H. Hoste, "Genetic markers for closely-related parasitic nematodes," *Molecular and Cellular Probes*, vol. 9, no. 5, pp. 315–319, 1995.

[123] T. O. Powers, T. C. Todd, A. M. Burnell et al., "The rDNA internal transcribed spacer region as a taxonomic marker for nematodes," *Journal of Nematology*, vol. 29, no. 4, pp. 441–450, 1997.

[124] R. Floyd, E. Abebe, A. Papert, and M. Blaxter, "Molecular barcodes for soil nematode identification," *Molecular Ecology*, vol. 11, no. 4, pp. 839–850, 2002.

[125] A. Eyualem and M. Blaxter, "Comparison of biological, molecular, and morphological methods of species identification in a set of cultured *Panagrolaimus isolates*," *Journal of Nematology*, vol. 35, no. 1, pp. 119–128, 2003.

[126] P. De Ley, I. T. De Ley, K. Morris et al., "An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding," *Philosophical Transactions of the Royal Society B*, vol. 360, no. 1462, pp. 1945–1958, 2005.

[127] N. R. R. Da Silva, M. C. Da Silva, V. F. Genevois et al., "Marine nematode taxonomy in the age of DNA: the present and future of molecular tools to assess their biodiversity," *Nematology*, vol. 12, no. 5, pp. 661–672, 2010.

[128] M. A. Smith, D. M. Wood, D. H. Janzen, W. Hallwachs, and P. D. N. Hebert, "DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 12, pp. 4967–4972, 2007.

[129] H. van Megen, S. van den Elsen, M. Holterman et al., "A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences," *Nematology*, vol. 11, no. 6, pp. 927–950, 2009.

[130] A. Yu. Ryss, "Phylogeny of the order Tylenchida (Nematoda)," *Russian Journal of Nematology*, vol. 1, no. 2, pp. 74–95, 1993.

[131] L. K. Carta, A. M. Skantar, and Z. A. Handoo, "Molecular rDNA phylogeny of telotylenchidae siddiqi, 1960 and evaluation of tail termini," *Journal of Nematology*, vol. 42, no. 4, pp. 359–369, 2010.

[132] R. A. Bedding, "Controlling the pine-killing woodwasp, *Sirex noctilio*, with nematodes," in *Use of Microbes For Control of Invasive Arthropods*, A. E. Hajek, T. R. Glare, and M. O'Callaghan, Eds., pp. 213–235, Springer, Amsterdam, The Netherlands, 2009.

[133] A. R. Maggenti, M. Luc, D. J. Raski, R. Fortuner, and E. Geraert, "A reappraisal of Tylenchina (Nemata). 2. Classification of the suborder Tylenchina (Nemata: Diplogasteria)," *Revue De Nématologie*, vol. 10, no. 2, pp. 135–142, 1987.

[134] V. N. Chizhov and S. N. Kruchina, "Phylogeny of the nematode order Tylenchida (Nematoda)," *Zoologichesky Zhurnal*, vol. 67, no. 9, pp. 1282–1293, 1988.

[135] M. Luc, A. R. Maggenti, R. Fortuner, D. J. Raski, and E. Geraert, "A Reappraisal of Tylenchina (Nemata) 1. For a New Approach to the Taxonomy of Tylenchina," *Revue De Nématologie*, vol. 10, no. 2, pp. 127–134, 1987.

[136] O. V. Slobodyanyuk, “Host specificity in tylenchids of fleas,” in *Abstracts of Papers Presented at the Russian Society of Nematologists 1st English Language International Symposium*, pp. 102–103, Zoological Institute of the Russian Academy of Sciences, 1995.

[137] K. D. Elsey, “Parasitism of some economically important species of chrysomelidae by nematodes of the genus howardula,” *Journal of Invertebrate Pathology*, vol. 29, no. 3, pp. 384–385, 1977.

[138] G. O. Poinar Jr., J. Jaenike, and D. D. Shoemaker, “*Howardula neocosmis* sp.n. parasitizing North American *Drosophila* (Diptera: Drosophilidae) with a listing of the species of Howardula Cobb, 1921 (Tylenchida: Allantonematidae),” *Fundamental and Applied Nematology*, vol. 21, no. 5, pp. 547–552, 1998.

[139] R. L. H. Dennis, L. Dapporto, S. Fattorini, and L. M. Cook, “The generalism-specialism debate: the role of generalists in the life and death of species,” *Biological Journal of the Linnean Society*, vol. 104, pp. 725–737, 2011.

[140] C. E. Richmond, D. L. Breitburg, and K. A. Rose, “The role of environmental generalist species in ecosystem function,” *Ecological Modelling*, vol. 188, no. 2-4, pp. 279–295, 2005.

[141] M. E. J. Woolhouse, L. H. Taylor, and D. T. Haydon, “Population biology of multihost pathogens,” *Science*, vol. 292, no. 5519, pp. 1109–1112, 2001.

[142] H. D. Loxdale, G. Lushai, and J. A. Harvey, “The evolutionary improbability of “generalism” in nature, with special reference to insects,” *Biological Journal of the Linnean Society*, vol. 103, no. 1, pp. 1–18, 2011.

[143] J. C. Castillo, S. E. Reynolds, and I. Eleftherianos, “Insect immune responses to nematode parasites,” *Trends in Parasitology*, vol. 27, no. 12, pp. 537–547, 2011.

[144] R. Kassen, “The experimental evolution of specialists, generalists, and the maintenance of diversity,” *Journal of Evolutionary Biology*, vol. 15, no. 2, pp. 173–190, 2002.

[145] N. I. Samia, K. L. Kausrud, H. Heesterbeek et al., “Dynamics of the plague-wildlife-human system in Central Asia are controlled by two epidemiological thresholds,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 35, pp. 14527–14532, 2011.

[146] H. E. Kaiser and A. J. Boucot, “Specialisation and extinction: cope’s law revisited,” *Historical Biology*, vol. 11, no. 1–4, pp. 247–265, 1996.

[147] C. Darby, A. Chakraborti, S. M. Politz, C. C. Daniels, L. Tan, and K. Drace, “*Caenorhabditis elegans* mutants resistant to attachment of *Yersinia* biofilms,” *Genetics*, vol. 176, no. 1, pp. 221–230, 2007.

[148] G. A. Eroshenko, N. A. Vidyaeva, and V. V. Kutyrev, “Comparative analysis of biofilm formation by main and nonmain subspecies *Yersinia pestis* strains,” *FEMS Immunology and Medical Microbiology*, vol. 59, no. 3, pp. 513–520, 2010.

[149] G. A. Eroshenko, N. A. Vidyaeva, L. M. Kukleva et al., “Studies of biofilm formation in non-pigmented and plasmid-deprived mutants of *Yersinia pestis* on biotic surfaces, in vivo and in vitro conditions,” *Problems of Particularliy Dangerous Infections*, vol. 113, pp. 45–49, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 642089, 22 pages
http://dx.doi.org/10.1155/2014/642089

Research Article

Reconciliation of Gene and Species Trees

L. Y. Rusin,[1,2] E. V. Lyubetskaya,[1] K. Y. Gorbunov,[1] and V. A. Lyubetsky[1]

[1] *Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Bolshoy Karetny Pereulok 19, Moscow 127994, Russia*
[2] *Faculty of Biology, Moscow State University, Leninskie Gory 1-12, Moscow 119234, Russia*

Correspondence should be addressed to V. A. Lyubetsky; lyubetsk@iitp.ru

Received 11 August 2013; Accepted 27 November 2013; Published 27 March 2014

Academic Editor: William H. Piel

The first part of the paper briefly overviews the problem of gene and species trees reconciliation with the focus on defining and algorithmic construction of the evolutionary scenario. Basic ideas are discussed for the aspects of mapping definitions, costs of the mapping and evolutionary scenario, imposing time scales on a scenario, incorporating horizontal gene transfers, binarization and reconciliation of polytomous trees, and construction of species trees and scenarios. The review does not intend to cover the vast diversity of literature published on these subjects. Instead, the authors strived to overview the problem of the evolutionary scenario as a central concept in many areas of evolutionary research. The second part provides detailed mathematical proofs for the solutions of two problems: (i) inferring a gene evolution along a species tree accounting for various types of evolutionary events and (ii) trees reconciliation into a single species tree when only gene duplications and losses are allowed. All proposed algorithms have a cubic time complexity and are mathematically proved to find exact solutions. Solving algorithms for problem (ii) can be naturally extended to incorporate horizontal transfers, other evolutionary events, and time scales on the species tree.

1. Reconciliation of Gene and Species Trees: A Brief Overview

This section of the paper does not intend to cover the vast diversity of published literature on the problem of trees reconciliation. Instead, the authors strived to overview the problem of defining algorithmic construction of the evolutionary scenario as a central concept in many areas of evolutionary research. Important definitions are discussed, and essential problems are highlighted. We believe that, despite many approaches to defining the scenario known today, its solid theoretical framework is still to be developed.

1.1. Evolutionary Scenarios and Fields of Their Application. The evolution of the genome, apart from the mutation process, is an entangled complex of individual and concerted evolutions of genes, their regulations, gene content and arrangement on chromosomes, genetic flows between the genome and intracellular organelles, and so forth. Their evolutionary histories often do not coincide with each other and with patterns of speciation giving the rise to a variety of evolutionary events, such as gene duplications, losses, gains, horizontal transfers, chromosome rearrangements, and others. These phenomena play a pivotal role in evolutionary plasticity of the genome, the emergence of genes and gene families with novel functions, maintenance of the molecular machinery of the cell, evolutionary adaptation of the organism, and so forth. As known today, various types of horizontal transfers were the key force to drive the evolution of prokaryotes [1–3], while duplications of genes, partial or entire genomes, and mass gene loss events formed the genotypes of many higher eukaryotes, including higher plants [4–6] and vertebrates [7–10]. The genomic change fixed in generations over time ultimately shapes the biological diversity.

Important information contained in the discrepancies between these evolutions can be extracted and studied with the methods of trees reconciliation. Knowledge of ancestral genomic events provides efficient instruments in a range of fields, like establishing orthology/paralogy relationships between gene families [11–14], functional gene annotations [15–18], reconstruction of ancestral genes and genomes and their dating [19, 20], accurate reconstruction of gene and species trees [18, 21–27], construction of phylogenies based on whole genome data [22, 23], event-based reconstruction of coevolution [28] and its applications in ecology and biogeography [29–31], phylogenetic approaches to predict protein interactions [32], and so forth. A particularly intriguing problem is the *coevolution* of species, genes, and their regulatory systems, including binding sites, protein and RNA factors, DNA and RNA secondary structures, and RNA triplexes, which is poorly understood even in its statement. Further research in this area will shed more light on understanding the principles of concerted evolution at various levels.

In complex studies of coevolution it is vital to develop reconciliation approaches that account for as many various evolution events as possible. Not only inferring the events per se but also their mutual arrangement in time is important. Such an arrangement is called the *evolutionary scenario*. An overview of approaches to define and construct the scenario with trees reconciliation is the scope of this section.

In earlier works, scenarios accounted for gene duplications and losses only [33–36], some later—for only transfers and losses [37–41]. Such incomplete scenarios are useful in certain cases, for example, in studies of the Metazoa where transfers are very scarce, with low-copied or functionally nonredundant gene families or under low rates of duplications and losses [42, 43]. Timing the species tree and, particularly, imposing time scales (slices) are used in recent models to incorporate horizontal transfers [44–46]. The problem of defining and constructing the evolutionary scenario in its broad sense is actively studied, although its full definition is by far not yet obtained. Section 2 of this paper contains some original results obtained on these problems.

Approaches to substitute the species or even gene trees with a forest or net (graph or hypergraph) and to identify their areas that cannot be described by a tree are important but remain poorly studied [47].

The accuracy of reconciliation methods depends on the quality of initial phylogenetic data, usually gene trees, and multiple alignments in selected cases. The traditional steps of building gene trees (constructing multiple alignments, the choice and configuration of inference methods, robustness verification, etc.) can be nontrivial, especially for the automated generation of phylogenies on genomic scales. These methods are ever developing and are not discussed here. Some approaches are proposed or overviewed, for example, in [48–52], with extensive further referencing provided therein.

To mention is the group of methods that does not rely on gene trees to construct the scenario. Instead, genetic data in extant species of the given species tree is used to reconstruct the same type of data at internal nodes. In [53, 54] the authors addressed the problem of constructing parsimonious scenarios for individual sets of orthologous genes on a fixed species tree. Duplication events are not considered, and a horizontal transfer is not scored separately from an *ab novo* gene gain.

An extensive corpus of studies is devoted to the reconstruction of ancestral molecular characters and properties, rather than inferring discrete evolutionary events on the species tree. Such can be ancestral sequences, their lengths, primary and secondary regulatory structures, the tree areas with potential genetic transfers, and so forth [55–59]. Deterministic and probabilistic models (in particular, the Gibbs field approach) to reconstruct ancestral sequences and secondary structures are discussed in [60–62] presents a dedicated web service. These works remain out of the scope, as do studies of the mutation process and various reconciliation applications. The reader is referred to the original cited works for further details.

1.2. Reconciling Gene and Species Trees: The Classic "Embedding α" as the Basis of Other Mappings and Mapping Costs.

Earlier scenarios accounted only for duplications and losses. Such is the classic definition of mapping α, usually referred to as the "embedding α." In [33, 63] it maps vertices of a gene tree into vertices of a species tree. Namely, each vertex g of a gene tree G is assigned a vertex $\alpha(g)$ of the species tree S that corresponds to the last common ancestor of the species containing the leaves descendants of g. Mapping α explicitly infers duplications and implicitly losses.

Define edges of tree S as *tubes* to distinguish between edges of S and G. Each root is supplied with an additional *root edge* (or root tube), which ends in a *superroot*; that is, the superroot is the only vertex with the single child. Henceforth, all trees are described as directed downwards from the root.

Consider another definition of the "embedding α." Define mapping f as a mapping of vertices in the gene tree G into vertices or tubes (often both) in species tree S that satisfies the conditions: the leaves in G map into leaves in S having the same species notations; the superroot of G maps into the root tube in S; mapping f preserves the *natural order* relation on G and S; which is defined on any tree by the branching order downwards from the root (i.e., this relation keeps the succession of lineages). Additional less determinative conditions are formulated in Sections 2.3 and 2.7 ([45, 64]). In this paper, most definitions are provided in Section 2 and the reader is expected to be acquainted with general terminology used throughout the text.

Definition of f continued. The total sum of duplications and losses (the "embedding cost") has the minimal value on α among all costs of possible mappings f of gene tree G into species tree S. The embedding cost of mapping α is denoted $c(\alpha)$; the analogous cost of f is denoted $c(f)$; that is, $c(\alpha) = \min\{c(f) \mid f\}$ (where f is a variable). In other words, $c(\alpha)$ and $c(f)$ are sums of the amounts of gluings and gaps in mappings α and f, respectively; these numbers can be weighted according to the costs of corresponding event types (in this case, duplications and losses). Thus, mapping α can be defined as a global minimum of the embedding cost functional $c(G, S, f) = c(f)$, where variable f runs over all mappings of G into S. Note that the list of event types and the localization of evolutionary events are defined on the species tree individually for each mapping f (refer to definitions in Section 2.4).

Algorithmically, mapping α is built by induction from leaves toward the root in linear computing time, and its cost is computed simultaneously [65, 66].

Study [45] describes a similar definition of mapping α, and a different construction algorithm is applied from the root toward the leaves. It is a useful definition in terms of its extensibility to scenarios with gene horizontal transfers and gains. The presented algorithm simultaneously computes the mapping and its cost.

In [36] all possible reconciliations of gene tree G and species tree S are considered, that is, all possible mappings f of G into S. This approach is further developed in [43], where f maps each vertex in G into a vertex or tube in S, thus inferring the speciation (if $f(g)$ is a vertex) and duplication (if $f(g)$ is a tube) events, respectively. An algorithm described in [43] generates a random reconciliation of G and S, enumerates all such possible reconciliations, and calculates exactly the minimal number of fixed operations needed to rearrange one reconciliation into the other.

Let only duplications and losses be considered, and let G be a binary gene tree with a predefined set of "reliable" edges. To find is a tree G' with the same set of leaves and containing all clades induced by reliable edges such that G' minimizes the embedding cost of its mapping α into a given binary species tree S. Algorithms to solve this problem are described in [35, 67]; in [35] the algorithm is proved to find exactly the optimal gene tree G' in cubic time, while [67] offers a heuristic solving algorithm. Similarly, in [68] duplications, losses and transfers are accounted for to find a gene tree G' such that it contains a predefined set of reliable edges (i.e., the induced clades) from G and minimizes the embedding cost of any mapping f of G' into a given binary species tree S. A heuristic solving algorithm is proposed.

An approach to reconcile gene and species trees based on information about synteny of corresponding genes in the genome is proposed in [69]. An algorithm is described to build a forest of trees that reflect the evolution of pairs of neighboring genes by minimizing the embedding cost of gains and losses of the gene pairs. Computing time of this algorithm has the order n^2k^2, where n is the number of gene trees and k is their maximal size.

1.3. The Binarization Problem for Fixed Gene and Species Trees.

The algorithm described in [70] has a linear time complexity, and, given a polytomous gene tree G, binary species tree S, and their mapping α, searches for a binarization G^* of G by first minimizing the total sum of duplications and then the total sum of losses in the obtained set of binarizations.

Study [71] describes a linear time algorithm to binarize the tree G against the tree S using mapping α, provided that only duplications and losses are allowed. A binary resolution G' of a polytomous G is constructed such that the resulting binarized gene tree G^* optimally reconciles with the species tree S; that is, it has the minimal embedding cost compared to other binarizations. Importantly, the algorithm is mathematically proved to find the global minimum of the embedding cost functional $c(G', S, \alpha)$ (G' is a variable). The authors of [71] reference the history of the binarization problem for the case of duplications and losses under fixed G and S.

Study [25] uses a similar minimization criterion for $\sum_j c(G'_j, S, f_j)$ to binarize many polytomous gene trees G_j against a binary species tree S when horizontal gene transfers are allowed, and the variable f_j is an arbitrary mapping (refer to Sections 2.7 and 2.12). In [25] the algorithm is proved to find the globally minimal binarization and possess the complexity determined as follows: if k is a maximal degree of polytomy among all vertices in G_j, then the computing time has the order of the product of the total number of vertices in initial trees G_j and S and coefficient 2^{2k}.

In [70] it is proved that the optimal binarization problem is NP-complete for the case of a polytomous species tree even if a gene tree is binary. However, heuristics is proposed to handle even nonbinary gene trees. In [72] another heuristic algorithm is proposed to solve the same problem, nevertheless requiring a binary gene tree.

The algorithm described in [73] computes all possible binarizations S' of a polytomous species tree S in order to find such S^* that minimizes the embedding cost for an input fixed binary gene tree G against the variable S'. In this search all event types are considered, including transfers; and the variable f_j is an arbitrary mapping. A new condition is imposed: let a vertex g in a gene tree be mapped into a vertex s in the species tree; then both child clades of s contain at least one species from the clade of g. The computing time of the algorithm is the product of a polynomial of degree 4 (a function of the number of leaves in the input data) and an exponential functional that depends on the maximal degree of polytomy in the species tree.

Sections 2.12 and 2.13 present an essentially different statement of the binarization problem (refer also to [25]).

1.4. Evolutionary Scenarios with Horizontal Transfers: Coevolution of Genes and Their Regulation Systems on a Species Tree.

Accounting for gene horizontal transfers in evolutionary models is vital for understanding the evolution of many life forms, especially prokaryotes [1–3]. It also provides efficient tools to study the evolution of molecular systems, establishing orthology/paralogy relationship between gene families [11–14], and so forth. In [74] the authors give a broad view of the perspectives to reconstruct the Tree of Life within the general framework of genome evolution, the role of gene horizontal transfers, duplications and losses in the emergence of new molecular functions, and evolutionary adaptation.

With only duplication events allowed, for a given set of binary gene trees G_j and a binary species tree S, consider any mapping f_j of G_j into S. In the approach in [75], for each G_j a duplication event $\alpha(g)$ is attempted closer to the root of a species tree but below $\alpha(g')$, where g' is the parent of g (if g is the root, $\alpha(g)$ is attempted closer to the root). A functional is proposed that depends on $\{f_j\}$ and equals the sum (over all vertices s in S) of maximal heights of subtrees in all G_j (not only those that reach the leaves) mapped by f_j into a vertex s. The desired are mappings f_j that minimize this functional. A linear complexity proved algorithm to find this global minimum is proposed. Historical references to this approach are provided in [75] and in review [76].

Event-based approaches to study coevolution of various elements are discussed in [28], and their applications in ecology and biogeography are discussed in [29–31]. For example, in [77, and unpublished materials] the authors present a model and an effective algorithm to reconstruct coevolution of genes and their regulatory systems (binding sites, protein and RNA factors, DNA and RNA secondary structures, RNA triplexes, etc.) under horizontal transfers and other events allowed on a species tree. A general coevolutionary scenario was constructed based on a universal functional that combines requirements specific for individual scenarios of the co-evolving elements. Evolutionary events inferred in individual scenarios within the general coevolutionary scenario appear to be biologically consistent (coordinated with each other). Inferring coevolutions is an important and complex problem, which we do not almost discuss in this paper.

1.5. Time Slices on the Species Tree as an Approach to Accounting for Horizontal Transfers. When horizontal transfers are included in the model, a gene cannot transfer between two tubes located anywhere on the species tree S, a transfer is possible only between the "contemporaries." To correctly describe transfers, the tree S must be partitioned in *time slices*, for example, by dating its tubes or vertices. An approach to do so is presented in [44], where each tube is associated with a time interval, and a transfer between tubes is allowed if their intervals have a non-empty intersection. A corresponding construction algorithm is described in [44], without the complexity assessment. A very complicated original description of the algorithm does not allow us to provide detailed comments.

Assume that to correctly define a transfer in time is to allow it to occur exactly within one time slice (a set of predefined time slices on the species tree S is to be fixed). If the correctness condition is not imposed but transfers are allowed, the fastest algorithm constructs a scenario in time of the order mn, where m and n are numbers of leaves in the input gene and species trees [78]. Finding a scenario defined correctly in time is an appealing challenge and requires an intricate imposition of time slices on the tree. Constructing the slices is a difficult problem of its own already at the level of definition. An algorithm with complexity n^3 that solves it is proposed in [45], albeit without a proper biological justification.

An approach to construct a time-correct scenario is finely elaborated in [45] and, independently, in [46]. The constructing algorithm in [46] uses a prefixed set of time slices and does not consider (similarly to [44]) the common case of a gene transfer with loss of the donor copy. The authors prove the polynomial time complexity of their algorithm, however not providing an exact assessment of the polynomial degree. In [45] the algorithm accounts for all types of transfers and differs conceptually from those proposed in [44, 46]; it is proved to have the complexity of mh, where h is the number of vertices in a species tree with preimposed time slices and is proved to find the exact global minimum (under certain conditions). The proof is given in [79].

In [80] the following condition (below referred to as the "tofig-condition") on mapping f is formulated. Assume there exists a linear order $<_T$ at vertices of the species tree S, for which: for any tube (u, v) in S the inequality $u <_T v$ is valid, and if for two edges, (u, v) and (u', v'), in a gene tree G one precedes the other in terms of the natural order on edges, then the upper terminus a of the tube that "contains" $f(u)$ (i.e., $f(u)$ is this same tube or its lower terminus) must "precede" (in the sense of $a <_T b$) the lower terminus b of the tube that contains $f(v')$. Under this condition the problem of finding a globally minimal scenario f is NP-complete [81, 82]. Strengthening this condition may simplify the situation. For example, let each time slice on S consist of tubes equidistant from the root, and let, as mentioned above, horizontal transfers be permitted only within the common slice. Such the condition on f implies the tofig-condition if $<_T$ is a width-first linear order. The problem of finding a globally minimal scenario f under the above-mentioned strong condition becomes polynomial in time [45].

The notion of the evolutionary scenario, specifically for a pair $\langle$gene tree G, species tree $S\rangle$, is very important in mathematic aspects of the theory of evolution. A realistic scenario is such that accounts for as many different types of gene evolutionary events as possible, including various types of horizontal transfers.

Analogously to mapping α, a candidate mapping (scenario) f is defined at vertices of the gene tree G, with its values being the vertices or tubes (often both) of the species tree S, such that f keeps the natural orders (the successions of lineages on the trees). Each mapping f defines its own set of evolutionary events (exact definitions are provided in Sections 2.3 and 2.7). As in mapping α, each event type is assigned a cost. Analogously to the embedding cost $c(\alpha)$ of mapping α, the cost $c(f)$ of a candidate mapping f is the sum of event costs defined by f, which may be weighted according to the reliability of corresponding vertices and the type of event. The problem is to find the mapping β (scenario β) that globally minimizes the total cost $c(f)$ under certain constrains, which almost always need to be imposed on its design.

The cost of the pair $\langle G, S\rangle$ is the cost of its minimal scenario β and is denoted $c(G, S)$. Therefore, one needs to minimize the functional $c(G, S, f)$ over all mappings f of G into S to obtain the desired scenario β and the value $c(G, S)$.

An alternative approach is to describe the scenario in terms of a stochastic process on a species tree. Selected relevant approaches are proposed in [25].

An algorithm to construct a scenario β for a gene tree G and a species tree S_0 derived by partitioning S in time slices, is described in [25, 45, 79]. The running time of the algorithm has the order of the product of m and the number of leaves in S_0. The algorithm, its proof, and all definitions from [45, 79] are reproduced in [83], where the algorithm was extensively tested on novel data. In [84] the same (as the authors perceive) algorithm is applied to different biological data.

The importance of taking into account suboptimal scenarios that can become optimal under slight variations of the costs of event types is demonstrated in [80]. An approach to deal with suboptimal scenarios is proposed in [25], where the authors also examine the case of gene gain using the outgroup approach (refer to the extended event list in Table 1 in [25]).

1.6. Constructing the Supertree. The definitions and algorithms of trees reconciliation and construction of the scenario (mapping) stated above can be applied to another long studied problem: given a set $\{G_j\}$ of gene trees, find the tree

S^*, for which the *total cost* $\sum_j c(G_j, S, f_j)$ of all events for pairs $\langle G_j, S^* \rangle$ reaches the global minimum (S and f_j are variables; usually $f_j = \alpha_j$, which gives the cost $\sum_j c(G_j, S)$). The tree S^* is called a *supertree*. In this statement, the supertree may be imposed certain constraints depending on the initial gene tree data that need to be taken into account when optimizing the total cost functional.

In the classic sense, the supertree is constructed with no constraints by merging input trees using a variety of heuristic methods based on various tree compatibility criteria. In distance approaches, the supertree is found by minimizing the average distance between it and all input trees. Defining a proper distance is therefore of importance. In the framework of trees reconciliation, this problem reduces to the minimization of the functional defined via the total cost of evolutionary events for trees G_j over all j. The classic and still commonly used distance was introduced in [33] as the total cost of duplications and losses (and transfers, if allowed).

In some approaches, the supertree construction step is preceded by filtering out leaves, subtrees, or entire gene trees that do not satisfy certain reliability conditions [85]. The discarded elements can later be used to detect areas of "active" evolution on the supertree. We do not discuss such kind of approaches here.

The problem of building the supertree is NP-complete if *no constraints* are imposed on the desired tree S^*, even when only duplications and losses are allowed [86]. This stimulated the development of heuristic methods and attempts to reformulate the problem itself.

Heuristic Approaches. Among such is the quartet method that consists of two phases. At the first phase, trees are built for all quartets of species; here the choice of the reliability function to assess quartet topologies plays the important role; refer, for example, to [87]. At the second phase, the supertree is built by optimally reconciling the quartet species trees using a heuristics. In different implementations of the second phase, the supertree is constructed either "from root to leaves" [88] or "from leaves to root" [89]. The method produces an unrooted tree.

Rooted supertrees are produced by the triplet method, an analog of the quartet method, where the final tree is obtained by assembling triplet trees also using heuristics, for example, as described in Phase 2 of the supertree building algorithm from [25]; refer to Section 2.10 below.

Other methods use heuristics to maximize the functional of clades matching among two trees (rooted supertrees are produced) [90] or use a matrix representation of multiple trees [91]. A simple method to root species trees is proposed in [25, Suppl. 1].

Out of the scope of this paper remain other approaches to infer a species tree, such as the supermatrix strategies, which are popularly used in many phylogenetic studies of particular groups as well as larger taxa. In the supermatrix design, sets of orthologous genes sampled across the compared species are aligned, concatenated into a "superalignment" (supermatrix) and processed for computing one tree. In so doing, this method combines partially overlapping species samplings in the input orthologous sets to accommodate all species in one tree. Although the supermatrix approach relies on the well-established methodology of inferring gene trees, there exist many pitfalls that limit its application to larger analyses on a genomic scale. Among them are the strict requirement on orthology, missing data in sparse supermatrices, and different modes of evolution exhibited by different supermatrix partitions (often exacerbated by disparities in their size) and even by individual positions in the alignments, which requires the usage of sophisticated evolutionary models and causes inevitable computational burden that may become intractable with larger datasets [92–95].

In this context, fine selection of orthologs has received much attention as a problem of high relevance and arduous both ideologically and computationally. Approaches to this problem diverge into reconciliation-based (e.g., [11–14]) and graph clustering methods (e.g., [96–100]). The authors in [100] proposed a quadratic in time complexity clustering algorithm to construct orthologous protein families based on sequence similarity (and local synteny in certain cases). It was applied to mitochondrial, plastid, and some other (unpublished) genomic data. The obtained clusters well conform with known protein functional annotations, independently constructed orthologous groups, and other protein characteristics. The clustering revealed some lineage-specific proteins. Thus, mitochondria of the vine *Vitis vinifera* were found to encode proteins also typical for plastids, which implies that a horizontal genetic flow between these organelles had happened in the past [100].

Reformulation of the Problem. The development of novel reconciliation approaches and their effective solving algorithms with low (polynomial) complexity that are mathematically proved to find the global minimum (presumably the correct supertree) holds a good perspective. The algorithm originally developed by the authors [23, 64] introduces a condition that allows to effectively find the global minimum of the total cost functional. The condition constrains the desired supertree S^* to contain only clades from the input gene trees and certain combinations of them. Under this condition and if only duplications and losses are allowed, the algorithm is mathematically proved to find the global minimum of the cost functional in time cubic of the input data size [64]. Solving the same problem for the case of transfers is an important perspective. This approach is based on a different principle compared to other known methods.

1.7. Probabilistic Definitions of the Evolutionary Scenario: Evolution as a Stochastic Process and Coalescent Approaches.

The definition of the clade probability as a fraction of trees containing a given clade was introduced in [101]. The authors argue that the correct supertree commonly contains all clades from the initial tree set with the probability >1/3.

The species tree reconstruction under the assumption of numerous transfers is discussed in [102]. Using a probabilistic approach, it is shown that the species phylogeny is tree-like even with a high transfers content, that is, when their number linearly depends on the average number of leaves per tree. Conversely, in [24] it is mathematically proved that the triplet method recovers the correct supertree with high probability only if transfers are not many. Studies [24, 102] well reference this approach.

A stochastic procedure to construct a scenario with all types of events, including transfers, is proposed in [25]. The authors describe an algorithm to compute expectation values of the event numbers in each tube and over all tubes of the species tree. The proposed approach can also be used to determine other characteristics of the process.

In the first subsection below we briefly overview two groups of publications operating with quite sophisticated probabilistic approaches that need to be further discussed in terms of the probability theory. The second subsection is devoted to the coalescent theory.

1.7.1. Evolution as a Stochastic Process. A type of stochastic processes other than in [25] is considered in [103]. Fix a gene tree G and a species tree S, with tube lengths corresponding to times; paths from the root to each leaf have equal lengths. An oracle is fixed that assigns to each natural number n and tube d the probability of the outcome "d contains exactly n duplications." Here, a mapping f of vertices in G into vertices and tubes in S is defined under the condition: if for any child g_1 of g the inequality $\alpha(g) \neq \alpha(g_1)$ is valid, then $\alpha(g) = f(g)$ or $f(g)$ is a tube having $\alpha(g)$ as its lower terminus.

A probability of f is the probability of tube d to contain exactly $|\{x \in G \mid f(x) = d\}|$ duplications multiplied over all d in S. Recall that the sign $|\cdot|$ stands for the number of the set elements, the cardinality of the set. To find are (i) the highest likelihood among all possible mappings f (ii) the mapping f^* with the highest likelihood itself, and (iii) the numbers of duplications in each tube d under the mapping f^*. The authors describe a polynomial of degree 5 heuristic algorithm for (i) and exponential complexity algorithms that find exact solutions for each of the three tasks.

The following statement is considered in [104]. Fix G and S (the root tube d_0 is located upwards the root in S) with tube lengths corresponding to times and λ and μ being the intensities of duplications and losses, respectively. The intensities are constant across all tubes and are parameters of a linear death-birth process (its formal definition is provided at the end of this subsection).

For each vertex g in G denote $A(g)$ as the set $A(g) = \{f(g) \mid f\}$; that is, $A(g)$ contains vertices and tubes $f(g)$ from S for a variable mapping f and a fixed argument g. Call mappings f and h adjacent if the following conditions are valid: f and h differ at exactly one vertex g, $f(g)$ and $h(g)$ are comparable in S (in terms of the natural order on tree S), and there exist no elements from $A(g)$ strictly in-between them.

In [104] a tree G' is defined and generated by the below stochastic process; G' is then compared with the initial tree G. Let us first informally describe the stochastic process for G'. The root tube of S contains the start of gene lineages that descend downwards and bifurcate at each vertex of S (the divergence events). In each tube, each gene lineage undergoes duplications or losses with given intensities λ and μ. In case of a loss of the lineage terminates, in case of a duplication, it bifurcates into two descendent lineages in this tube. All lineages terminate in leaves, and only then the process ends to generate a tree G' (inside the tubes) and its natural mapping into S; all lineages terminated before leaves are discarded and not included in G'. The tree G' and its natural mapping into tree S are generated in any realization of this random process from the root toward leaves in S. More precisely, arrange slices by ascending order when the current total amount of lineages changes by 1. Let the root part of the tree G' be generated at instant t. If at that instant the number of lineages in a tube increases by 1, a lineage is chosen equiprobably in the tube and bifurcated; if it decreases by 1, this lineage terminates.

The probability $P(G, f)$ of mapping f is the probability that the random process generates a tree G' isomorphic to the tree G through mapping f. The probability $P(G)$ of tree G is the sum of probabilities of all its mappings in S; *a conditional probability* $P(f \mid G)$ is defined as $P(G, f)/P(G)$. By substituting $P(G)$ in the denominator with a sum over a given subset of mappings (defined K), we obtain the definition of a K-*approximated* conditional probability and denote it $P_K(f \mid G)$.

Define a graph, where all vertices are mappings f of G into S and edges connect adjacent vertices. Fix an arbitrary spanning tree T in the graph that is rooted by mapping α; and let K be a connected subgraph with k vertices in T. In [104] the authors prove the following: for all mappings f from K, the probability $P(f \mid G)$ is computed with the time and memory of $O(|G|^2|S| + k(|S| + |G|))$ and the K-approximation $P_K(f \mid G)$—with the time and memory of $O(|S||G| + k(|S| + |G|))$.

Experiments with biological data were performed to obtain realistic values of intensities λ and μ of duplications and losses.

A d-probability is the sum of conditional probabilities $P(f \mid G)$ of all mappings f from T, which are separated from the root α by maximum d edges; such mappings are called d-mappings. Computer simulations showed that, (i) with the increase of d (from 0), the d-probability soon reaches the plateau and (ii) for each mapping f before the plateau, the value $P_K(f \mid G)$ approximates $P(f \mid G)$ with high accuracy if K includes all mappings before the plateau.

An algorithm realizing the approach of [104] was developed and applied to biological data in [105]. Earlier related results are in [104, 105].

Probabilistic modeling of gene evolution can also be applied to model sequence divergence, as described in [106] and, with more detail, in [107]. A model and an algorithm are proposed in [107] to simultaneously infer gene trees, the species tree, and expectations of duplications and losses in each tube of the species tree, given a set of multiple alignments. Further relevant references are provided in [107].

A Formal Description of the above Described Process. Let P be a linear death-birth process applied to the tree S. The process argument is time τ taking on a value from 0 to τ_0, where τ_0 is the path length between the root and a leaf. At each vertex s define time $t(s)$ as the length of the path from the vertex to the root; tube $d = (s_1, s_2)$ (s_1 closer to the root) "contains the instant" τ if $t(s_1) \leq \tau \leq t(s_2)$. The value of $P(\tau)$ is a set of pairs: a tube d possessing instant τ and the number $d(\tau)$ of gene lineages in the tube at instant τ. The definition of $d(\tau)$ is as follows: $P(0)$ is a set consisting of the single pair $\langle$root tube $d_0, 1\rangle$; that is, $d_0(0) = 1$; let for a nonroot tube $d = (s_1, s_2)$ hold $\tau = t(s_1)$; then, by induction from root to leaves assume that $d(\tau)$ equals $d'(\tau)$, where d' is the parent tube for d; determine the change of $d(\tau)$ in a small time interval δt of the argument such that d contains $\tau + \delta t$. For each tube $d = (s_1, s_2)$ possessing instant τ, define conditional (transition) probabilities of the number $d(\tau + \delta t)$ of gene lineages at the instant $\tau + \delta t$ if at the previous value $n = d(\tau)$ is known:

$$\Pr\{d(\tau + \delta t) = n + 1 \mid d(\tau) = n\} = n\lambda\delta t + o(\delta t), \quad (1)$$

where λ is duplications intensity, and

$$\Pr\{d(\tau + \delta t) = n - 1 \mid d(\tau) = n\} = n\mu\delta t + o(\delta t), \quad (2)$$

where μ is losses intensity,

$$\Pr\{|d(\tau + \delta t) - n| > 1 \mid d(\tau) = n\} = o(\delta t). \quad (3)$$

The end of process definition.

1.7.2. Coalescent Approaches. In this group of studies, modeling the evolution of genes along a species tree includes a novel approach and an evolutionary event of novel type, the incomplete lineage sorting (refer to [108, 109] for the theory and references provided therein). Below we go with some detail into this important concept, which is grounded on the mathematical theory of the reverse time. On trees, the "direct time" refers to the time directed from the root to the leaves, and the "reverse time" reverses this direction.

The problem of reversing time in stochastic processes was first visited by Kolmogorov [110]. Kingman [111] had found that the probability distribution on phylogenies in large populations is described by a special type of random processes named coalescence and analyzed them in reverse time using the earlier model of population evolution proposed by Wright [112] and Fisher [113]. These works laid the foundation of the coalescence theory, and Kingman gave it further development and formulated it for continuous time.

The Central Idea of the Model in Short. Consider the sets of "parents" and "children," G_n and G_{n+1}, at nth and (n + 1)th generations, each consisting of N elements. Assume that a multivalued mapping D_n of G_n into G_{n+1} is surjective and satisfies the condition: the values of any two different elements from G_n are disjoint subsets of G_{n+1}. Such the mapping D_n is an inverse of the single-valued mapping F_n of G_{n+1} into G_n (informally, children are mapped into their parent). There is exactly N^N different mappings F_n, which are equiprobable, and thus each F_n has the probability N^{-N}. It is also assumed that for all generations G_n all F_n are independent random maps.

This simple description is equivalent to conventional definitions of the Wright-Fisher-Kingman model, which we describe below for the comparison.

The jth individual ("parent") from G_n produces ν_j individuals ("children") in generation G_{n+1} and dies; ν_j are supposed to be mutually independent (over j) and follow the Poisson distribution:

$$P\left\{\nu_j = r\right\} = e^{-\lambda}\frac{\lambda^r}{r!}. \quad (4)$$

Let a population contain N individuals at each generation G_n; that is, the condition $\sum_j \nu_j = N$ must be imposed to fix the population size over generations. The joint distribution of ν_j is multinomial:

$$\begin{aligned} P\left\{\nu_j = r_j, 1 \le j \le N\right\} &= \frac{\prod_j \left(e^{-\lambda} \cdot \left(\lambda^{r_j}/r_j!\right)\right)}{e^{-N\lambda} \cdot \left((N\lambda)^N/N!\right)} \\ &= \frac{\binom{N}{r_1,\ldots,r_N}}{N^N} = \frac{N!}{r_1! \cdots r_N! N^N}, \end{aligned} \quad (5)$$

where $\sum_j r_j = N$.

The map F_n can be interpreted as the random choice of a parent from G_n by each individual from G_{n+1}; this choice is equiprobable. The latter formula defines the probability of the event "the first parent has r_1 children, the next parent has r_2 children, and so on down to r_N".

The evolution in reverse time is a transition from G_{n+1} to G_n and deeper toward the root. The probability of any two individuals from G_{n+1} having different parents is $(1 - (1/N))$; and having different parents in s preceding generations (down to G_{n-s+1}) is $(1 - (1/N))^s$. The probability of k fixed different individuals from G_n having k different parents in one preceding generation is

$$P_k = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{k-1}{N}\right), \quad (6)$$

and having k different parents in s preceding generations (down to G_{n-s}) is P_k^s. Consider a limit of P_k^s with variable s and $N \to \infty$. The lifetime of one generation is assumed to be $\Delta t = 1/N$; that is, in time $t = s \cdot \Delta t = s/N$, one observes s generations, $s = Nt$. The probability of k fixed individuals having k different parents (in the limit under $N \to \infty$) over fixed time t (the lifetime of $[Nt]$ generations) is

$$\begin{aligned} \lim_{N \to \infty} P_k^{[Nt]} &= \lim_{N \to \infty} \left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{N}\right)^{Nt} \\ &= \exp\left(-\frac{k(k-1)}{2} t\right). \end{aligned} \quad (7)$$

A simple interpretation of the last formula: k individuals can form $k(k-1)/2$ pairs, the probability that any pair of individuals over time t does not share a common parent is $\exp(-t)$. In random time t (with the parameter k) a random pair of individuals is chosen and assigned a parent. Time t is defined by the random variable T distributed exponentially as $P\{T > t\} = \exp(-(k(k-1)/2)t)$. The choice of the pair is equiprobable because the probability of choosing a pair from $k(k-1)/2$ possible pairs is $(k(k-1)/2)^{-1}$. An unpaired individual is a parent of itself. The obtained parents are further paired with each other analogously until no further pairing is possible.

This process can be described as building a phylogenetic tree for given k individuals (leaves). In a next (inductive) tree level in the direction root ward, a pair is chosen from m current parent individuals and coupled under a new common node to form two new edges with the same length t equal to the value of random variable T distributed exponentially with the parameter m (at the start of induction $m = k$):

$$P\{T > t\} = \exp\left(-\frac{m(m-1)}{2}t\right). \quad (8)$$

An unpaired individual is projected on the next level and forms a new vertex and a new edge of length t connecting the two vertices. Thus, an unpaired individual is parental to itself.

This process is called the *coalescence* and ends with building a rooted tree with k leaves.

Other coalescences are considered in [109, 114]. Namely, fix k individuals at a time instant, with k_1 mutants and k_2 wild type, $k_1 + k_2 = k$. Assume that all mutants evolve from one parent that acquired a single mutation and all its descendants are mutants. Denote A, $|A| = k_1$ the extant population of mutants and by B, $|B| = k_2$ the population of extant wild types. The genealogy of A bearing the mutation is a subtree in the genealogy of the whole population, the union of A and B. The coalescence process is used to find a phylogenetic tree such that lineages of A coalesce earlier than any lineage from A forms a common parent with any lineage from B. A coalescence that satisfies this constraint is called *conditional* [114]. Algorithmically, the coalescence tree building process is running multiple times until the described tree containing the clade A is built.

Another constraint imposed on the tree building process is studied in detail in [109]. Namely, consider a species tree S with lengths (times) given for all tubes such that all paths from any node to the leaves are equal, thus defining the age of the node. Conditional coalescence can be applied to build a gene tree G along with its mapping into S, that is, the evolution of the gene inside the species tree [109]. At the start of induction, each gene leaf is assigned to its corresponding species leaf in the tree S. In *reverse* time, in the resulting tree any two gene leaves existing in two different species tubes form the common gene parent of at least the age of the common parent of the corresponding species. Thus, gene lineages may coalesce much later than their containing species. Such an event is called the *incomplete lineage sorting*.

The described process is applied to the case when the mutant is replaced with a duplicated copy of a gene that acquired a mutation after the duplication [108] had occurred. If a part of a population undergoes genetic change, it may result in the formation of subspecies. After the speciation event, the change is usually fixed in this subspecies. The duplicated copy of a gene survives, in contrast with the models like mapping α that operate with species as discrete units.

Study [108] also introduces the *interim locus tree* concept based on conditional coalescence. A gene tree is mapped into the interim locus tree, which then maps into the species tree. The species tree evolves in direct time, from root to leaves. However certain ideas in the description of this approach remain hard to understand.

2. Constructing the Evolutionary Scenario and the Supertree: Algorithms and Proofs

This section (Sections 2.1–2.13) of the article describes the original solutions and corresponding mathematical proofs proposed by K. Y. Gorbunov and V. A. Lyubetsky for the two problems in the field of trees reconciliation: inferring gene evolution along a species tree and trees reconciliation into a single tree (including the case of polytomous trees). These developments apply to a diverse and important subject of the evolution of species, genes, and their regulatory systems considered in concert or separately.

2.1. Statement of Two Problems. Studies [23, 25, 45, 64, 79] tackle two important and sophisticated problems in bioinformatics. The obtained results are partially reviewed in Section 1 of the paper, which also provides an extended biological background and relevant references.

The first problem is to reconstruct a gene evolution along a species tree or, in other words, to construct a mapping of a gene tree into a species tree and to build the scenario. *The second problem* is to reconcile a set of gene trees into one common species tree. A specific facet of the second problem is to build a supertree (by globally minimizing a suitable functional commonly referred to as the "cost") for the given set of trees. This problem is extended to the hard case of polytomous data, especially polytomous input trees.

In the above-mentioned works [23, 25, 45, 64, 79] only concise formulations are provided, while in this section we give mathematical statements and proofs to describe the two problems on the case when only gene duplication, loss, and divergence during speciation are the considered evolutionary events. Following on, we describe in detail the extension of the developed algorithms to incorporate other types of gene evolution events and/or the case of polytomous gene trees.

The first problem is solved in polynomial (often linear, at maximum cubic) time even for the case of incorporating time slices and horizontal gene transfers. In Section 2.7 it is proved that the corresponding original algorithm of cubic time complexity finds exactly the global minimum; that is, the model is exactly solvable.

In its traditional statement, the second problem cannot be solved algorithmically in polynomial time, as it is proved to be NP-complete. Known exponential solutions (based on various enumerators) are computationally too intensive, and do not guarantee that the optimal solution (the global minimum of a functional) is found if heuristics are applied to stop the search. Moreover, the accuracy of approximating the global minimum by a heuristic solution is not clear at all.

Complete proofs are first given for the case of no time slices and gene transfers. The discussion of the second problem follows next. A solving algorithm cubic of the initial data size is suggested that finds the exact *conditional* (refer to Section 2.5) global minimum under no gene transfers; that is, in this case the model is also exactly solvable. However, for the case of transfers the algorithm is not mathematically proved to find the exact conditional global minimum, which remains an important open problem. The heuristic solution for this case and its usage are described in [23, 25, 45, 64, 79].

We end this section with giving a solid mathematical background for the second problem for a fixed set of polytomous rooted gene trees. This problem is also discussed in [25].

2.2. Auxiliary Definitions. Let a gene tree G and a species tree S be given. The trees are rooted and binary, and oriented downwards from the roots. Recall that edges of the tree S are referred to as *tubes* to distinguish between the edges of S and G. Each root is supplied with an additional *root edge* (or root tube), which initiates in a *superroot* and ends in the root; that is, the superroot is the only vertex inducing the single child. Each leaf is labeled with a species name. Species names in S are unique; species names in G may duplicate if it contains several genes from the same species (paralogous genes). Species names in G are a subset of those in S. A *subtree* is a part of a tree that consists of a vertex, an edge entering the vertex from above (the subtree root edge), and all vertices and edges descending downwards. A *clade* of a subtree is a set of species names present in all its leaves; a clade of a vertex is the clade of its subtree. For a clade V, the corresponding tree is referred to as a *tree over V. A paralogous subtree* (with respect to a species) in G is such a maximal subtree that has all leaves marked with one species (i.e., its clade is a singleton; the paralogs are in-paralogs for this species). Pruning of a subtree T from tree G is a deletion of all edges and vertices in G belonging to T followed by merging of the two edges, incoming in and outcoming from the upper terminus of the root edge in T. A *child* of a vertex is another vertex located directly downwards, that is, at a distance of one edge. Remember that $\geq$ and $>$ mean the natural order on any tree as defined in Section 1.2. The *natural order* relation is defined analogously on a set of edges (tubes), a set of vertices, or a united set of edges and vertices. The terms "lower" and "upper" refer to the natural order of the tree branching downwards from the root.

Let e_+ be the lower terminus of edge e, and let e^+ be its upper terminus.

2.3. Definition of Mapping with Duplications and Losses Only: Reconciling Gene and Species Trees. A *mapping* f of a gene tree G into a species tree S is an assignment of each vertex in G to a vertex or tube in S, the superroot is mapped into the root tube, and each leaf is mapped into a leaf with the same species name. Two conditions are imposed on f: if a vertex is mapped into a tube, its child is mapped into the same tube or downwards (lower); if $f(g)$ is a vertex in S, then for children g_1 and g_2 of g, the values $f(g_1)$ and $f(g_2)$ are in the two different descendent (lower) subtrees of $f(g)$ in S.

Examples and illustrations of mappings are given in [23, 25, 45, 64].

2.4. Definitions of Gene Duplication and Loss and Their Localization on the Species Tree. Let mapping f be fixed. In f, a *duplication* is a nonsuperroot vertex g for which $f(g)$ is a tube, a *divergence* is a nonleaf vertex g for which $f(g)$ is a vertex, and a *loss* is a pair $\langle e, s\rangle$ of edge e in G and vertex s in S such that, for the upper terminus e^+ and the lower terminus e^+ of e, we observe $f(e_+) < s < f(e^+)$. If the clade of a child of s contains no species from G, the loss is called *implicit* (as it is induced by species in S but not in G). Otherwise, the loss is called *explicit*. A duplication is *located* in the corresponding tube in S, a divergence in the corresponding vertex in S, and a loss in vertex s.

Each event type (duplication, loss, divergence, etc.) is assigned a nonnegative cost value. A *cost* of mapping f of G into S is the sum of event costs inferred in this mapping. A *cost* of mapping $\{f_j\}$ of a set of gene trees G_j into a species tree S is the total cost of mappings f_j of G_j into S. Denote these costs $c(G, S, f)$ and $c(\{G_j\}, S, \{f_j\}) = \sum_j c(G_j, S, f_j)$, respectively. The variables f and $\{f_j\}$ are often implied but not written explicitly.

A mapping with the minimal cost is called *canonic* and designated α, [33, 63]. A linear algorithm to construct it is described in [65, 66]; more details can be found in [20].

Denote V_0 a set of all species names in all given gene trees G_j.

2.5. Formulating the Problems of Reconciling Two (Gene and Species) and Many (Gene) Trees. During the reconciliation of two trees, for given gene G and species S trees, a mapping f is sought for such that it globally minimizes the functional $c(G, S, f)$ over the variable f.

During the reconciliation of many trees, for a given set $\{G_j\}$ of gene trees, a set of mappings $\{f_j\}$ and a tree S^* are sought for such that they globally minimize the functional $c(\{G_j\}, S, \{f_j\})$ over the variables $\{f_j\}$ and S. This minimization is done under the ad hoc *condition*: each S must contain only clades belonging to a *predefined* set P of subsets of set V_0; all clades from $\{G_j\}$ are by default already contained in P.

Traditionally, the second problem requires the *unconditional* (absolute) minimization. We refer to the introduced reformulation as to the *parametric* (over the parameter P) or *conditional* minimization (optimization).

2.6. The First Problem under Gene Duplications and Losses Only: Reconciling Gene and Species Trees. If g is not the superroot vertex g_0 in tree G, denote LCA(g) the last common ancestor in S of a clade defined in G by a subtree with the root vertex g. A *second definition* of canonic mapping α slightly differs from the definition provided in Section 1.2 as follows. Let $\alpha(g_0) = d_0$, where d_0 is the root tube. If for both children g_1 of vertex g holds the inequality $\text{LCA}(g) \neq \text{LCA}(g_1)$, then $\alpha(g) = \text{LCA}(g)$; otherwise $\alpha(g)$ is a tube incoming to LCA(g) from the upwards. Informally, $\alpha(g)$ may be visualized as located "inside the tube." Hereafter, only the second definition of canonic mapping α is used. Analogously, a set $\{f_j\}$ of mappings f_j is canonic if each f_j (of G_j into S) is canonic. From the remark to Lemma 2 it follows that the second definition of α and its definition based on the global cost minimization are equivalent. The second definition is given in [33, 63].

Lemma 1. *If mapping f is not canonic α, then for each vertex g the inequality $f(g) \geq \alpha(g)$ is valid, and, at least for one g, $f(g) > \alpha(g)$.*

Proof. Clade $f(g)$ contains clade g, that is proved with induction from leaves to g. The first inequality follows from the statement above and the observation: if $\alpha(g)$ is a tube, then $f(g)$ is not its lower terminus, as the terminus already contains a descendant of g. By definition, f cannot map two comparable vertices to one. The condition $f \neq \alpha$ implies the last statement of the lemma. □

Lemma 2. *For any mapping f different from α, the amount of duplications for f is not less than for α, and the amount of losses is strictly greater for f than for α.*

Proof. Consider a duplication for α; that is, $\alpha(g) = d$, where d is a tube. Then $f(g)$ cannot be a vertex $s > d$, as then, by definition of mapping, one of the children of g must map in a descendent subtree of s not containing d. It is impossible, as the clade g does not intersect with the clade of the descendent subtree (it is further referred to as the *bifurcation effect*). According to Lemma 1, $f(g) \geq d$, therefore $f(g)$ is a tube. Consequently, a duplication for α remains a duplication for f. Note that a divergence for α may become a duplication for $f \neq \alpha$.

First prove that the amount of losses is not less for f than for α. Consider a loss (e, s) for α. If $f(e_+) < s$, it remains a loss for f, because, according to Lemma 1, $f(e^+) \geq \alpha(e^+) > s$. The equality $f(e_+) = s$ is false due to the bifurcation effect.

Next, due to $f(e_+) \geq \alpha(e_+)$ obtain that $f(e_+)$ is comparable with s. If $f(e_+) > s$, the loss (e, s) corresponds to at least two losses, (e_1, s) and (e_2, s), in s for f, where $e_1 \neq e_2$ and both $e_1, e_2 < e$. Indeed, on any path from e_+ downwards, an edge will induce a loss in s (a divergence cannot occur on the path due to the bifurcation effect). If (e', s') is another loss for α, it corresponds to two losses in f differing by s or e, given that e' is incomparable with e. Thus, there exists a multivalued injective mapping that maps each loss in α to one or two losses in f, with nonintersecting images. Since for a fixed vertex s the property of being an explicit or implicit loss in s depends on the tree G only, and losses for f are of the same type (explicit or implicit) as for α.

The Last Statement of the Lemma. By the condition and Lemma 1, there exists a vertex g in G, for which $f(g) > \alpha(g)$. The two cases are possible: (i) $f(g)$ and $\alpha(g)$ are tubes, (ii) $f(g)$ is a tube, and $\alpha(g)$ is a vertex. Indeed, for a vertex $f(g)$ a contradiction arises according to the bifurcation effect.

Case (i). Let s be an arbitrary vertex, for which $f(g) > s > \alpha(g)$. Consider two nonoverlapping paths from g to the leaves. On both paths there occur edges e_1 and e_2 inducing for f losses l_1 and l_2 in s (the bifurcation effect). As $s > \alpha(g)$, the paths from e_1 and e_2 to the root contain either none or one coincident loss in s for α. Consequently, the losses l_1 and l_2 either are not contained in the mapping image μ or constitute the image of one coincident loss. In both alternatives, the amount of losses is greater for f than for α.

Case (ii). Consider an arbitrary path from g to a leaf. According to the bifurcation effect, it contains an edge e such that $(e, \alpha(g))$ is a loss for f. This loss is not contained in the mapping image μ, as there exists no edge e' on the path from e to the root such that $(e', \alpha(g))$ is a loss for α. □

Remark 3. Let f and h be two different mappings. If for any vertex g holds the inequality $f(g) \geq h(g)$, then by substituting α to h in Lemma 2 we prove that the amounts of duplications for f is not less than for h, and the amount of losses is greater for f than for h. An analogous statement is proved in [115] for vertex-to-vertex mapping functions.

Henceforth, *assume* that the cost of a divergence is less than the cost of a duplication; this condition is likely to be biologically justified. Then, by Lemma 2, a canonic mapping α is a solution of the first problem, that is, the two definitions of α coincide. Further, if in a set $\{f_j\}$ a mapping f_j is not canonic, then its replacement with a canonic mapping will reduce the total cost. Thus, in the second problem the only true variable is the desired species tree S.

Lemmas 1-2 solve the first problem only for the case when the gene duplication, loss, and divergence during speciation are considered. Lemmas 4-7 prove certain properties of α and will be used in the proofs of Theorems 8-9.

Lemma 4. *If a gene tree G is obtained from a species tree S by pruning some subtrees from S, then for a canonic mapping α of G into S duplications and explicit losses are absent, and each pruned subtree (with the root tube d) induces an implicit loss in d^+. Conversely, if for a canonic α there are no duplications, then G is obtained from S by pruning some subtrees.*

Proof. Prove the lack of duplications with induction on the amount of pruned subtrees. At the start of induction, $G = S$, and only a divergence event is possible.

An induction step from G_n to G_{n+1}, where n is the number of pruned subtrees. If in G_n it is true that $\alpha(g)$ is a vertex s and a vertex g is not pruned, then both clades of its children in G_n and G_{n+1} are subsets of the clades of corresponding children of vertex s. These children in G_{n+1} still map in α strictly below s. Therefore, in G_{n+1} also $\alpha(g) = s$, vertices in G_{n+1} map into vertices, and a duplication does not occur.

Prove the absence of explicit losses by contradiction. An induction step. Let a vertex s contain an explicit loss $\langle e, s\rangle$ after pruning a $(n+1)$th subtree T. Then in tree S both clades of the children of s contain species from G_{n+1}. Consequently, there exists a vertex g, for which $\alpha(g) = s$, as such g existed in G_0 and was not pruned. Thus, edge e does not exist.

A vertex of tree S, a former image of the upper terminus of the root edge of a pruned subtree, contains an implicit loss in G_{n+1} induced by a new edge in the tree G_{n+1} formed after merging of two initial edges.

Let us prove that, for any vertex s in S, a set $M_s = \{g \mid \alpha(g) \in T_s\}$, where T_s is a subtree rooted in s, defines a tree obtained from T_s by pruning certain subtrees. If s is a root, this statement is obtained. Prove it with induction. If s is a leaf, the statement is obvious. An induction step. Assume a vertex g, for which $\alpha(g) = s$. Then the sought subtrees set is the union of the corresponding sets for children s_1 and s_2 of vertex s. Assume there is none such g. Then the images of members of M_s belong to one child subtree of vertex s (put it the child s_1); otherwise s will contain the last common ancestor of members of M_s. The sought set of subtrees consists of a subtree rooted in s_2 and the set of subtrees for s_1. □

Lemma 5. *In a canonic α of G into S, each leaf tube terminating in species s contains the number of duplications equal to the number of nonleaf vertices in a paralogous subtrees for s in the gene tree G.*

Proof. Denote the number of nonleaf vertices in Lemma 5 by $\mathrm{Par}(G, s)$. Any internal vertex of a paralogous subtrees induces a duplication according to the definition of α. And, conversely, such a duplication corresponds to a vertex in G that is contained in a paralogous subtree for s. □

Lemma 6. *Fix a gene tree G over a subset of V_0. Let species trees S_1 and S_2 be both defined over V_0, each containing a certain subtree S. Then the two canonic mappings of G into S_1 and G into S_2 produce the same set of events in S; that is, the set of events in a subtree does not depend on the subtree's complement (the rest of the tree).*

Proof. Let V be a clade of the subtree S. Vertices mapped in S coincide in both mappings, as their clades belong to V; the image of such vertices coincides in both mappings, as it depends only on S. By definition of the duplication, the set of duplications in S coincides in both mappings. Let $\langle e, s\rangle$ be a loss in one mapping α, where s is a vertex in S. Then in another mapping α the image of edge e_+ remains constant, and the image of e^+ also remains constant (if belongs to S) or remains external to S (if does not belong to S) above the image of e_+ and, therefore, above s. In both cases, $\langle e, s\rangle$ is a loss also in another mapping α. Thus, the set of losses also coincides between two mappings α. □

Lemma 7. *Fix a gene tree G over a subset of V_0. Let V be a subset of V_0, and a species tree S_1 over V_0 contains a subtree T_1 over V. Let a tree S_2 be derived from S_1 by substituting the subtree T_1 with a subtree T_2 over V. Then the canonic mappings of G into S_1 and G into S_2 produce the same set of events in the complements to the subtrees T_1 and T_2; that is, the set of events in a complement to a subtree does not depend on this subtree.*

Proof. By definition of mapping α, vertices mapped outside S_i are the same for $i = 1, 2$, as their clades do not belong to V, or, equivalently, their LCA images are not contained in S_i. Each such vertex g has the same α-image. Indeed, the values of $\mathrm{LCA}(g)$ coincide on S_1 and S_2; that is, if on one of the S_i the $\alpha(g)$ is a tube, it is a tube on the other. By definition of a duplication, the set of duplications outside S_i coincides between the two mappings. Let $\langle e, s\rangle$ be a loss in one mapping, where s does not belong to S_i. Then in the other mapping the image of e^+ does not change, and that of e_+ either does not change (if not belong to S_i) or remains in S_i (if belongs to S_i) and thus is below s. In both cases, $\langle e, s\rangle$ is a loss also in the other mapping. Consequently, the set of losses also coincides between the two mappings. □

2.7. The First Problem under Gene Duplications, Losses, and Horizontal Transfers with Imposed Time Slices: An Algorithm to Reconcile Gene and Species Trees (Building an Evolutionary Scenario). The generalization of mapping α to incorporate gene transfers has long been a daunting task. Here we describe an original approach to solve it.

Let the species tree S impose certain time slices; refer to Sections 1.4-1.5; the slices are ranked from the root to leaves. The slices must satisfy the single condition: if $d_1 \le d_2$, then the rank of d_1 is not less than the rank of d_2. For example, a kth slice contains all tubes distanced by the amount of k tubes from the root; in [25] the slices are constructed with an additional condition: all leaf tubes belong to one slice. The latter condition is inessential in further definitions and is accepted without discussion. *Denote* $d_1 \sim d_2$ for tubes d_1 and d_2 if $d_1 \ne d_2$ and d_1, d_2 belong to the same time slice.

With horizontal transfers, we formulate a similar (refer to Sections 2.3, 2.6) but inductive definition of *mapping* f of a gene tree G into a species tree S and its *cost* [64]. Simultaneously with f, an additional tree G' is defined as derived from G by inserting *new vertices with a single child*. The number n of new vertices on an edge defines the number of transfers: if n is even, a gene (more precisely, edge e in G; see below) underwent $n/2$ transfers without retention of the gene donor copy, and if n is odd, a single transfer with and $(n-1)/2$ transfers without retention.

Let e be any edge in G,and let d be any tube in S. The definition of f and G' is based on an important auxiliary definition of the *inner tree* and its *cost* for *any pair* $\langle e, d\rangle$. All pairs of the form $\langle$edge from G, tube from $S\rangle$ are partially ordered: a pair $\langle e, d\rangle$ is lower than $\langle e', d'\rangle$ if $e < e'$ or $e = e'$ and the rank of tube d is greater than that of d'. Pairs $\langle e, d\rangle$ are visited from leaves to the root in the linear order consistent with the described partial order. Remember that any vertex is identified by its incoming edge.

2.7.1. Defining the Inner Tree for the Pair $\langle e, d\rangle$. The start of induction. Let e and d be any leaf edge and any leaf tube, respectively, and let d' be a tube with the species of gene e in its lower terminus. If $d \ne d'$, the inner tree contains the pair $\langle e, d\rangle$ and its single child, the pair $\langle e, d'\rangle$; this corresponds to a transfer without retention of the donor copy from d into d'. If $d = d'$, the inner tree consists of the single pair $\langle e, d\rangle$. The *cost* of this tree is the cost of a transfer without retention if $d \ne d'$ and is zero otherwise (for more details on transfers refer to Sections 1.4-1.5) [23, 25, 45, 64, 79].

Thus, the inner tree is a marked tree; the mark of a vertex has the form $\langle e, d\rangle$.

2.7.2. An Induction Step. Let e and d be a nonleaf edge and tube, respectively. Then the *inner tree* and its *cost* for the pair $\langle e, d\rangle$ are defined as follows depending on the sequential choices listed below. Namely, for any $\langle e, d\rangle$, the outcome is selected according to rules 1–6 below, with some of the rules describing a choice. In square brackets is the description of applicability. Otherwise said, a set of inner trees is defined, with each inner tree describing an alternative evolution of gene e inside species d.

(1) [Tube d has the single child d_1]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d_1\rangle$ that roots the already known inner tree for $\langle e, d_1\rangle$. This tree has the cost equal to that of $\langle e, d_1\rangle$. Descriptively, lineage e enters the next tube.

(2) [Tube d has two children, d_1 and d_2]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d_1\rangle$ that roots the inner tree for $\langle e, d_1\rangle$ or the child $\langle e, d_2\rangle$ that roots the inner tree for $\langle e, d_2\rangle$ (only one case must be chosen). The cost of this tree is the cost of the chosen $\langle e, d_i\rangle$ plus the cost of a loss (explicit if the other child d_j possesses at least one leaf from G and implicit otherwise). Descriptively, lineage e survives only in one of the two tubes.

(3) [Edge e has children e_1 and e_2]. The inner tree consists of the pair $\langle e, d\rangle$ with two children, $\langle e_1, d\rangle$ and $\langle e_2, d\rangle$, which root the inner trees for pairs $\langle e_1, d\rangle$ and $\langle e_2, d\rangle$. Its cost is the sum of costs of trees $\langle e_1, d\rangle$ and $\langle e_2, d\rangle$ and a duplication. Descriptively, lineage e is duplicated in d.

(4) [Edge e has children e_1 and e_2; tube d has children d_1 and d_2]. The inner tree consists of the pair $\langle e, d\rangle$ with two children, $\langle e_1, d_1\rangle$ and $\langle e_2, d_2\rangle$, which root the inner trees for pairs $\langle e_1, d_1\rangle$ and $\langle e_2, d_2\rangle$. Its cost is the sum of costs of trees $\langle e_1, d_1\rangle$, $\langle e_2, d_2\rangle$, and a divergence. In the alternative choice, e_1 and e_2 swap. Descriptively, lineage e diverges in d.

(5) [Edge e has children e_1 and e_2]. The inner tree consists of the pair $\langle e, d\rangle$ with two children, $\langle e_2, d'\rangle$ and $\langle e_1, d\rangle$, which root the trees for pairs $\langle e_2, d'\rangle$ and $\langle e_1, d\rangle$, where $d' \sim d$. Its cost is the sum of costs of the trees for $\langle e_1, d\rangle$, $\langle e_2, d'\rangle$, and a transfer with retention. In the alternative choice, e_1 and e_2 swap. Descriptively, lineage e duplicates in d with a subsequent transfer into d' and retention of the donor copy in d.

In rule 6 the definition of d' is used in the same sense.

(6) In this rule, descriptively, lineage e duplicates in d with subsequent transfers into d' and losses of the donor copy in d.

(6.1) [Tube d' has the single child d'_1]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d'\rangle$, which also produced the single child $\langle e, d'_1\rangle$ that roots the tree for $\langle e, d'_1\rangle$. The cost of this tree is the sum of costs of $\langle e, d'_1\rangle$ and a transfer without retention. Descriptively, lineage e enters from d' into the next tube d'_1.

(6.2) [Tube d' has two children, d'_1 and d'_2]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d'\rangle$, which also produces the single child $\langle e, d'_1\rangle$ that roots the tree for $\langle e, d'_1\rangle$. The cost of the tree is the sum of costs: $\langle e, d'_1\rangle$, a transfer without retention, and a loss in d'_2 (explicit if d'_2 possesses at least one leaf from G and implicit otherwise). The alternative is the choice for $\langle e, d'_2\rangle$. Descriptively, lineage e survives only in one of the two tubes.

(6.3) [Edge e has children e_1 and e_2]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d'\rangle$, which produces two children, $\langle e_1, d'\rangle$ and $\langle e_2, d'\rangle$, which root the trees for $\langle e_1, d'\rangle$ and $\langle e_2, d'\rangle$. The cost of the tree is the sum of costs of $\langle e_1, d'\rangle$, $\langle e_2, d'\rangle$, a transfer without retention, and a duplication in d'. Descriptively, lineage e duplicates in d'.

(6.4) [Edge e has children e_1 and e_2; tube d' has children d'_1 and d'_2]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d'\rangle$ that produces two children, $\langle e_1, d'_1\rangle$ and $\langle e_2, d'_2\rangle$, which root the trees for $\langle e_1, d'_1\rangle$ and $\langle e_2, d'_2\rangle$. The cost of this tree is the sum of costs: $\langle e_1, d'_1\rangle$, $\langle e_2, d'_2\rangle$, a transfer without retention, and a divergence. In the alternative choice, e_1 and e_2 swap. Descriptively, lineage e transfers in d' and then diverges in the lower terminus of d'.

(6.5) [Edge e has children e_1 and e_2]. The inner tree consists of the pair $\langle e, d\rangle$ with the single child $\langle e, d'\rangle$ that produces two children, $\langle e_2, d''\rangle$ and $\langle e_1, d'\rangle$, which root the trees for $\langle e_2, d''\rangle$ and $\langle e_1, d'\rangle$, where $d'' \sim d' \sim d$ (tube d'' differs from tubes d and d'). The cost of the tree is the sum of costs of $\langle e_1, d'\rangle$, $\langle e_2, d''\rangle$, and transfers with and without retention. Descriptively, lineage e transfers in d', duplicates in d' with a subsequent transfer into d'' and retention of the donor copy in d'. The end of the inner tree definition.

Remember the notation: subscript and superscript indices of "+" designate lower and upper termini, respectively, or edges and tubes; e_0 and d_0 are the root edges in trees G and S.

The inner tree T for the pair $\langle e_0, d_0\rangle$ is used to construct a candidate mapping $f = f_{T,\langle e_0, d_0\rangle}$ and simultaneously a candidate tree G', which vertices are mapped into vertices and tubes of tree S. Namely, when running the vertices of an inner tree T for the pair $\langle e_0, d_0\rangle$ from its leaves upwards to the root consider the following. Let e_1 and e_2 be children of edge e, and let d_1, d_2 be children of tube d. In square brackets is the description of applicability followed by the rule formulation. Each pair $\langle e, d\rangle$ marks the corresponding vertex in tree T:

(0) $f(e_0^+) = d_0$;

(1) [leaf vertex $\langle e, d\rangle$] $f(e_+) = d_+$;

(2) [vertex $\langle e, d\rangle$ has a child of the form $\langle e_i, d\rangle$] $f(e_+) = d$. If the other child has the form $\langle e_j, d'\rangle$, a new vertex g' (with the single child) is inserted on edge e_j in current G' and $f(g') = d'$. If the edge e_j already received a number of single-child vertices, a new single-child vertex is inserted in the edge upwards of the already received;

(3) [$\langle e, d\rangle$ has the children $\langle e_1, d_1\rangle$ and $\langle e_2, d_2\rangle$] $f(e_+) = d_+$;

(4) [$\langle e, d\rangle$ has the single child $\langle e, d'\rangle$]. Insert two vertices g' and g'' on edge e in current G'(each with the single child; g' is higher than g'') and $f(g') = d$, $f(g'') = d'$.

The set of candidate *mappings* f of G' into S is obtained. Candidate *partial mappings* $f_{T,\langle e,d\rangle}$ for any pair $\langle e, d\rangle$ are obtained analogously, as well as candidate partial trees $G'_{T,\langle e,d\rangle}$. The end of the candidate mapping definition.

A scenario (*mapping*) f^* is a candidate mapping that minimizes the total cost of its evolutionary events.

The role of the inner tree for $\langle e_0, d_0\rangle$ is to describe the evolution of a gene described by a tree G inside the species described by a tree S; if a pair $\langle e, d\rangle$ is a vertex of the inner tree then edge e evolves inside tube d at least along its certain segment.

An algorithm to build the scenario trivially repeats the same induction that was used to define the inner tree: for every pair $\langle e, d\rangle$, the choice will minimize the cost over all possible choices. The same induction is used to build the mapping that coincides with canonic α when transfers are not considered.

To account for gene gain events, we introduce an auxiliary outgroup, a tube d^* connecting the root of S with an auxiliary outgroup species d^*. Introducing time slices generates tubes on the outgroup tube with single children, which we also denote d^*. Gene lineage that evolves into the outgroup tube and later transfers back into the initial species tree S is considered as *gained*. The start of induction is modified as follows: for d^*, the cost of a transfer without retention is replaced with a fixed gain cost. Induction steps are also modified. In rules 2-3, the costs of loss and duplication are zeroed for d_0 and d^*, respectively. Rule $3'$ is added: for a pair $\langle e, d_0\rangle$, the inner tree consists of $\langle e, d_0\rangle$ with two children, $\langle e_1, d_0\rangle$ and $\langle e_2, d^*\rangle$, which root the inner trees for pairs $\langle e_1, d_0\rangle$ and $\langle e_2, d^*\rangle$, where e_1 and e_2 are children of e, and d^* is the upper outgroup tube. The cost of this tree is the sum of costs for $\langle e_1, d_0\rangle$ and $\langle e_2, d^*\rangle$. In the alternative choice, e_1 and e_2 swap. In rule 4, the cost of a divergence is zero for d_0. A condition is added in rules 5, 6.1, 6.2, 6.3, 6.4, and 6.5: tubes d', d'' are not in the outgroup; for d^*, the cost of a transfer with retention (rule 5) or without retention (rule 6) is replaced by the gain cost.

In [25] we describe an even more extended list of evolutionary events. The nontrivial definitions and algorithm above were proposed and thoroughly tested in [45]. In [79] the *complexity of the algorithm* was mathematically proved to be *cubic* with respect to the number of vertices in the species tree that contains time slices. In [45] it was mathematically proved that the algorithm finds the minimal mapping and its cost under the presence of horizontal transfers.

The first problem is solved for the general case.

2.8. The Second Problem: Phase 1 of the Supertree Building Algorithm under Gene Duplications and Losses Only. Hereafter, all mappings are canonic α. Only duplication, loss, and divergence events are considered.

Consider a set of gene trees $\{G_j\}$ with a set of species called V_0. To find is a species tree S^* over V_0, for which the total cost of individual tree mappings is globally minimal. It is an NP-complete problem. To overcome this limitation, we reformulate the problem of unconstrained optimization into a biologically justified constrained (conditional) optimization problem. Constrain the solution space to contain only species trees S satisfying the condition: all clades of S belong to a predefined set P, which includes at least all clades of input gene trees. Thus, S^* must also satisfy this condition. The parameter P is nontrivial and is introduced to overcome the NP-complete nature of the problem. A "true" species tree may not exist in this solution space, depending on the degree of consistency of the input set of clades.

The proposed original algorithm of solving the second problem consists of two phases. An exact solution is obtained during Phase 1, provided that the conditional optimization problem is solved under a certain condition.

If the condition is not valid, a follow-up heuristic procedure implemented in Phase 2 can be invoked, which outcome depends on the data generated during Phase 1. As with real data the existence of the unconstrained solution in the solution space for a fixed P is usually unknown, one can either empirically expand the set P or take the heuristic solution obtained during Phase 2. In computer simulations the latter strategy produced better results (data not shown).

Description of Phase 1. Standard approaches are used to define algorithmic relations over sets from P: the "inclusion of one set into another," "intersection of two sets is empty," and "cardinality of a set." Also, the algorithmic relation is defined between vertices of G_j (separately for each j) and their clades from P. Different vertices (even within one tree) may correspond to the same clade; the set P may contain sets that do not correspond to any clade in the input gene trees.

For each set V from P the set of all its partitions is defined. A partition is a pair $\langle V_1, V_2\rangle$ of nonempty nonintersecting subsets V_1, V_2 of set V that belong to P and their union equals V; partitions are easily calculated by verifying the condition $|V_1| + |V_2| = |V|$. Sets from P that can be so partitioned down to singletons are defined as *basic*; all singletons are also defined as *basic*. The set P may contain nonbasic sets. Thus, an initial V_0 may be nonbasic, which invokes Phase 2 of the algorithm. By induction, we enumerate all basic sets according to the increasing of their cardinality. For each basic set, Phase 1 constructs a tree $S(V)$ over V, called a *basic tree*, and computes its *cost*. In the algorithm implementation, the construction of basic trees and computing their costs are naturally combined. For any singleton s from P, tree $S(s)$ contains the single leaf (the root) s and the root tube; its *cost* is zero if there are no paralogous trees for s and is the cost of one duplication multiplied by $\sum_j \mathrm{Par}(G_j, s)$ otherwise (refer to Lemma 5).

2.8.1. Definition of Basic Trees S(V) and Their Costs: The Induction Step. Fix nonsingleton basic set V from P and enumerate all its partitions into basic sets V_1 and V_2 with lesser cardinality.

For each partition, compute a *new cost* $c(V, V_1, V_2)$ as follows. Denote $V(g)$ the clade of a vertex g. Let g_1 and g_2 be children of g; if g is a superroot, then $g_1 = g_2$. Run each g in all G_j and compute the following numbers q_1, q_2, q_3', and q_3''.

The number q_1 of vertices g in all G_j, for which $V(g_1) \subseteq V_1$ and $V(g_2) \subseteq V_2$; (or otherwise: $V(g_1) \subseteq V_2$ and $V(g_2) \subseteq V_1$ (the sign $\subseteq$ stands for "a subset")); the number q_2 of vertices g in all G_j, for which $V(g)$ is a subset of V and at least one of the sets $V(g_1)$ or $V(g_2)$ has non-empty intersection both with V_1 and with V_2.

Select gene trees G_j for which (i) the root clade intersects with both sets V_1 and V_2 and (ii) the root clade intersects with one of the sets and not with the other.

Compute the number q_3' of edges $e = (e^+, e_+)$ in all G_j satisfying (i) and the new condition (iii): $V(e_+)$ is a subset of V_1 or V_2, and either e^+ is the superroot, or for the child $g \neq e_+$ of e^+, the set $V(g)$ is a subset neither of V_1 nor of V_2. Also compute the number q_3'' of edges in all G_j satisfying (ii) and (iii).

Define a new cost

$$c(V, V_1, V_2) = c(V_1) + c(V_2) + c_{\text{div}} \cdot q_1 + c_{\text{dup}} \cdot q_2 + c_{\text{los1}} \cdot q_3' + c_{\text{los2}} \cdot q_3'', \tag{9}$$

where c_{div} is the cost of a divergence, c_{dup} is the cost a duplication, c_{los1} is the cost of an explicit loss, and c_{los2} is the cost of an implicit loss.

Assume that $c(V, V_1^*, V_2^*)$ is the minimal cost among $c(V, V_1, V_2)$ for all partitions $\langle V_1, V_2 \rangle$ of V. The tree $S(V)$ is *obtained* by merging trees $S(V_1^*)$ and $S(V_2^*)$ under the join root, where $\langle V_1^*, V_2^* \rangle$ is one of the pairs satisfying the minimal cost requirement. The *cost* of $S(V)$ is defined as $c(V, V_1^*, V_2^*)$.

Phase 1 outputs a set $\{S(V) \mid V\}$ of basic trees $S(V)$ for each basic set V. The end of Phase 1.

2.9. Justification of Phase 1. Let S_1 be an arbitrary species tree over V_0 that includes a subtree $S(V)$. *Denote* $c(V)$ the total cost of events in $S(V)$ in canonic mappings of all gene trees G_j in S. The cost $c(V)$ differs from the total cost $c(\{G_j\}, S_1)$ as it accounts only for the events in $S(V)$; of course, if $V = V_0$, the costs are equal. If any tree S_2 over V_0 is considered that contains $S(V)$ as a subtree, the cost $c(V)$ will remain the same as for S_1 according to Lemma 6. Thus, the cost $c(V)$ is a function of the tree $S(V)$ and does not depend on its comprising tree S_1.

Evidently, if the second conditional problem is solvable, then V_0 is a basic set, and the tree $S(V_0)$ is the solution according to Theorem 8.

Theorem 8. *A basic tree $S(V)$ globally minimizes the functional $c(V)$ in the conditional problem for V if the problem is solvable. The algorithm constructs $S(V_0)$ in time $|P|^3 + |P|^2 \cdot |V_0| \cdot n$, where n is the number of input trees G_j.*

Proof. Obviously, the solution exists if and only if V is a basic set. The time complexity is proved in [64].

By induction, enumerate basic sets according to the increasing of their cardinality. For a singleton set, the statement of Theorem 8 follows from Lemma 5. Let V be a nonsingleton set. Prove that, for each partition of V into V_1 and V_2, the computed value $c(V, V_1, V_2)$ equals the sum of event costs in a tree T, where T is a result of merging trees $S(V_1)$ and $S(V_2)$ under the common root (as mentioned above, the value $c(T)$ depends on T only). Denote r the common root, and d—the tube entering the root (the root tube). There are three groups of considered events: (i) events in $S(V_1)$, (ii) events in $S(V_2)$, and (iii) events occurring in r or in d. By inductive assumption, the total event cost of groups (i) and (ii) is $c(V_1) + c(V_2)$.

Examine the total event cost of group (iii). From definitions of mapping α and the events, it easily follows that

(1) $\alpha(g) = r$ (a divergence event) if and only if the condition on g corresponding to the number q_1 in the algorithm description is satisfied;

(2) $\alpha(g) = d$ (a duplication event) if and only if the condition on g corresponding to the number q_2 in the algorithm description is satisfied;

(3) pair $\langle e, r \rangle$ is a loss if and only if condition (iii) in the algorithm description is satisfied; the loss is explicit if condition (i) on G_j is satisfied and implicit if condition (ii) is satisfied.

Thus, the algorithm finds the numbers of duplications in d, divergences in r, explicit and implicit losses in r, and their total cost. Consequently, the value $c(V, V_1, V_2)$ is computed correctly.

Let a certain tree $T(V)$ be the global minimum of the functional $c(V) = c(T(V))$ if all its clades belong to the set P. The root bifurcation corresponds to a partition of V into two basic sets, V_1 and V_2. If subtrees $T(V_1)$ and $T(V_2)$ are replaced with trees $S(V_1)$ and $S(V_2)$, respectively, then by Lemma 7 the functional $c(V)$ does not decrease (indeed, if, e.g., $T(V_1)$ is replaced by $S(V_1)$, the cost of the events from group (i) does not decrease, and the total cost of groups (ii) and (iii) remains constant). Consequently, such a replacement does not affect the global minimum, and trees over V_1 and V_2 in the desired solution can be legitimately considered those $S(V_1)$ and $S(V_2)$ that are already constructed at previous steps of the algorithm. The algorithm will output as $S(V)$ the global minimum of the functional $c(V)$. □

2.9.1. Remark. According to Lemma 5, the cost $c(V)$ includes the total cost of duplications in all paralogous subtrees over all G_j over all species from V. Therefore, the costs of singletons can be any constants, as the optimal tree $S(V)$ does not depend on them. The set $\{G_j\}$ can also be simplified by replacing all paralogous subtrees with singleton subtrees.

Phase 1 of the algorithm produces a set $\{S(V) \mid V\}$ of basic trees, where V runs over all basic sets. If the set V_0 of all species is not basic, it will not contain a tree over V_0. In this case, Phase 1 returns no conditional supertree; that is, the conditional problem has no solution.

A natural question is "how to determine if the degree of consistency of the input set of trees suffices for the correct supertree to exist?" An empiric directive for the moment can be that the trees are consistent enough if V_0 is a basic set.

2.10. The Second Problem: Phase 2 of the Supertree Building Algorithm. The set P is not unambiguously defined by the initial set of gene trees G_j. For this reason, a heuristics is implemented in Phase 2 of the algorithm to solve the unconditional problem and assemble basic trees $S(V)$ into one species tree S^* over V_0 under a certain fixed P. This heuristic solution largely depends on the outcome of Phase 1. The assembling can be done using a variety of known methods. We propose an original *ad hoc* "augmentation" method described below.

Consider a tree S over a set $V \subseteq V_0$. Its *cost* $c(S)$ is defined as the total cost of mappings of *all basic trees* (with two or more leaves) pruned to contain only species from the set V.

Let V contain only three species. The *basic cost* $c(V)$ is the minimal cost $c(S)$ among all trees S over V. The *subbasic cost* $c'(V)$ is the minimal cost $c(S)$ strictly greater than $c(V)$. The *reliability* $R(V)$ is defined as $(c' - c)/c'$. By enumerating all such V, find a tree S over V with a nonzero reliability and the minimal value of $c(V) \cdot (2 - R(V))$. If for any V the cost $c'(V)$

does not exist, the algorithm terminates. The final tree S is the result of the basis of the induction.

An inductive step is similar. Let a tree S with $n \geq 3$ species be obtained. Consider all pairs: species s from V_0 not contained in S and edge d from S including its root edge. The edge d is broken in two by inserting a new vertex connected with a newly added leaf s, thus generating a new tree S'. The basic cost $c(s)$ is the minimal cost $c(S')$ when s is fixed and d is a variable. The subbasic cost $c'(s)$ is the minimal cost $c(S')$ strictly greater than $c(s)$. The reliability $R(s)$ is defined as above. By enumerating all s find a tree S' with a nonzero reliability and for which $c(s) \cdot (2 - R(s))$ is minimal. If c' does not exist for a species s, the species is marked as unreliable and not used in Phase 2. An augmentation step is a transition from S to S'; the steps are continued until the current S' contains all successfully attempted species from V_0. The resulting species tree is the output of Phase 2 of the algorithm.

The correctness of Phase 2 is proved by Theorem 9. Informally, the topologies of trees G_j in Theorem 9 are assumed to share at least some topological similarity.

Theorem 9. *Let the cost of an implicit loss be zero. If there exists a tree S' over V_0 such that each basic tree $S(V)$ can be obtained by pruning S' to contain only species from V, then the augmentation leads to a species tree with the zero cost, and the conditional problem is solved. The converse statement is also true.*

Proof. In the first statement additionally, intermediate trees also have zero costs.

If a tree T over V is obtained by pruning the tree S, then all basic trees pruned to V are also prunings of T, and, by Lemma 4, the tree T has the zero cost. Thus, the augmentation, in where all trees are prunings of S, is the desired process. Obviously, such the process exists. The converse statement follows from Lemma 4. □

2.11. Modification of Phase 1. If topologies of the initial trees G_j strongly contradict (an example is provided in [64]), then Phase 2 produces a tree with a nonzero cost; that is, according to Theorem 9, there exists a basic tree that cannot be obtained by pruning the output of Phase 2 to contain only species from the set V. This situation occurs because the basic trees are optimal in terms of the functional $c(V)$, not in terms of the more accurate total mapping cost.

Computer simulations suggest (data not shown) that Phase 2 performs more accurately in the below case. Let V be a fixed subset of V_0 and an element of P. Prune each initial gene tree G_j to V (denote the result $T_j : V$) and each element A from P to V(denote the result $A : V = A \cap V$; $P : V = \{A : V \mid A \text{runs over} P\}$). For a fixed V, apply Phase 1 to the sets $\{T_j : V \mid j\}$ and $P : V$. Let $T(V)$ be a basic tree over V, if such exists. Apply Phase 2 to the set $\{T(V) \mid V \text{ runs over} P\}$ and *denote* the result S^{**}. An analog of Theorem 9 is easily proved for S^{**} with Lemma 10 stated below. If a set V is basic for $\langle\{G_j\}, P\rangle$ and $\langle\{T_j : V \mid j\}, P : V\rangle$, the basic trees over V may be different.

Lemma 10. *If a set V is basic for $\langle\{G_j\}, P\rangle$, then it is basic for $\langle\{T_j : V \mid j\}, P : V\rangle$.*

Proof. Since V belongs to P, it also belongs to $P : V$. Use induction on the increase of $|V|$. Singleton sets are always basic. If V is a nonpruned basic set, it can be partitioned into two nonpruned basic subsets. By inductive assumption, the subsets are pruned basic. Then V is also pruned basic. □

The running time of modified Phase 1 is obviously $|P|$ times greater compared to standard Phase 1. For both versions of Phase 1, the complexity of Phase 2 has the order of $|P| \cdot |V_0|^5$, which is proved in [25].

2.12. Definitions of Binarization and Paralogous Binarization. Hereafter, only a canonic mapping α is considered and applied to polytomous trees (in the definition of α "for both children" is naturally replaced with the "for all children", refer to Section 2.6). Fix a polytomous gene tree G. Describe the procedure that starts from the initial G and iteratively derives G'. Let in this procedure a tree G' be already derived and possess a polytomous vertex g. Then arbitrarily divide the children of g with their incoming edges into two nonempty parts A and B, and for each part (with the corresponding subtrees) introduce an intercalating edge connecting a new vertex (the ancestor of this part) with g; if a part is a singleton, the corresponding new vertex is eliminated (none of the trees contains edges with one child). The tree G' so acquires two or one new vertices and keeps the ones inherited from G, and the vertex g becomes binary. The described operation is the *step of binarization* of vertex g against partition (A, B). Repeat the operation until all polytomous vertices are found. Name the obtained "resolved" tree G' a *candidate binarization* of G.

Fix a binary species tree S and the polytomous gene tree G. Among all candidate binarizations G' of G, find such $G^{\#} = G^{\#}(S)$ that has the minimal embedding cost among the values $c(G', S, \alpha)$ (G' is a variable); name $G^{\#}$ a *binarization* of G *against* S.

By definition, for given G and S, an edge e from G *enters* (downwards) a tube d in S if

$$\alpha(e^{+}) \geq d^{+} > d \geq \alpha(e_{+}) \tag{10}$$

and henceforth designated $e \downarrow d$.

For a vertex g from G or G' designate $d(g)$ a tube that equals $\alpha(g)$ (if $\alpha(g)$ is a tube) or the tube incoming in $\alpha(g)$ (if $\alpha(g)$ is a vertex). For each vertex from G, its clades in G and G' are equal. For g from G the tube $d(g)$ depends only on clade g in G; that is, $d(g)$ is the same in G and in G'. Note that the triple inequality above is equivalent to

$$d(e^{+}) > d \geq d(e_{+}). \tag{11}$$

A *paralogous binarization* $G^{\#\#}$ of G *against* S is a candidate binarization G', in which for each tube d the number of entering edges is minimal among all candidate binarizations G'. Intuitively, it minimizes the number of paralogs.

A *paralogous binarization* $G^{\#\#}$ of G exists and is produced from the initial G with the following iterative procedure. Let a certain G' be already obtained. Choose arbitrarily a polytomous vertex g in G', and let $d(g)$ produce two child tubes, d_1 and d_2. Divide all children g' of vertex g into three parts defined according to the conditions $d(g') = d(g)$, $d(g') \leq d_1$, $d(g') \leq d_2$, respectively. The parts are disjoint. If only the first part is nonempty, arbitrarily divide it in two nonempty sets. If the first and at least one of the other two parts are nonempty, the first set coincides with the first part, and the second set is the union of the second and third parts. If the first part is empty, the two sets are the second and third parts, correspondingly; both are nonempty by definition of $d(g)$. Perform a step of binarization of vertex g against partition (A, B), where A is the first set and B is the second set. A new G' is thus derived. Apply the procedure until all polytomous vertices are visited; the result, according to Lemma 11, is the paralogous binarization $G^{\#\#}$ of G against S.

A *bundle of edges* for d in G is a nonempty maximal on inclusion set of edges e in G that have the common upper terminus e^+ (the *vertex parent* of the bundle), and all *e enter d*.

Denote $p(G, d)$ the amount of bundles in G for d. The *vertex parent* of a bundle F is denoted by F^+. Obviously, a bundle has a unique vertex parent; and vertex parents of different bundles for d are different in G (and $G^{\#\#}$); edges of different bundles for d are incomparable in G.

A complement F' of bundle F is a set of edges e, for which $e^+ = F^+$ and e does not belong to F. For the paralogous binarization $G^{\#\#}$, an edge $e < F^+$ (where e and F^+ are in $G^{\#\#}$) is called a *parent* of bundle F in G for d, if e_+ is the last common ancestor of the lower termini of all edges in F and e_+ is not the ancestor of the lower termini of all edges in F'.

Lemma 11. *For any candidate binarization G' and mapping G' into S, at least $p(G, d)$ edges enter each tube d. For $G^{\#\#}$ (against S) and for each bundle (in G for d) and the mapping $G^{\#\#}$ into S, its parent in $G^{\#\#}$ exists and enters the tube d. Conversely, each edge entering tube d is the parent of a bundle for d. Consequently, $G^{\#\#}$ is a paralogous binarization.*

Proof. The first statement. In the mapping of $G\prime$ into S, each bundle in G for d induces at least one edge entering d; different bundles induce different edges. Indeed, let e be any edge in the bundle. Then on a path in $G\prime$ connecting e^+ and e_+, there exists an edge in $G\prime$ that enters d. As for any two bundles for d, their vertex parents are different; for any two corresponding paths in $G\prime$, the set of edges in one path does not intersect with the set of edges in another path. Consequently, at least $p(G, d)$ edges enter d.

The second statement. Let F be a bundle for tube d, and $g = F^+$. By definition of the bundle, $d(g) > d$, and for all lower termini g_i of edges in F, we observe $d(g_i) \leq d$. If the vertex g is binary or is the superroot, then $|F| = 1$, and the assertion is obvious. Let vertex g be polytomous. If $|F| = 1$, the assertion is obvious. Otherwise, consider in $G^{\#\#}$ a maximally long path L of vertices $g_1 = g, g_2, \ldots, g_k$, where each vertex descends directly from the other and is ancestor of the lower termini of all edges in the bundle F. Observe $d(g_k) \leq d$; otherwise during partitioning, the set of children of g_k in the constructed $G^{\#\#}$, all children from the bundle F belong in one part (the second or the third one), which contradicts the assumption of maximal L. It follows that the edge (g_{k-1}, g_k) enters the tube d and is the parent of the bundle F.

The third statement. Let $e \downarrow d$, where e is an edge in $G^{\#\#}$. By constructing the candidate binarization, there exists such vertex g in G that $g > e > g'$, where g' is a child of g in G. Consider the set of children g' of vertex g, for which $g' < e$ in $G^{\#\#}$. The edges in G having lower terminus in this set form a nonempty subset of a certain bundle F for d, where $F^+ = g$. Let e' be the bundle parent. Then e is comparable with e', and e' enters d according to Lemma 11. Any two comparable edges cannot enter the same tube; therefore $e = e'$. □

The described paralogous binarization procedure runs in linear time.

Lemma 12. *Let F_1 be a bundle for d_1, let F_2 be a bundle for d_2 (both in G), and $d_1^+ = d_2^+$. If $F_1^+ = F_2^+$, then in the paralogous binarization $G^{\#\#}$, the parents of F_1 and F_2 share the common upper terminus. And, conversely, if the parents of F_1 and F_2 share a common upper terminus in $G^{\#\#}$, then $F_1^+ = F_2^+$.*

Proof. Define with d a tube such that $d_+ = d_1^+ = d_2^+$, and with g a vertex $F_1^+ = F_2^+$. If g is a binary vertex in G, the assertion is obvious.

Otherwise, consider in $G^{\#\#}$ a maximally long path L of consecutive vertices $g_1 = g, g_2, \ldots, g_k$, where each g_i is an ancestor of a set C of all lower termini of edges in the union of F_1 and F_2. Observe $d(g_k) = d$; otherwise during partitioning the set of children of g_k in the constructed $G^{\#\#}$, all its children from C would belong in one part (the second or third one), which contradicts the assumption of maximal L.

The parents of F_1 and F_2 share a common upper terminus, as in the constructed $G^{\#\#}$ the bundles F_1 and F_2 correspond to the second and third parts of the g_k children set (the first part is empty, as follows from the assumption of maximal L). By the procedure, the parts are induced by separate edges, the parents of the corresponding bundles, and the two edges share the common upper terminus.

Prove the converse statement by contradiction. Denote the parents of bundles F_1 and F_2 as e_1 and e_2. Consider in $G^{\#\#}$ a path p_1 connecting F_1^+ with the lower terminus of an arbitrary edge from F_1 and a path p_2 connecting F_2^+ with the lower terminus of an arbitrary edge from F_2. Then p_1 contains e_1 and p_2 contains e_2. By our assumption, $F_1^+ \neq F_2^+$. Consequently, no two edges, one belonging to p_1 and the other belonging to p_2, can share a common upper terminus. The contradiction is obtained. □

The number of different $G^{\#\#}$ is exponential of the maximal number of edges e in G sharing the upper terminus e^+, for which $d(e_+) = d(e^+)$. Importantly, any $G^{\#\#}$ can be legitimately used in Section 2.13, according to Lemma 13. Our algorithm of constructing one $G^{\#\#}$ for any G can be easily extended to enumerate any portion of the binarization solutions space.

Lemma 13. *Embedding costs of all $G^{\#\#}$ against a fixed S are equal.*

Proof. Each paralogous binarization $G^{\#\#}$ possesses the same amount of vertices. Note that in a canonic mapping each edge e entering a tube d corresponds either to a divergence (if $\alpha(e^+) = d^+$) or loss (otherwise). Conversely, a divergence corresponds to a pair of edges with a common upper terminus entering the tubes a with common upper terminus, and a loss corresponds to edge e entering tube d, where $\alpha(e^+) \neq d^+$. Hence, draw two bijective correspondences:

(1) between divergences and unordered pairs $\{\langle e_1, d_1\rangle, \langle e_2, d_2\rangle\}$, where $i = 1, 2$, $e_i \downarrow d_i$, $e_1^+ = e_2^+$, and $d_1^+ = d_2^+$;

(2) between losses and pairs $\langle e, d\rangle$, where $e \downarrow d$, which do not fall in correspondence (1) with any divergence.

According to these correspondences and Lemmas 11-12, in a mapping of a paralogous binarization into S, there exist as many divergences as there are unordered pairs of bundles of the form $\langle$bundle F_1 for d_1, bundle F_2 for $d_2\rangle$ in G, where $d_1^+ = d_2^+$, $F_1^+ = F_2^+$. Other nonleaf vertices are duplications; therefore their amount does not depend on $G^{\#\#}$. The amount of losses is also $G^{\#\#}$-independent; according to the correspondences above and Lemma 11, there exist as many losses as there are bundles that do not fall in the correspondence with any divergence. □

Lemma 14. *In a paralogous binarization $G^{\#\#}$, let an edge e be a parent of a bindle F. The number of leaves contained in the clade of e_+ in $G^{\#\#}$ does not depend on $G^{\#\#}$.*

Proof. By definition of the bundle parent, the set of leaves contained in $G^{\#\#}$ below e is the set of leaves contained in G below the edges of bundle F. This set depends only on G and F. □

For canonic mappings of $G^{\#\#}$ (against S) into S, hold the following analogs of Lemmas 5–7.

Lemma 15. *In a canonic mapping of $G^{\#\#}$ into S, each leaf tube d of species s contains the amount of duplications equal to the difference between the total amount L of leaves below the edges of the bundles from G for d and the amount of all bundles for d in G.*

Proof. Denote this amount of duplications $\mathrm{Par}'(G, s)$. In a canonic mapping of $G^{\#\#}$ into S, the edges parental to the bundles for a leaf tube d enter the tube d, and all nonleaf vertices in the tree $G^{\#\#}$ lower to these edges are duplications. By Lemma 14, for each such bundle F, there are L leaf vertices lower to the parent of F. A binary tree contains $n - 1$ internal vertices compared with the number n of leaves; therefore, the number of duplications is also $n - 1$ of the number n of edges in a bundle. □

Lemma 16. *Fix a polytomous tree G over a subset of V_0. Let species trees S_1 and S_2 be both defined over V_0, each containing a certain subtree S. The total costs of all events in the mappings of $G^{\#\#}$ into S_1 and$G^{\#\#}$ into S_2, having place in S, are equal. In other words, the total cost depends only on the subtree S and not on its complement.*

Proof. In a canonic mapping of $G^{\#\#}$ into S_1 or S_2, the edges of tree $G^{\#\#}$, the parents of the bundles for the root tube d in a subtree S, enter the tube d. All vertices below these edges map into S. Conversely, if a vertex g maps into S, then on the path connecting it with the superroot, there exists an edge entering d and, by Lemma 11, being a parent of a bundle for d. If an edge e from $G^{\#\#}$ is parental to a bundle for d, then by Lemma 14 the set of leaves below e is defined by the bundle and does not depend on $G^{\#\#}$. The number of vertices in a binary subtree is determined by the number of leaves. Consequently, the amount of vertices mapped into S does not depend on $G^{\#\#}$. According to correspondence (1) stated in the proof of Lemma 1 and to Lemma 11, the amount of divergences in these vertices is exactly the amount of the unordered pairs $\langle$bundle F_1 for d_1, bundle F_2 for $d_2\rangle$ in G, where $F_1^+ = F_2^+$, $d_1^+ = d_2^+$, and d_1^+ lies in S. Other nonleaf vertices are duplications. According to correspondence (2) stated in the proof of Lemma 1 and to Lemma 11, the number of losses in S is also $G^{\#\#}$-independent and is exactly the number of bundles for tubes d, which do not fall in a correspondence with any divergence where d^+ lies in S. □

Lemma 17. *Fix a polytomous tree G over a subset of V_0. Let V be a subset of V_0, and a species tree S_1 over V_0 contains a subtree T_1 over V. Let a tree S_2 be derived from S_1 by substituting the subtree T_1 with a subtree T_2 over V. The total costs of all events in the mappings of $G^{\#\#}$ into S_1 and S_2 having place in the complements of T_1 in S_1 and T_2 in S_2 are equal. In other words, the total event cost does not depend on a subtree.*

Proof. The mapping α maps into each T_i vertices in a tree $G^{\#\#}$ that are below the edges parental to the bundles for the root tube d in T_i. By definition of the bundle, the set of such bundles depends only on the clade of d_+ that does not depend on index i. According to the argument in the proof of Lemma 16, the same amount of such vertices is mapped in each T_i, regardless of $G^{\#\#}$. Consequently, the complement of T_i receives the same amount of vertices. Among these vertices, the number of divergences equals exactly the number of the unordered pairs $\langle$bundle F_1 for d_1, bundle F_2 for $d_2\rangle$ in G, where $F_1^+ = F_2^+$, $d_1^+ = d_2^+$, and d_1^+ does not lie in S_i. Other nonleaf vertices are duplications. The number of losses in S is also $G^{\#\#}$-independent and is exactly the number of bundles for tubes d, which do not fall in a correspondence with any divergence where d^+ does not lie in T_i. □

2.13. The Second Problem for a Fixed Set of Polytomous Gene Trees. The general problem for a given set of polytomous gene trees G_j is to find a species tree $S^{\#}$ that minimizes the total sum (over binarizations $G_j^{\#}$ of all G_j) of the mappings of $G_j^{\#}$ into $S^{\#}$. The unconditional (absolute) problem imposes no constraint on the solution space. In the conditional problem, the search space of trees (including $S^{\#}$) is limited to the clades belonging to a prefixed parameter P; all clades from all G_j are included in P by default. The found binarization $G_j^{\#}$ may not be unique, but its choice does not affect the functional. The authors are only aware of an exponential complexity algorithm that solves both the unconditional and conditional problems. However, such complexity renders it of little use.

We formulate a simplification of the conditional problem; paralogous binarizations $G_j^{\#\#}$ are used instead of arbitrary candidate binarizations as described in Section 2.12.

A simplified problem is to construct a tree $S^{\#\#}$ (also containing clades from the set P) that minimizes the functional $c(\{G_j^{\#\#}(S)\}, S, \{f_j\})$, where $G_j^{\#\#}(S)$ is any paralogous binarizations of the initial trees G_j against S. By Lemma 13, the functional value is independent of the choice of $G_j^{\#\#}(S)$.

This $S^{\#\#}$ does not generally provide a global solution but can be useful, as paralogous binarizations are often biologically realistic.

Our solving algorithm for the simplified problem is similar to the case of binary trees and consists of two phases, with Phase 2 being identical. Phase 1 uses the same induction to build basic trees $S(V)$. The start of induction is identical to the binary case, with replacing $\sum_j \operatorname{Par}(G_j, s)$ to $\sum_j \operatorname{Par}'(G_j, s)$. In the induction step, the only difference with the binary case is the calculation of the cost of the events from the third group. By enumerating all vertices in all given G_j, compute the numbers of all bundles in $\{G_j\}$ for each d, d_1, and d_2 and denote those numbers n, n_1, and n_2, respectively. Analogously find the number k of pairs of all bundles of the form $\langle$bundle F_1 for d_1, bundle F_2 for $d_2\rangle$ in $\{G_j\}$ for each d_1 and d_2, where $F_1^+ = F_2^+$. Find values n_i and k in $\{G_j\}$, for which (i) the root clade intersects with both sets V_1 and V_2, and (ii) the root clade intersects with one of the sets and not with the other. Designate n_i', k' the numbers for (i), and n_i'', k'' the numbers for (ii).

Define

$$\begin{aligned} c(V, V_1, V_2) = {} & c(V_1) + c(V_2) \\ & + c_{\text{div}} \cdot k + c_{\text{dup}} \cdot (n_1 + n_2 - n - k) \\ & + c_{\text{los1}} \cdot (n_1' + n_2' - 2k') \\ & + c_{\text{los2}} \cdot (n_1'' + n_2'' - 2k''). \end{aligned} \tag{12}$$

Justification of the Algorithm. Let S be an arbitrary species tree over V_0 that includes a subtree $S(V)$. The total cost of events in $S(V)$ undercanonic mappings of all $G_j^{\#\#}$ (against S) into S is *designated* $c(V)$, analogously to the binary case. Obviously, if $V = V_0$, then $c(V_0) = c(\{G_j^{\#\#}\}, S, \{f_j\})$. According to Lemma 16, $c(V)$ also depends on the subtree $S(V)$ only and not on its complement (in the tree S). Theorem 18 is analogous to Theorem 8.

Theorem 18. *A basic tree $S(V)$ globally minimizes the functional $c(V)$ in the conditional problem for V, if a solution exists.*

Proof. For a singleton set V, the assertion of the theorem follows from Lemma 15.

According to Lemma 11, the number of edges entering a tube in a canonic mapping of $G^{\#\#}$ into S equals the number of the bundles in G for this tube. The mapping in d or r involves exactly the vertices of $G^{\#\#}$ that are both the descendants of one of the edges entering d and ancestors of at least one edge entering d_1 or d_2. Obviously, there are $n_1 + n_2 - n$ such vertices. Among them, the number of divergences (mappings in r) is exactly the number of pairs of the bundles $\langle$bundle F_1 for d_1, bundle F_2 for $d_2\rangle$ in G, where $F_1^+ = F_2^+$. The number of losses in r is exactly the number of bundles for d_1 or d_2 that do not fall in a correspondence with any divergence. Therefore, under k divergences, there exist $n_1 + n_2 - n - k$ duplications and $n_1 + n_2 - 2k$ losses. Consequently, the value $c_{\text{div}} \cdot k + c_{\text{dup}} \cdot (n_1 + n_2 - n - k) + c_{\text{los1}} \cdot (n_1' + n_2' - 2k') + c_{\text{los2}} \cdot (n_1'' + n_2'' - 2k'')$ is the total cost of events in the third group.

Further justification of the algorithm is identical to the binary case (considering Lemma 17). The remark to the proof of Theorem 8 and modification of Phase 1 (refer to Section 2.11) are still valid. □

The solution of the simplified conditional problem is obtained. The running complexity of the algorithm has the same order as specified in Theorem 8.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The subsection "Coalescent approaches" in Section 1.7 is written together with S. A. Pirogov, whose help was precious. The authors are indebted to the scientific editor (Professor W. H. Piel) and the reviewers for support and valuable comments. The work was partly funded by the Ministry for Education and Science of Russia (Grants 14.740.11.1053, 8481) and the Russian Foundation for Basic Research (Grant 13-04-40196-H).

References

[1] T. J. Treangen and E. P. C. Rocha, "Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes," *PLoS Genetics*, vol. 7, no. 1, Article ID e1001284, 2011.

[2] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Current Opinion in Microbiology*, vol. 11, no. 5, pp. 472–477, 2008.

[3] T. Dagan, Y. Artzy-Randrup, and W. Martin, "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 10039–10044, 2008.

[4] M. Hegarty, J. Coate, S. Sherman-Broyles, R. Abbott, S. Hiscock, and J. Doyle, "Lessons from natural and artificial polyploids in higher plants," *Cytogenetic and Genome Research*, vol. 140, no. 2–4, pp. 204–225, 2013.

[5] T. E. Wood, N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg, "The frequency of polyploid speciation in vascular plants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 33, pp. 13875–13879, 2009.

[6] J. E. Bowers, B. A. Chapman, J. Rong, and A. H. Paterson, "Unravelling angiosperm genome evolution by phylogenetic

analysis of chromosomal duplication events," *Nature*, vol. 422, no. 6930, pp. 433–438, 2003.

[7] K. S. Kassahn, V. T. Dang, S. J. Wilkins, A. C. Perkins, and M. A. Ragan, "Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates," *Genome Research*, vol. 19, no. 8, pp. 1404–1418, 2009.

[8] P. Dehal and J. L. Boore, "Two rounds of whole genome duplication in the ancestral vertebrate," *PLoS Biology*, vol. 3, no. 10, Article ID e314, 2005.

[9] O. Jatllon, J. M. Aury, F. Brunet et al., "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype," *Nature*, vol. 431, no. 7011, pp. 946–957, 2004.

[10] A. Christoffels, E. G. L. Koh, J. M. Chia, S. Brenner, S. Aparicio, and B. Venkatesh, "Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1146–1151, 2004.

[11] A. M. Altenhoff and C. Dessimoz, "Inferring orthology and paralogy," in *Evolutionary Genomics*, M. Anisimova, Ed., vol. 855 of *Methods in Molecular Biology*, chapter 9, pp. 259–279, Humana Press, 2012.

[12] A. Kuzniar, R. C. H. J. van Ham, S. Pongor, and J. A. M. Leunissen, "The quest for orthologs: finding the corresponding gene across genomes," *Trends in Genetics*, vol. 24, no. 11, pp. 539–551, 2008.

[13] M. S. Poptsova and J. P. Gogarten, "BranchClust: a phylogenetic algorithm for selecting gene families," *BMC Bioinformatics*, vol. 8, article 120, 2007.

[14] C. E. V. Storm and E. L. L. Sonnhammer, "Automated ortholog inference from phylogenetic trees and calculation of orthology reliability," *Bioinformatics*, vol. 18, no. 1, pp. 92–99, 2002.

[15] E. V. Koonin, "Orthologs, paralogs, and evolutionary genomics," *Annual Review of Genetics*, vol. 39, pp. 309–338, 2005.

[16] H. Mi, Q. Dong, A. Muruganujan, P. Gaudet, S. Lewis, and P. D. Thomas, "Panther version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D204–D210, 2009.

[17] B. Sennblad and J. Lagergren, "Probabilistic orthology analysis," *Systematic Biology*, vol. 58, no. 4, pp. 411–424, 2009.

[18] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney, "Ensemblcompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates," *Genome Research*, vol. 19, no. 2, pp. 327–335, 2009.

[19] L. A. David and E. J. Alm, "Rapid evolutionary innovation during an archaean genetic expansion," *Nature*, vol. 469, no. 7328, pp. 93–96, 2011.

[20] J. Ma, A. Ratan, B. J. Raney et al., "Dupcar: reconstructing contiguous ancestral regions with duplications," *Journal of Computational Biology*, vol. 15, no. 8, pp. 1007–1027, 2008.

[21] M. D. Rasmussen and M. Kellis, "A Bayesian approach for fast and accurate gene tree reconstruction," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 273–290, 2011.

[22] J. G. Burleigh, M. S. Bansal, O. Eulenstein, S. Hartmann, A. Wehe, and T. J. Vision, "Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees," *Systematic Biology*, vol. 60, no. 2, pp. 117–125, 2011.

[23] K. Y. Gorbunov and V. A. Lyubetsky, "Fast algorithm to reconstruct a species supertree from a set of protein trees," *Molecular Biology*, vol. 46, no. 1, pp. 161–167, 2012.

[24] M. Steel, S. Linz, D. H. Huson, and M. J. Sanderson, "Identifying a species tree subject to random lateral gene transfer," *Journal of Theoretical Biology*, vol. 322, pp. 81–93, 2013.

[25] V. A. Lyubetsky, L. I. Rubanov, L. Y. Rusin, and K. Yu. Gorbunov, "Cubic time algorithms of amalgamating gene trees and building evolutionary scenarios," *Biology Direct*, vol. 7, article 48, no. 1, pp. 1–20, 2012.

[26] G. J. Szöllösi, B. Boussau, S. S. Abby, E. Tannier, and V. Daubin, "Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 43, pp. 17513–17518, 2012.

[27] K. M. Swenson, A. Doroftei, and N. El-Mabrouk, "Gene tree correction for reconciliation and species tree inference," *Algorithms for Molecular Biology*, vol. 7, article 31, no. 1, 2012.

[28] D. Merkle, M. Middendorf, and N. Wieseke, "A parameter-adaptive dynamic programming approach for inferring cophylogenies," *BMC Bioinformatics*, vol. 11, supplement 1, article S60, 2010.

[29] C. Nieberding, E. Jousselin, and Y. Desdevises, "The use of co-phylogeographic patterns to predict the nature of interactions, and vice-versa," in *The Geography of Host-Parasite Interactiones*, S. Morand and B. Krasnov, Eds., Oxford University Press, New York, NY, USA, 2010.

[30] M. A. Charleston and S. L. Perkins, "Traversing the tangle: algorithms and applications for cophylogenetic studies," *Journal of Biomedical Informatics*, vol. 39, no. 1, pp. 62–71, 2006.

[31] D. R. Brooks and A. L. Ferrao, "The historical biogeography of co-evolution: emerging infectious diseases are evolutionary accidents waiting to happen," *Journal of Biogeography*, vol. 32, no. 8, pp. 1291–1299, 2005.

[32] R. Jothi, M. G. Kann, and T. M. Przytycka, "Predicting protein-protein interaction by searching evolutionary tree automorphism space," *Bioinformatics*, vol. 21, supplement 1, no. 1, pp. i241–i250, 2005.

[33] R. Guigó, I. Muchnik, and T. F. Smith, "Reconstruction of ancient molecular phylogeny," *Molecular Phylogenetics and Evolution*, vol. 6, no. 2, pp. 189–213, 1996.

[34] C. Chauve, J. P. Doyon, and N. El-Mabrouk, "Gene family evolution by duplication, speciation, and loss," *Journal of Computational Biology*, vol. 15, no. 8, pp. 1043–1062, 2008.

[35] D. Durand, B. V. Halldórsson, and B. Vernot, "A hybrid micro-macroevolutionary approach to gene tree reconstruction," *Journal of Computational Biology*, vol. 13, no. 2, pp. 320–335, 2006.

[36] P. Górecki and J. Tiuryn, "DLS-trees: a model of evolutionary scenarios," *Theoretical Computer Science*, vol. 359, no. 1–3, pp. 378–399, 2006.

[37] R. G. Beiko and N. Hamilton, "Phylogenetic identification of lateral genetic transfer events," *BMC Evolutionary Biology*, vol. 6, no. 1, article 15, p. 17, 2006.

[38] S. S. Abby, E. Tannier, M. Gouy, and V. Daubin, "Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests," *BMC Bioinformatics*, vol. 11, no. 1, article 324, 2010.

[39] A. Boc, H. Philippe, and V. Makarenkov, "Inferring and validating horizontal gene transfer events using bipartition dissimilarity," *Systematic Biology*, vol. 59, no. 2, pp. 195–211, 2010.

[40] T. Hill, K. J. Nordström, M. Thollesson et al., "Sprit: identifying horizontal gene transfer in rooted phylogenetic trees," *BMC Evolutionary Biology*, vol. 10, no. 1, article 42, 2010.

[41] G. Jin, L. Nakhleh, S. Snir, and T. Tuller, "Parsimony score of phylogenetic networks: hardness results and a linear-time

heuristic," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 3, pp. 495–505, 2009.

[42] Ö. Åkerborg, B. Sennblad, L. Arvestad, and J. Lagergren, "Simultaneous Bayesian gene tree reconstruction and reconciliation analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5714–5719, 2009.

[43] J. P. Doyon, C. Chauve, and S. Hamel, "Space of gene/species trees reconciliations and parsimonious models," *Journal of Computational Biology*, vol. 16, no. 10, pp. 1399–1418, 2009.

[44] D. Merkle and M. Middendorf, "Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information," *Theory in Biosciences*, vol. 123, no. 4, pp. 277–299, 2005.

[45] K. Y. Gorbunov and V. A. Lyubetsky, "Reconstructing the evolution of genes along the species tree," *Molecular Biology*, vol. 43, no. 5, pp. 881–893, 2009.

[46] R. Libeskind-Hadas and M. A. Charleston, "On the computational complexity of the reticulate cophylogeny reconstruction problem," *Journal of Computational Biology*, vol. 16, no. 1, pp. 105–117, 2009.

[47] P. Puigbo, Y. I. Wolf, and E. V. Koonin, "Seeing the tree of life behind the phylogenetic forest," *BMC Biology*, vol. 11, article 46, 2013.

[48] B. Robbertse, R. J. Yoder, A. Boyd, and J. Reeves, "Hal: an automated pipeline for phylogenetic analyses of genomic data," *PLoS Currents*, 2011.

[49] A. Dereeper, S. Audic, J. M. Claverie, and G. Blanc, "BLAST-EXPLORER helps you building datasets for phylogenetic analysis," *BMC Evolutionary Biology*, vol. 10, article 8, 2010.

[50] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, 2009.

[51] T. Frickey and A. N. Lupas, "Phylogenie: automated phylome generation and analysis," *Nucleic Acids Research*, vol. 32, no. 17, pp. 5231–5238, 2004.

[52] V. A. Lyubetsky, K. Yu. Gorbunov, L. Yu. Rusin, and V. V. V'yugin, "Algorithms to reconstruct evolutionary events at molecular level and infer species phylogeny," in *Bioinformatics of Genome Regulation and Structure II, Part 1*, N. Kolchanov, R. Hofestaedt, and L. Milanesi, Eds., pp. 189–204, Springer, New York, NY, USA, 2006.

[53] B. G. Mirkin, T. I. Fenner, M. Y. Galperin, and E. V. Koonin, "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes," *BMC Evolutionary Biology*, vol. 3, no. 1, article 2, 2003.

[54] A. Bolshoy and V. Kirzhner, "Algorithms of an optimal integer tree labeling," *BioMed Research International*, In press.

[55] D. A. Liberles, Ed., *Ancestral Sequence Reconstruction*, Oxford University Press, Oxford, UK, 2007.

[56] K. Y. Gorbunov and V. A. Lyubetsky, "Reconstruction of ancestral regulatory signals along a transcription factor tree," *Molecular Biology*, vol. 41, no. 5, pp. 836–842, 2007.

[57] K. Y. Gorbunov, O. N. Laikova, D. A. Rodionov, M. S. Gelfand, and V. A. Lyubetsky, "Evolution of regulatory motifs of bacterial transcription factors," *In Silico Biology*, vol. 10, article 0012, no. 3-4, pp. 163–183, 2010.

[58] K. Y. Gorbunov, E. V. Lyubetskaya, E. A. Asarin, and V. A. Lyubetsky, "Modeling evolution of the bacterial regulatory signals involving secondary structure," *Molecular Biology*, vol. 43, no. 3, pp. 485–499, 2009.

[59] K. Y. Gorbunov and V. A. Lyubetsky, "Identification of ancestral genes that introduce incongruence between protein- and species trees," *Molecular Biology*, vol. 39, no. 5, pp. 741–751, 2005.

[60] V. A. Lyubetsky, E. A. Zhizhina, and L. I. Rubanov, "Gibbs field approach for evolutionary analysis of regulatory signal of gene expression," *Problems of Information Transmission*, vol. 44, no. 4, pp. 333–351, 2008.

[61] T. Pupko, A. Doron-Figenboim, D. A. Liberles, and G. M. Cannarozzi, "Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences," in *Ancestral Sequence Reconstruction*, D. A. Liberles, Ed., chapter 4, Oxford University Press, Oxford, UK, 2007.

[62] H. Ashkenazy, O. Penn et al., "FastML: a web server for probabilistic reconstruction of ancestral sequences," *Nucleic Acids Research*, vol. 40, no. 1, pp. W580–W584, 2012.

[63] R. D. M. Page, "Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas," *Systematic Biology*, vol. 43, no. 1, pp. 58–77, 1994.

[64] K. Y. Gorbunov and V. A. Lyubetsky, "The tree nearest on average to a given set of trees," *Problems of Information Transmission*, vol. 47, no. 3, pp. 274–288, 2011.

[65] L. Zhang, "On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies," *Journal of Computational Biology*, vol. 4, no. 2, pp. 177–187, 1997.

[66] B. Ma, M. Li, and L. Zhang, "From gene trees to species trees," *SIAM Journal on Computing*, vol. 30, no. 3, pp. 729–752, 2000.

[67] A. C. Berglund-Sonnhammer, P. Steffansson, M. J. Betts, and D. A. Liberles, "Optimal gene trees from sequences and species trees using a soft interpretation of parsimony," *Journal of Molecular Evolution*, vol. 63, no. 2, pp. 240–250, 2006.

[68] T. H. Nguyen, J. P. Doyon, S. Pointet, A. M. A. Chifolleau, V. Ranwez, and V. Berry, "Accounting for gene tree uncertainties improves gene trees and reconciliation inference," in *Algorithms in Bioinformatics*, B. Raphael and J. Tang, Eds., vol. 7534 of *Lecture Notes in Computer Science*, pp. 123–134, Springer, Berlin, Germany, 2012.

[69] S. Bérard, C. Gallien, B. Boussau, G. J. Szöllösi, V. Daubin, and E. Tannier, "Evolution of gene neighborhoods within reconciled phylogenies," *Bioinformatics*, vol. 28, no. 18, pp. i382–i388, 2012.

[70] Y. Zheng, T. Wu, and L. Zhang, "Reconciliation of gene and species trees with polytomies," *Bioinformatics*, http://arxiv.org/abs/1201.3995.

[71] M. Lafond, K. M. Swenson, and N. El-Mabrouk, "An optimal reconciliation algorithm for gene trees with polytomies," in *Algorithms in Bioinformatics*, vol. 7534 of *Lecture Notes in Computer Science*, pp. 106–122, Springer, Berlin, Germany, 2012.

[72] B. Vernot, M. Stolzer, A. Goldman, and D. Durand, "Reconciliation with non-binary species trees," *Journal of Computational Biology*, vol. 15, no. 8, pp. 981–1006, 2008.

[73] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand, "Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees," *Bioinformatics*, vol. 28, no. 18, pp. i409–i415, 2012.

[74] B. Boussau and V. Daubin, "Genomes as documents of evolutionary history," *Trends in Ecology and Evolution*, vol. 25, no. 4, pp. 224–232, 2010.

[75] C.-W. Luo, M. C. Chen, Y. C. Chen, R. W. L. Yang, H. F. Liu, and K. M. Chao, "Linear-time algorithms for the multiple gene duplication problems," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 1, pp. 260–265, 2011.

[76] O. Eulenstein, S. Huzurbazar, and D. A. Liberles, "Reconciling phylogenetic trees," in *Evolution After Gene Duplication*, K. Dittmar and D. Liberles, Eds., chapter 10, pp. 185–206, Wiley-Blackwell, New York, NY, USA, 2010.

[77] K. V. Lopatovskaya, K. Yu. Gorbunov, L. Yu. Rusin, A. V. Seliverstov, and V. A. Lyubetsky, "The evolution of proline synthesis transcriptional regulation in gammaproteobacteria," *Moscow University Biological Sciences Bulletin*, vol. 65, no. 4, pp. 211–212, 2010.

[78] M. S. Bansal, E. J. Alm, and M. Kellis, "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss," *Bioinformatics*, vol. 28, no. 12, pp. i283–i291, 2012.

[79] K. Yu. Gorbunov and V. A. Lyubetsky, "An algorithm of reconciliation of gene and species trees and inferring gene duplications, losses and horizontal transfers," *Information Processes*, vol. 10, no. 2, pp. 140–144, 2010 (Russian).

[80] A. Tofigh, *Using trees to capture reticulate evolution, lateral gene transfers and cancer progression [Ph.D. thesis]*, KTH Royal Institute of Technology, 2009.

[81] Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas, "The cophylogeny reconstruction problem is NP-complete," *Journal of Computational Biology*, vol. 18, no. 1, pp. 59–65, 2011.

[82] A. Tofigh, M. Hallett, and J. Lagergren, "Simultaneous identification of duplications and lateral gene transfers," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 2, pp. 517–535, 2011.

[83] J.-P. Doyon, C. Scornavacca, K. Yu. Gorbunov, G. J. Szeollosi, V. Ranwez, and V. Berry, "An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers," in *Comparative Genomics*, S. Istrail, P. Pevzner, and M. Waterman, Eds., vol. 6398 of *Lecture Notes in Computer Science*, pp. 93–108, Springer, Berlin, Germany, 2010.

[84] L. A. David and E. J. Alm, "Rapid evolutionary innovation during an Archaean genetic expansion," *Nature*, vol. 469, no. 7328, pp. 93–96, 2011.

[85] V. A. Lyubetsky, K. Yu. Gorbunov, and L. Yu. Rusin, "Detecting conflicts in large sets of phylogenetic trees," in *Proceedings of the BioSyst.EU, Global Systematics Conference*, p. 131, Vienna, Austria, February 2013.

[86] B. Ma, M. Li, L. Zhang et al., "On reconcstructing species trees from gene trees in term of duplications and losses," in *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology (RECOMB '98)*, pp. 182–191, ACM, New York, NY, USA, 1998.

[87] B. Chor and S. Snir, "Analytic solutions of maximum likelihood on forks of four taxa," *Mathematical Biosciences*, vol. 208, no. 2, pp. 347–358, 2007.

[88] S. Snir and S. Rao, "Quartet MaxCut: a fast algorithm for amalgamating quartet trees," *Molecular Phylogenetics and Evolution*, vol. 62, no. 1, pp. 1–8, 2012.

[89] D. G. Brown and J. Truszkowski, "Fast error-tolerant quartet phylogeny algorithms," *Theoretical Computer Science*, vol. 483, pp. 104–114, 2013.

[90] M. S. Bansal, J. G. Burleigh, O. Eulenstein, and D. Fernández-Baca, "Robinson-Foulds supertrees," *Algorithms for Molecular Biology*, vol. 5, no. 1, article 18, 2010.

[91] N. Nguyen, S. Mirarab, and T. Warnow, "MRL and SuperFine+MRL: new supertree methods," *Algorithms for Molecular Biology*, vol. 7, article 3, 2012.

[92] B. Roure, D. Baurain, and H. Philippe, "Impact of missing data on phylogenies inferred from empirical phylogenomic data sets," *Molecular Biology and Evolution*, vol. 30, no. 1, pp. 197–214, 2013.

[93] S. Buerki, F. Forest, N. Salamin, and N. Alvarez, "Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study," *Systematic Biology*, vol. 60, no. 1, pp. 32–44, 2011.

[94] M. P. Simmons, "Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data," *Molecular Phylogenetics and Evolution*, vol. 62, no. 1, pp. 472–484, 2012.

[95] M. P. Simmons, "Misleading results of likelihood-based phylogenetic analyses in the presence of missing data," *Cladistics*, vol. 28, no. 2, pp. 208–222, 2012.

[96] M. Remm, C. E. V. Storm, and E. L. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1041–1052, 2001.

[97] D. L. Fulton, Y. Y. Li, M. R. Laird, B. G. S. Horsman, F. M. Roche, and F. S. L. Brinkman, "Improving the specificity of high-throughput ortholog prediction," *BMC Bioinformatics*, vol. 7, no. 1, article 270, 2006.

[98] D. P. Wall, H. B. Fraser, and A. E. Hirsh, "Detecting putative orthologs," *Bioinformatics*, vol. 19, no. 13, pp. 1710–1711, 2003.

[99] A. C. J. Roth, G. H. Gonnet, and C. Dessimoz, "Algorithm of OMA for large-scale orthology inference," *BMC Bioinformatics*, vol. 10, no. 1, article 220, 2009.

[100] V. A. Lyubetsky, A. V. Seliverstov, and O. A. Zverkov, "Construction of homologous plastid-encoded protein families separating paralogs in the Magnoliophyta," *Mathematical Biology and Bioinformatics*, vol. 8, no. 1, pp. 225–233, 2013 (Russian).

[101] E. S. Allman, J. H. Degnan, and J. A. Rhodes, "Determining species tree topologies from clade probabilities under the coalescent," *Journal of Theoretical Biology*, vol. 289, no. 1, pp. 96–106, 2011.

[102] S. Roch and S. Snir, "Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis," *Journal of Computational Biology*, vol. 20, no. 2, pp. 93–112, 2013.

[103] P. Górecki, G. J. Burleigh, and O. Eulenstein, "Maximum likelihood models and algorithms for gene tree evolution with duplications and losses," *BMC Bioinformatics*, vol. 12, supplement 1, article S15, 2011.

[104] J. P. Doyon, S. Hamel, and C. Chauve, "An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 26–39, 2012.

[105] J. Sjostrand, B. Sennblad, L. Arvestad, and J. Lagergren, "DLRS: gene tree evolution in light of a species tree," *Bioinformatics*, vol. 28, no. 22, pp. 2994–2995, 2012.

[106] L. Arvestad, A. C. Berglund, J. Lagergren, and B. Sennblad, "Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution," in *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB ' 04)*, pp. 326–335, ACM, March 2004.

[107] B. Boussau, G. J. Szöllösi, L. Duret, M. Gouy, E. Tannier, and V. Daubin, "Genome-scale coestimation of species and gene trees," *Genome Research*, vol. 23, no. 2, pp. 323–330, 2013.

[108] M. D. Rasmussen and M. Kellis, "Unified modeling of gene duplication, loss, and coalescence using a locus tree," *Genome Research*, vol. 22, no. 4, pp. 755–765, 2012.

[109] J. H. Degnan and L. A. Salter, "Gene tree distributions under the coalescent process," *Evolution*, vol. 59, no. 1, pp. 24–37, 2005.

[110] A. N. Kolmogorov, "Zur Umrehrbarkeit der statistischen Naturgesedze," *Mathematische Annalen*, vol. 113, no. 1, pp. 766–772, 1937.

[111] J. F. C. Kingman, "On the genealogy of large populations," *Journal of Applied Probability*, vol. 19, pp. 27–43, 1982.

[112] S. Wright, "Evolution in Mendelian populations," *Genetics*, vol. 16, pp. 97–159, 1931.

[113] R. A. Fisher, *The Genetical Theory of Natural Selection*, Oxford University Press, New York, NY, USA, 1st edition, 1930.

[114] C. Wiuf and P. Donnelly, "Conditional genealogies and the age of a neutral mutant," *Theoretical Population Biology*, vol. 56, no. 2, pp. 183–201, 1999.

[115] T. Wu and L. Zhang, "Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree," *BMC Bioinformatics*, vol. 12, supplement 9, article S7, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 857329, 11 pages
http://dx.doi.org/10.1155/2014/857329

Research Article

Differences in Brain Transcriptomes of Closely Related Baikal Coregonid Species

Oksana S. Bychenko,[1] Lyubov V. Sukhanova,[2] Tatyana L. Azhikina,[1] Timofey A. Skvortsov,[1] Tuyana V. Belomestnykh,[2] and Eugene D. Sverdlov[1]

[1] *M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Miklukho-Maklaya Street, 16/10, V-437, 117997 Moscow, Russia*
[2] *Limnological Institute, Russian Academy of Sciences, Ulan-Batorskaya 3, 664033 Irkutsk, Russia*

Correspondence should be addressed to Lyubov V. Sukhanova; lsukhanova@yandex.ru

Received 1 April 2013; Revised 2 October 2013; Accepted 28 October 2013; Published 29 January 2014

Academic Editor: Vassily Lyubetsky

The aim of this work was to get deeper insight into genetic factors involved in the adaptive divergence of closely related species, specifically two representatives of Baikal coregonids—Baikal whitefish (*Coregonus baicalensis* Dybowski) and Baikal omul (*Coregonus migratorius* Georgi)—that diverged from a common ancestor as recently as 10–20 thousand years ago. Using the Serial Analysis of Gene Expression method, we obtained libraries of short representative cDNA sequences (tags) from the brains of Baikal whitefish and omul. A comparative analysis of the libraries revealed quantitative differences among ~4% tags of the fishes under study. Based on the similarity of these tags with cDNA of known organisms, we identified candidate genes taking part in adaptive divergence. The most important candidate genes related to the adaptation of Baikal whitefish and Baikal omul, identified in this work, belong to the genes of cell metabolism, nervous and immune systems, protein synthesis, and regulatory genes as well as to DTSsa4 Tc1-like transposons which are widespread among fishes.

1. Introduction

Adaptation to different environmental conditions through filling available ecological niches may lead to population divergence and subsequently to the emergence of new species. Recently diverged populations are appropriate to study the genetic basis of primary evolutionary changes [1].

Having the unusual propensity for rapid speciation and adaptive radiation, Coregonidaeare becoming a model system for studying the genomic basis of adaptive divergence and reproductive isolation [2, 3]. Of particular relevance is the occurrence of both in North America (*Coregonus clupeaformis* complex) and Eurasia (*Coregonus lavaretus* complex) of lacustrine forms of whitefish that live in sympatry [3]. Phylogeographic studies confirmed the young age of these sympatric forms that evolved postglacially, less than 15,000 yr BP ([4, 5] and reviewed in [3]). For example, the limnetic dwarf ecotype of lake whitefish, *Coregonus clupeaformis,* derived in from the ancestral benthic normal ecotype evolved in parallel and independently in several postglacial lakes [6, 7]. Moreover, recent extensive gene expression studies have shown that, for some genes, this parallel phenotypic evolution of whitefish morphs is accompanied by parallelism in expression of the genes potentially underlying phenotypic divergence [1, 2, 8].

The subjects of our investigation were Baikal (lacustrian) whitefish *Coregonus baicalensis* Dybowski (bathypelagic bentophage) and Baikal omul *C. migratorius* Georgi (pelagic planctophage). The genomic sequences of the two organisms are still unavailable, thus excluding direct nucleotide-by-nucleotide comparison. However, analyses of Palearctic and Nearctic whitefishes [9, 10] as well as genealogical reconstruction of Baikal whitefishes [11, 12] showed that Baikal whitefish and omul are sister taxa and that they diverged from a common ancestor just 10–20 thousand years ago, probably due to the occupation of different trophic niches.

Baikal whitefish and omul arose after the last cool period (Sartanian glaciation, 34–10 thousand years ago) and represent one more case of sympatric postglacial whitefish divergence into pelagic and benthic niches [12]. However, their ancestor has diverged from all other coregonid fishes, including Baikal lacustrine-riverine whitefish, *Corgonus pidschian*, about 1.5 million years ago, when simultaneous speciation (or fast cladogenesis) took place in the south of East Siberia [13, 14]. This event gave rise to the several main clades of true whitefishes, among them the *C. clupeaformis* complex, the *C. lavaretus* complex, and all Lake Baikal coregonid fishes. The Baikal whitefish/omul pair is the only representative of its own clade, one of the main clades in the cluster of true whitefishes. The origin of the Baikal whitefish and omul ancestral form is likely Lake Baikal itself. This means that Baikal whitefish and the omul ancestral form inhabited Lake Baikal at least from the time of the formation of the true whitefishes group, moreover probably before the appearance of the genus *Coregonus* [14]. From at least the beginning of the Pleistocene, Lake Baikal has become not only large and deep but also oligotrophic and oxygenated water body inhabitable for higher animals [15]. It was hypothesized that isolation of pelagic and benthic forms within Lake Baikal was caused many times by Pleistocene climatic oscillations during 1.5 million years. Thus, Lake Baikal is the only place where true whitefish sympatric ecological divergence has been replicated many times within the same water body over such a long period of time. Another peculiarity consists in the multilevel pattern of intraspecific phenotypic divergence, especially pronounced in the pelagic form. Such multilevel structure is determined by availability of multiple ecological niches in the large, deep oligotrophic lake with a highly structured water body [13].

A whole genome comparison of Baikal whitefish and omul using subtractive hybridization did not reveal any difference between the two genomes [16]. All differential fragments found had a polymorphic character differed by single oligonucleotide substitutions or short (up to 35 nucleotides) insertions or deletions (indels). They belonged mostly to noncoding genomic regions: introns, and micro- and minisatellites as well as transposon-like structures but there was no divergence in protein coding genes detected. This result was additional evidence of the close relationship of Baikal whitefish and omul and might suggest that although genetic differences between coregonids at early stages of evolution are minimal in terms of genomic DNA sequence, they can still affect expression of some genes through, for example, mutations or heritable epigenetic modifications in *cis*-regulatory sequences. Recent analyses of transcription in closely related organisms using microarray technology confirmed that changes in expression of metabolic and regulatory genes might have a much larger impact on morphological traits than changes in structural genes [17, 18]. These changes may be further followed by large-scale alterations in genomic DNA [19].

It has been repeatedly reported that variations in neuron transcriptomes directly or indirectly correlate with behavioural differences between organisms [20, 21]. Therefore, by comparing the sets of genes transcribed in the brain of recently diverged fishes, one can hope to reveal candidate genes involved in adaptive divergence of populations [1].

Using the SAGE method (Serial Analysis of Gene Expression) [22], we performed a comparative study of Baikal whitefish and omul brain transcriptomes and found quantitative differences between them. Most of these differences belonged to the genes of cell metabolism, nervous and immune systems, protein synthesis, and regulatory genes as well as to Tc1-like transposons of the DTSsa4 family, which are widespread among fishes. Some of the differences revealed here may contribute to the phenotypic divergence of the two populations. The presence of parallelisms in phenotypic adaptation of Lake Baikal and other whitefishes toward the use of the pelagic niche accompanied by parallelism in differential pattern of expression is discussed.

2. Methods

2.1. Plasmid DNA. Growth and transformation of *E. coli* cells, preparation of plasmid DNA, agarose gel electrophoresis, and other nucleic acid manipulations were performed according to standard protocols [23] or to the recommendations of the manufacturers. The cells for plasmid extraction were grown overnight at 37°C in 5 mL of Luria-Bertani medium supplemented with ampicillin (0.1 mg/mL). Plasmid DNA was isolated using a Wizard Plus Miniprep DNA Purification System (Promega) according to the manufacturer's recommendations. Clone inserts were sequenced with an Amersham Biosciences/Molecular Dynamics MegaBACE 4000 Capillary Sequencer.

2.2. Oligonucleotides. Oligonucleotides were synthesized using an ASM-102U DNA synthesizer (Biosset Ltd., Russia). The primers specific for differential sequences were designed using the Primer3 software (http://bioinfo.ut.ee/primer3-0.4.0/).

2.3. Sample Collection. Live immature adult specimens of each Baikal whitefish and omul were collected by gill nets in August 2005 in the Maloye More strait region of Lake Baikal. The fishes were assigned to the two species according to the main diagnostic characteristics (i.e., counting the gill raker number on the first left gill arch and evaluation of the mouth position). Baikal lacustrine whitefish from the Maloye More region is a benthophage that has a subterminal mouth and 25–33 gill rakers on the first gill arch [23]. Omul is a planktophage with a terminal mouth and 37–51 gill rakers [23]. Whitefish individuals of our sample had a typical subterminal mouth, and the number of gill rakers varied from 25 to 31 (28 on average). As for omul individuals, they had a typical terminal mouth, and the number of gill rakers varied from 40 to 49 (44 on average). Six adult individuals were randomly collected in each species. Mean fork length and body mass for Baikal omul were 27.5 cm (SD = 3.1 cm) and 235 g (SD = 21 g), respectively. Mean fork length and body mass for Baikal whitefish were 34.5 cm (SD = 2.9 cm) and 489 g (SD = 28.5 g). Fishes were euthanized with 0.001% clove oil. Brain tissue samples were frozen immediately in liquid nitrogen and stored at −80°C. One specimen of each species was used for SAGE. Five specimens of each species were used for real-time PCR estimation of whitefish and omul brain cDNA transcription levels.

2.4. RNA Isolation and cDNA Synthesis. Total RNA was isolated from each sample of brain tissues using an SV Total RNA Isolation System (Promega) according to the manufacturer's recommendations. All RNA samples were treated with DNaseI to remove residual DNA. cDNA synthesis was performed using random hexamer primers with (RT+) or without (RT−) addition of PowerScript II reverse transcriptase (Clontech). The hexamer primers (12 pmol) were annealed in 11 μL of a mixture containing 2 μg total RNA. The mixture was heated for 2 min at 70°C and then chilled on ice for 10 min. To synthesize cDNA, the RT+ and RT− reaction mixtures were incubated at 37°C for 10 min and then at 42°C for 120 min.

2.5. Serial Analysis of Gene Expression. SAGE [22] was performed with an I-SAGE Long kit (Invitrogen) according to the manufacturer's protocol, starting with 100 μg of total RNA. First, RNA samples were isolated from brain tissues of whitefish and omul (one sample for each species). The isolated RNA was incubated with magnetic beads (Dynabeads) and used to synthesize double-stranded cDNA according to a standard protocol. The cDNAs obtained were treated with NlaIII (NEB) restriction enzyme, and $3'$ cDNA fragments were separated from other restriction fragments. The selected fragments were divided into two parts and ligated to two different adapters, each containing a MmeI restriction site. The ligated fragments were then treated with MmeI restrictase (NEB) to produce tags; the two mixtures were pooled together, and the tags were ligated to each other to give ditags. The ditags were PCR amplified, adapters were removed by digestion with NlaIII and separated from ditags in a 12% polyacrylamide gel, and the purified ditags were ligated to form linear concatemers.

The obtained concatemers, containing abundant representative cDNA fragments, were ligated into the pZErO-1 vector (Invitrogen) and cloned into *E. coli* cells. Independent recombinant clones were used to generate clone libraries arrayed in 96-well plates. Clones for sequencing were selected based on PCR screening.

2.6. Identification of cDNAs Corresponding to Tags Found by SAGE. To find the cDNA of a representative sequence (tag) from the SAGE libraries, double-stranded cDNA of the whitefish and omul brain (of the same two individuals that were used for SAGE, 3 μg each) was digested overnight with PvuII restriction enzyme (100 U, Fermentas) in a volume of 300 μL. The restricted cDNA was purified using a QIAquick PCR Purification Kit (Qiagen) and then ligated to the oligonucleotide adapters T7NotRsa (Table 1A) obtained by annealing the corresponding oligonucleotide and a short template (800 pmol each) in TM buffer (10 mmol L^{-1} Tris-HCl, pH 7.8, 10 mmol L^{-1} $\mathrm{MgCl_2}$). The ligation was performed overnight at 16°C using T4 phage DNA ligase (Promega). Then, two rounds of PCR were performed using the external T7 (round 1) and internal NotRsa (round 2) adapter primers (Table 1A), with the second primer being a tag oligonucleotide selected from the SAGE library. The PCR products were cloned and sequenced, and the sequences obtained were further used to design primers for real-time PCR.

2.7. Real-Time PCR. Real-time PCR was performed on all specimens of both species using an EVA green RT-PCR kit (Sintol, Russia) with 10 ng of the total cDNA samples as templates. Primers (Table 1B) were used at final concentrations of 0.2 μmol L^{-1} each. The reactions were performed in a 25 μL volume as follows: preincubation at 95°C (10 min) and then 40 cycles of 95°C for 30 s, 63°C for 30 s, and 72°C for 40 s. All real-time experiments were repeated in triplicate. The expression level of each gene was normalized using *GAPDH* as a reference gene. *GAPDH* oligonucleotide primers were designed against cDNA sequences annotated in GenBank records. At the end of the amplification, a dissociation curve was plotted to confirm the specificity of the product. To exclude contamination by genomic DNA, RT-experiments were done in parallel. The amplification efficiency of each primer set was determined using LinRegPCR [24]. The results were processed using the lin_reg_psr.exe and REST 2005 software [25].

2.8. Computer and Statistical Analysis. The libraries obtained for whitefish and omul were analysed using RIDGES software [26] as well as BlastAll and FormatDB programs from the BLAST software (NCBI, USA) [27]. RIDGES was used to extract tags from concatamer sequencing data and to calculate their amounts in libraries. 250 clones from each SAGE library were sequenced; 1894 and 2670 tags were identified in the whitefish and omul libraries, respectively (see Web Supplementary File Appendix 1) (Supplementary Material available online at http://dx.doi.org/10.1155/2014/857329). As the sequencing depth was not high enough to cover all transcripts, only the tags originating from the genes with the highest levels of expression were discovered in this way.

We found differentially expressed tags using online-based versions (Statistical Analysis of Transcript profiles, http://www.igs.cnrs-mrs.fr/spip.php?article168&lang=fr, and IDEG6, http://telethon.bio.unipd.it/bioinfo/IDEG6_form/) of the statistical test developed by Audic and Claverie [28]. This test gives the conditional probability of observing y number of tags in library B, given that x tags have been observed in libraries A, if N_A and N_B are the total number of tags for, respectively, library A and B, under the assumption that the null hypothesis is true and the null hypothesis is that the tag is expressed equally in both the conditions A and B. This method has been successfully used in several experiments to analyse data sets obtained by SAGE and RNA-Seq methods.

Due to the insufficient depth of sequencing, only 36 tags (see Web Supplementary File Appendix 2) were found to be differentially expressed between whitefish and omul, while the differences in the numbers of another 2732 unique tags were too small to be statistically significant. This fact also prevented our attempts to perform multiple testing corrections. Therefore, in order to confirm the obtained

Table 1: Oligonucleotide primers and adapters. (a) Oligonucleotides used to find the cDNAs corresponding to tags. (b) Oligonucleotides used in real-time PCR.

(a)

Name of the oligonucleotides	Oligonucleotides sequence (5′-3′)	Annealing temperature (°C)
External primer T7	CTAATACGACTCACTATAGGGC	66
Internal primer NotRsa	AGCGTGGTCGCGGCCGAGGT	68
GAPDH for.	GTGGACGGCCCCTCTGC	63
GAPDH rev.	TCTGGTGGGCACCACGG	
Suppression adapter (equimolar mixture of preliminary annealed oligonucleotides T7NotRsa and Rsa_10)	T7NotRsa CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGT Rsa_10 ACCTGCCCGG	
Tag no. 89	GTACTTTAATGGATGATCTCC	56
Tag no. 81	GTACAAAAAGACTGCTGTTCC	60

(b)

Name of the oligonucleotides	Oligonucleotides sequence (5′-3′)	Annealing temperature (°C)
GAPDH for.	GTGGACGGCCCCTCTGC	63
GAPDH rev.	TCTGGTGGGCACCACGG	
1ep for.	TGGTGTAGGTCTCCTGGAC	63
1ep rev.	CTCCACCTATGAGGACCAG	
2net for.	CAGCTCCTCCAGACGCACC	63
2net rev.	CGCTGGAGCTGTTCGACAAC	
3st for.	TGGAGGAGAAGATAGTGGACTTGT	63
3st rev.	AGTAATCTGACTTTGTTGGTGAACT	
4fgf for.	CGGTACAGCGTCGATGAGTAG	63
4fgf rev.	CACCATAAATGGGACCAAGG	
5tnf for.	CATTGCAGTCCTAGTCTCTCT	63
5tnf rev.	AGTGGTATCAACGCAGAGTAC	
6tr for.	CAGGACAATGACCCAACACAC	63
6tr rev.	TGAGGTCGGGGGATTGTG	
7deh for.	GAGGAGGTATTTAAAGAATCGG	63
7deh rev.	CCTCAGCTTTAAATTTGACCAC	
8atp for.	GGCAGGTGAACTCCACAATC	63
8atp rev.	CAGGAGTGCTGGAATCAAGG	

results, eight of the 36 differentially expressed genes were selected for qPCR instead of performing multiple testing corrections.

3. Results

3.1. Baikal Whitefish and Omul Brain SAGE Libraries. The transcriptomes of the Baikal coregonids were compared by the SAGE method [22], an experimental technique designed to gain a direct and quantitative measure of gene expression. The SAGE method is based on the isolation of unique sequence tags (21 bp in length) of cDNAs followed by their concatenation serially and sequencing of the resulting long DNA molecules. It can give not only nucleotide sequences of many short fragments at a time and qualitative composition of transcriptomes, but also quantitative correlation between transcriptome components.

We obtained and characterized two libraries of short representative cDNA sequences (tags) from the brains of Baikal whitefish and omul (one library for one individual of each species). 250 clones from each library were sequenced that gave 1894 and 2670 tags for whitefish and omul, respectively (listed in Web Supplementary File Appendices 1 and 2).

A comparative quantitative characteristic of the whitefish tag library is shown in Table 2. The data obtained showed that only 3.9% of tags fall within the range of significance. The sequences of these “significant” cDNA fragments were compared to cDNA sequences of different organisms annotated in available databases. The results for tags with 70% to 100% identity level to known genes are presented in the Web Supplementary File Appendix 3. The table shows that most of tags that are more represented in one of the two libraries (relative value between 2 and 3, predominantly whitefish tags) are similar to segments of protein synthesis (ribosomal proteins 40S, S5, S11 and 60S L7, L13a, L39, L15, etc.) and regulatory (*Sox9a2, Cdc23, tnfsf5ip1, GRB10, btf2p44, PRKA*, etc.) genes. In contrast, the tags with a relative value between

Table 2: A comparative quantitative characteristic of the whitefish tag library.

Representation of a given tags in the whitefish library relative to omul ($\mathrm{Dif}_{\mathrm{wh}}$)	Relative content in the whitefish library, %
2.7–3.7	2.0
1.7–2.6	1.2
0.6–1.6	96.1*
0.3–0.5	0.7
In total	3.9

*Tags which turned outside the range of significance. We considered as statistically significant differences between whitefish and omul libraries in the range $0.6 > \mathrm{Dif}_{\mathrm{wh}} > 1.6$ that corresponds to the statistical significance level $P > 0.999$. $\mathrm{Dif}_{\mathrm{wh}}$ is the relative content of a given tag in the whitefish library (relative to omul), $\mathrm{Dif}_{\mathrm{wh}} = (N_{\mathrm{wh}}/N_{\mathrm{om}}) \times 1.41$, N_{wh} and N_{om} are the numbers of the given tag in the libraries of Baikal whitefish and omul, respectively, and 1.41 is the correction coefficient equal to the ratio 2670/1894.

0.5 and 0.3 (predominantly omul tags) are mostly similar to segments of metabolism genes (ATP synthase, H+ transporting, mitochondrial F0 complex, subunit b, and isoform 1; carbonic anhydrase 5B, Cytochrome P450 (CYP94C7); Na/K ATPase, alpha subunit, isoform 1c, etc.). However, a number of tags corresponding to cDNAs of metabolism genes were better represented in whitefish, while cDNA of 60S ribosomal protein L31 (clone ssal-rgh-513-300, *Salmo salar*) was similar to tags prevailing in omul.

Selected genes of the nervous and immune systems are more frequent in the whitefish library and others in the omul library (see Web Supplementary File Appendix 3). A number of tags more frequent in the whitefish library are similar to DTSsa4 Tc1-like DNA transposons.

For several of these apparently significant tags, we also determined sequences of the corresponding cDNAs and then confirmed the differences' reliability by real-time PCR.

3.2. Sequencing of cDNAs Corresponding to Tags. Available databases contain no data on Baikal whitefish and omul cDNA sequences, so we selected some whitefish tags to partially sequence the corresponding cDNAs (the experiment schematic is shown in Figure 1). The obtained nucleotide sequences were compared to known cDNAs of different organisms. They were found to be highly (70–90%) similar to cDNA segments of such genes as fibroblast growth factor 12 (*Danio rerio*), ependymin-related protein 1 (*Danio rerio*), Na/K ATPase alpha subunit isoform 1c mRNA (*Oncorhynchus mykiss*), and other genes (Table 3).

3.3. Real-Time PCR Estimation of Whitefish and Omul Brain cDNA Transcription Levels. Real-time PCR of each cDNA fragment was performed with unique oligonucleotide primers. cDNA of those whitefish and omul individuals for which SAGE was carried out was used as a template. The PCR results confirmed that the expression of genes of the nervous (similar to netrin-G1 ligand, ependymin-related protein 1) and immune (tumour necrosis factor receptor superfamily, member 9) system as well as of the regulatory fibroblast growth factor 12 gene was two- to threefold enhanced in the brain of whitefish compared with that of omul.

The expression levels were also confirmed for the metabolism genes of NADH dehydrogenase subunit 5 and Na/K ATPase (higher in the whitefish brain) and alpha subunit isoform 1c (higher in omul brain).

To prove the identities of the amplified cDNA fragments, the products of real-time PCR were ligated into a vector, cloned, and sequenced (10 clones for each fish). For each analysed cDNA, the sequences obtained were at least 99% similar to each other. The differences were only in single clones and rare single-nucleotide substitutions did not affect the reading frames. This might be due to either DNA polymerase errors during amplification or to allelic variants of identical genes.

To exclude the possibility that the differences revealed resulted from individual polymorphisms in the expression level, we performed the real-time PCR analysis of the expression levels of the genes using cDNAs prepared from five individuals of each species (Figure 2).

4. Discussion

Data on genomic sequence and transcriptome comparison for the two important inhabitants of the Lake Baikal are currently unavailable. Earlier, we performed a whole-genome comparison using subtractive hybridization but were unable to detect any species-specific differences between Baikal whitefish and omul within the accuracy of the technique [16]. This result confirmed the close similarity of the two genomes, in accordance with the very recent divergence of the two whitefish species. However, it did not exclude the existence of undetected minor genetic or epigenetic differences differently affecting the level of gene expression.

The Baikal whitefish and omul pair represents one more case of multiple divergent true whitefish morphs, which evolved within lakes postglacially, less than 15,000 yr BP [5]. In all studied North American and Eurasian lakes, such sympatric morphs differ in the number of gill rakers, a highly heritable trait related to trophic utilization. Individual growth rate, age and size at maturity, diet, and habitat use also differ between morphs within lakes, but are remarkably similar across lakes within the same morph. Most of sympatric morphs are genetically different, and similar morphs from different lakes likely have a polyphyletic origin. These results are most compatible with the process of parallel evolution through recurrent postglacial divergence into pelagic and benthic niches in each of these lakes [3]. Moreover, during the last few years, gene expression studies have shown, for some genes, that this parallel phenotypic evolution of whitefish morphs accompanied by parallelism in expression of the genes potentially underlies phenotypic divergence [1, 2, 8, 29, 30].

The life history and ecology of the Lake Baikal sympatric pair have also been well documented [13, 31]. A similar pattern of phenotypic character displacement has contributed to the evolution of the Lake Baikal pelagic and benthic ecotypes. Is the presence of parallelism in phenotypic adaptation of Lake Baikal and other whitefishes toward the

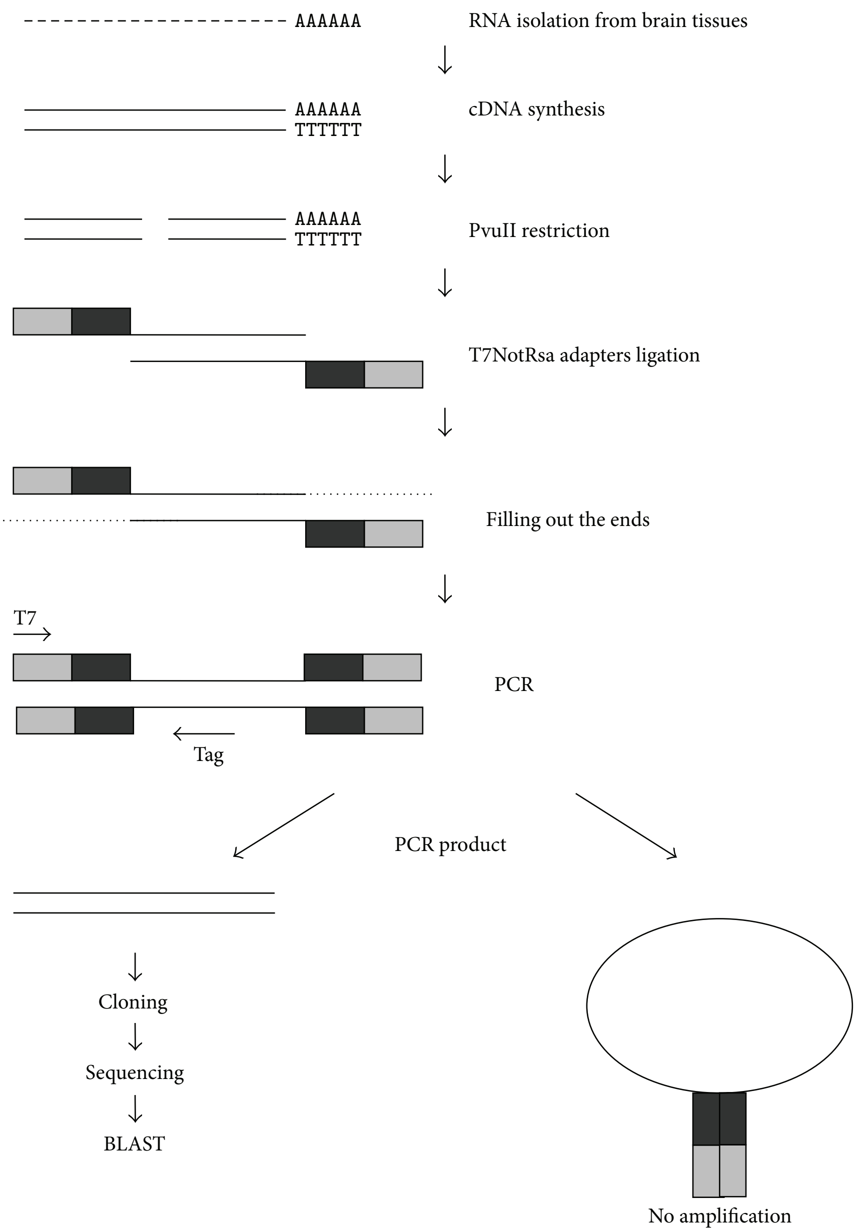
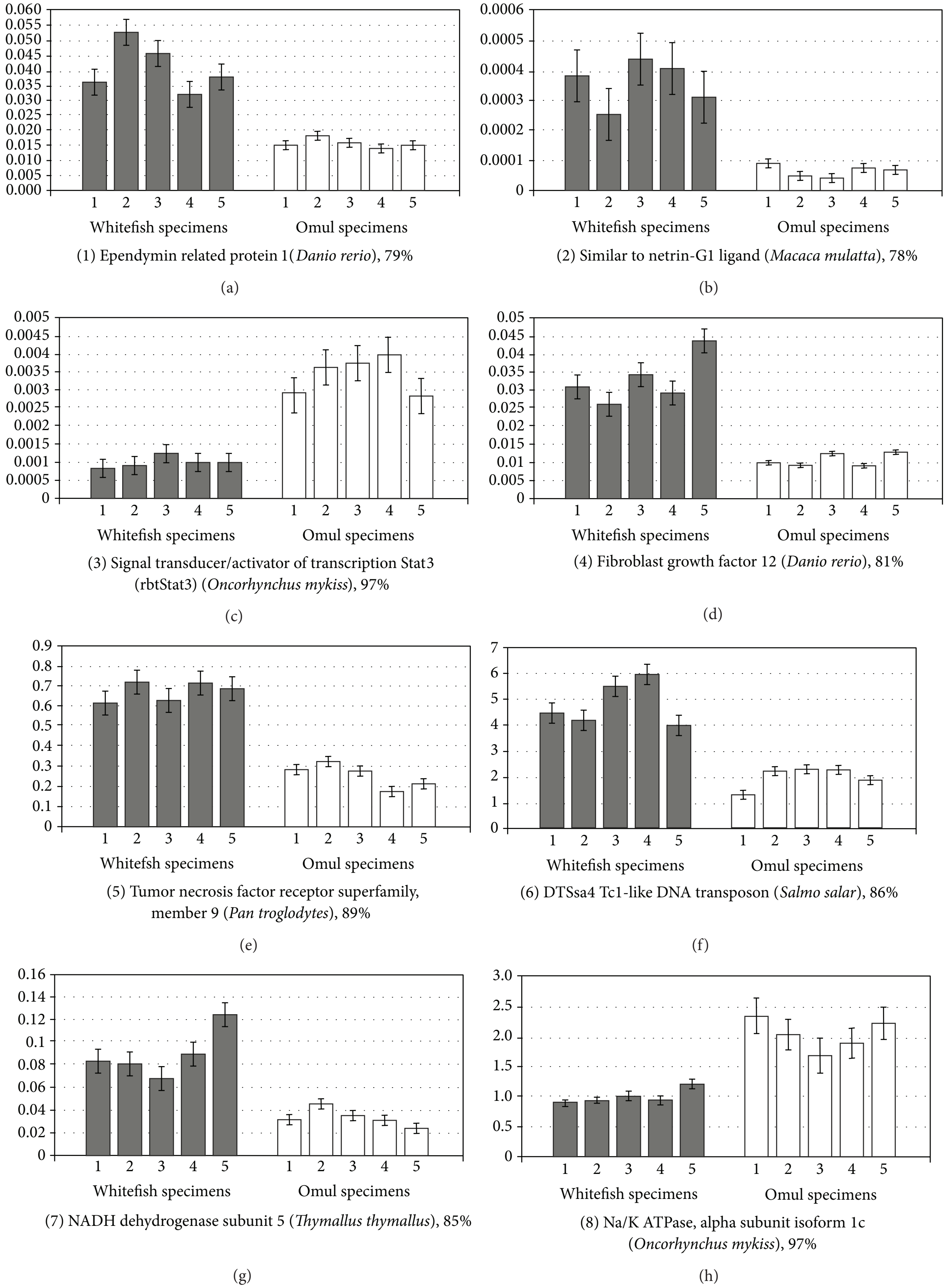


Figure 1: Identification of cDNAs corresponding to selected tags. cDNA synthesis was performed using the oligo(dT) primer. Double-stranded cDNA was hydrolyzed with PvuII followed by ligation to T7NotRsa suppression adapters (grey boxes). PCR primers are designated by arrows. "Tag" denotes an oligonucleotide primer designed according to the tag sequence found by SAGE.

use of the pelagic niche accompanied by any parallelism in the differential pattern of expression? We used the SAGE approach to reveal the difference in brain gene expression levels between Baikal whitefish and omul and identified some of candidate genes involved in the divergence of these species.

4.1. Differential Expression in Brain of Baikal Whitefish and Omul. At least ~4% of cDNA tags revealed quantitative differences between the fishes under study. Differential expression in the brains of closely related salmonids was also reported by other authors. For example, Aubin-Horth et al. found that 15% of transcripts were differentially expressed in the brain of salmonids (Atlantic salmon) with alternative developmental paths [32]. A differential expression was reported also for 10.5% of transcripts in the brain of salmonids (Atlantic salmon) living in natural conditions and those kept in captivity [33]. Finally, 11% of brain transcripts were differentially expressed in benthic and pelagic forms of whitefishes inhabiting lakes in the northeast of North America and diverged from a common ancestor. Candidate genes involved in the species divergence of these fishes belonged to functional categories of energy metabolism, protein synthesis, neural development, and, in some cases, regulatory genes, for example, a zinc-finger protein gene [1].

We show here that most tags that were two- to threefold more represented in whitefish than in omul were similar



(1) Ependymin related protein 1(*Danio rerio*), 79%

(a)

(2) Similar to netrin-G1 ligand (*Macaca mulatta*), 78%

(b)

(3) Signal transducer/activator of transcription Stat3 (rbtStat3) (*Oncorhynchus mykiss*), 97%

(c)

(4) Fibroblast growth factor 12 (*Danio rerio*), 81%

(d)

(5) Tumor necrosis factor receptor superfamily, member 9 (*Pan troglodytes*), 89%

(e)

(6) DTSsa4 Tc1-like DNA transposon (*Salmo salar*), 86%

(f)

(7) NADH dehydrogenase subunit 5 (*Thymallus thymallus*), 85%

(g)

(8) Na/K ATPase, alpha subunit isoform 1c (*Oncorhynchus mykiss*), 97%

(h)

Figure 2: Transcription levels of genes in the whitefish and omul brain estimated by real-time PCR. %: similarity of the PCR product to mRNA of known genes. The *y*-axis shows the number of transcripts normalized to the content of the *GAPDH* housekeeping gene. Data represent mean ± S.E. of six independent experiments.

Table 3: Transcription levels of genes in the whitefish and omul brain estimated by real-time PCR.

	Accession	Similarity to mRNA of known genes, %	Accession of mRNA of known genes	Qwf ± S.E.	Qomul ± S.E.	R ± S.E.
1	GenBank: GR918015	Ependymin related protein 1 (*Danio rerio*), 79	GenBank: NM_001002416.1	$3.7 \times 10^{-2} \pm 0.4 \times 10^{-2}$	$1.5 \times 10^{-2} \pm 0.12 \times 10^{-2}$	2.79 ± 0.6
2	GenBank: GR918016	Similar to netrin-G1 ligand (*Macaca mulatta*), 78	GenBank: XM_001090447.1	$3.9 \times 10^{-4} \pm 0.9 \times 10^{-4}$	$0.89 \times 10^{-4} \pm 0.2 \times 10^{-4}$	4.12 ± 0.8
3	GenBank: GR918017	Signal transducer/activator of transcription Stat3 (rbtStat3) mRNA (*Oncorhynchus mykiss*), 97	GenBank: OMU60333	$8.2 \times 10^{-4} \pm 3.1 \times 10^{-4}$	$29.09 \times 10^{-4} \pm 5 \times 10^{-4}$	0.33 ± 0.05
4	GenBank: GR918020	Fibroblast growth factor 12 (*Danio rerio*), 81	GenBank: BC124640.1	$3.1 \times 10^{-2} \pm 0.42 \times 10^{-2}$	$1.05 \times 10^{-2} \pm 0.09 \times 10^{-4}$	2.62 ± 0.6
5	GenBank: GR918021	Tumor necrosis factor receptor superfamily, member 9 (*Pan troglodytes*), 89	GenBank: XM_001157779.1	0.61 ± 0.04	0.27 ± 0.02	2.45 ± 0.6
6	GenBank: GR918018	DTSsa4 Tc1-like DNA transposon (*Salmo salar*), 86	GenBank: EF685957.1	4.43 ± 0.41	1.3 ± 0.18	3.48 ± 0.5
7	GenBank: GR918022	NADH dehydrogenase subunit 5 (*Thymallus thymallus*), 85	GenBank: AF270855	$8.2 \times 10^{-2} \pm 0.9 \times 10^{-2}$	$3.5 \times 10^{-2} \pm 0.42 \times 10^{-2}$	2.27 ± 0.24
8	GenBank: GR918019	Na/K ATPase alpha subunit isoform 1c mRNA (*Oncorhynchus mykiss*), 97	GenBank: AY319389.1	0.86 ± 0.05	2.38 ± 0.21	0.35 ± 0.03

Data represent mean ± S.E. (standard error) of six independent experiments.
Q: the number of transcripts normalized to the number of *GAPDH* gene transcripts.
R: relative expression ~ Qwhitefish/Qomul.

to protein synthesis and regulatory genes. In contrast, tags that were two- to threefold more represented in omul were mostly similar to segments of metabolism genes. Some tags differentially expressed in brain of whitefish and omul were similar to cDNA of genes of the nervous and immune systems and of DTSsa4 Tc1-like DNA transposons.

A considerable difference in expression was displayed by genes of the fish nervous system. An increased transcription level in whitefish was characteristic of two genes: ependimin related protein 1 (ERP) and a gene resembling the gene of netrin-G1 ligand (NGL-1). ERP is present at a high concentration in the cerebrospinal fluid of bony fishes, and it is associated with neuroplasticity, regeneration, and learning processes [34]. NGL-1 is mostly located in the cerebral cortex. Surface-bound NGL-1 stimulates the growth of embryonic thalamic axons, but free in solution the protein suppresses the growth [35]. In the brain of omul, an enhanced transcription level was observed for the *Stat3* gene of a signal transducer/activator of transcription. STAT3 is a multifunctional transcription factor of the central and peripheral nervous systems [36] and is necessary for differentiation of glial cells [37].

The identified genes of the nervous system might directly or indirectly affect the behavioural mechanisms of fishes, and the differential expression of these genes might facilitate the adaptation of the fishes to changeable environmental conditions. Two more genes of the nervous system known from the literature—the genes of troponin and SPARC—are thought to take part in the adaptive divergence of a lake whitefish species (*Coregonus* sp.) pair, dwarf (limnetic) and normal (benthic) whitefish [1]. These genes might be directly related to behaviour. SPARC is involved in neural development through signalling that allows neurons to end developmental migrations [38]. SPARC has also been proposed to modulate synaptic functions and to be involved in higher cortical functions in adult mammalian brains [39], making this a candidate for depth-preference behaviour [1]. Troponin is associated with actin and tropomyosin in the actin scaffold of muscle tissue [40]. In neurons, these molecules are collectively associated with neural development and growth [32, 41], thus potentially providing a link between troponin and behavioural differences between species pairs [1]. Thereby, in the nervous system, troponin and SPARC genes are needed for the development and growth of neurons. The differential expression of these genes in whitefishes is also suggested to affect the behaviour of fish, for example, the choice of depth of habitation. Thus, one of the potential targets of natural selection leading to behavioural differences might be the modulation of the expression of genes involved in the development of the nervous system.

Another way of adapting to changing environmental conditions can be accomplished by changing the expression levels of metabolism genes. One such gene is the gene of Na/K ATPase whose expression in omul is two- to threefold higher than that in whitefish. In nervous cells, Na/K ATPase is important for establishing the electrochemical gradient necessary for electroexcitability [42]. The enzyme is composed of a catalytic α-subunit and a glycoprotein β-subunit, which are suggested to be involved in the transport and stabilization of the enzyme complexes in membranes. The expression level of the catalytic α-subunit in Baikal omul was found to be ~twofold higher than that in whitefish. A possible explanation is that Baikal omul leads a more active life than whitefish, thus needing a faster neural conduction velocity.

Genes of the respiration chain, such as Cytochrome c oxidase, Cytochrome P450, and ATP synthase, were also differentially expressed in whitefish and omul. Moreover, the transcript ratios for different families or subunits of these proteins in whitefish and omul were not equal. For instance, the brain of whitefish contained a larger number of ATP synthase d subunit transcripts than that of omul, whereas the content of b and g subunit transcripts was higher in omul than in whitefish. This could also be due to metabolism and life features of benthic Baikal whitefish and pelagic omul.

A considerable portion of tags that were two- to threefold better represented in the whitefish than in the omul library had a resemblance to segments of regulatory genes, for example, the gene of fibroblast growth factor (FGF) that plays a key role in proliferation and differentiation of various cells and tissues. In this work, we analysed young, rapidly growing and immature fish individuals of approximately the same age (3-4- and 4-5-year-old whitefish and omul, resp.). Therefore, the observed differential expression of regulatory genes might be, first of all, due to different growth rates of whitefish and omul: whitefish weighs more because the absolute body mass increment of whitefish is higher than that of omul during practically all the life [31, 43]. The same explanation may also be valid for the tags similar to mRNA segments of the genes of 60S and 40S ribosomal proteins, such as S5, S11, L7, and L13. An exception is the tags similar to mRNA of 60S ribosomal protein L31.

The whitefish library of representative sequences also contains more tags similar to the gene of TNF receptor and to some genes of the immune system, for example, the gene of MHC H-2 class II histocompatibility antigen γ chain. TNF receptor plays an important role in regulation of a wide spectrum of physiological process including the immune response.

It should be noted that a comparison of the Baikal whitefish and omul genomes using subtractive hybridization revealed quite a few differences in noncoding regions of the immune system genes, such as *TCR, MHC,* and *IgA*. In addition, many of the differential fragments located close to coding regions of these genes were 65–85% similar to TC1-like transposons [16]. We show here that this family's transposons have differential expression in brain tissues of the fishes under study, with the corresponding tags being ~threefold better represented in the whitefish library. In the genomes of *Salmo salar* and *Danio rerio*, the Tc1-like transposon cDNA fragments identified in our work were in some cases mapped close to important regulatory genes, such as the genes of growth hormone, steroidogenic acute regulatory protein (StAR), coiled-coil transcriptional coactivator b2, homeobox protein HoxC13bb, guanine nucleotide binding protein G(o), and retinoic acid receptor γ b.

4.2. Parallelism in Differential Patterns of Expression between Pelagic and Benthic Ecotypes across Whitefish Sympatric Pairs. Over the last few years, gene expression studies have been intensively used to investigate the molecular basis of adaptive divergence between whitefish ecotypes in North American (taxon *Coregonus clupeaformis*) lakes [2, 8, 44]. These microarray studies were conducted both in controlled (laboratory) and natural (two lakes) environments, involved three life stages (embryos, juveniles, and adults) and three tissues (white muscle, liver, and brain). The hundreds of genes that showed differential patterns of transcription between pelagic and benthic whitefish across the three tissues were classified into at least 30 different functional groups. Of particular interest are those functional groups that were overrepresented in terms of number of parallel genes showing differences between pelagic and benthic whitefish relative to the total number of genes that were expressed in both ecotypes for each functional group. Pelagic whitefish consistently showed significant overexpression of genes potentially associated with survival through enhanced activity (energy metabolism, muscle contraction, homeostasis, lipid metabolism, and detoxification) whereas genes associated with growth (protein synthesis, cell cycle, and cell growth) were generally upregulated in benthic relative to pelagic whitefish [3]. In general, these transcriptomic studies combined with physiological data [45] show that energy metabolism is the main biological function involved in the divergence between pelagic and benthic whitefish. There is mounting evidence that selection has been acting more strongly on pelagic than benthic whitefish [8, 44].

In general, SAGE of Baikal whitefish and omul brain transcriptomes revealed a similar pattern of gene expression. Even with a low absolute amount of sequenced tags, it is evident that tags more represented in omul (pelagic ecotype) were mostly similar to segments of metabolism genes. In contrast, tags more represented in Baikal whitefish (benthic ecotype) had a resemblance to protein synthesis and regulatory genes.

Obviously, in Lake Baikal, in comparison with North American lakes, selection has been acting even more strongly on pelagic ecotype. The following peculiarities of Lake Baikal whitefish pair testify to it: (1) complete reproductive isolation of ecotypes/species by spawning times (autumn/winter) and places (rivers/lake shoals); (2) multilevel pattern of intraspecific phenotypic divergence, pronounced in the pelagic Baikal omul [13].

Transcriptome sequencing [46] of pelagic and benthic whitefish ecotypes in North American (taxon *Coregonus clupeaformis*) lakes displays even more clear parallelisms with the results of our SAGE on Lake Baikal whitefish and omul. The most salient finding of this work was that 14 genes involved in energy metabolism (both mitochondrial and nuclear) showed pronounced allele frequency differences and were also identified in several previous gene expression studies as differentially expressed in parallel between pelagic and benthic whitefish [46]. They are seven mitochondrial genes (cytochrome C subunits 1, 2, and 3; NADH-dehydrogenases 1, 4, and 5; and cytochrome b) and seven nuclear genes (cytochrome b-c1 complex subunit 6, ATP synthase subunit d, malate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, creatine kinase, succinyl-CoA ligase, and angiopoietin-related protein 3 precursor). Special attention is given to genes of metabolic genes associated with the mitochondrion machinery. In our study, genes of the respiration chain were also differentially expressed in the Baikalian sympatric pair. Namely, the transcript ratios for different families or subunits of such proteins as Cytochrome c oxidase, Cytochrome P450, and ATP synthase in whitefish and omul were not equal. In general, as mentioned above, SAGE tags more represented in the Baikal omul (pelagic ecotype) were mostly similar to segments of metabolism genes.

"Nonmodel" species studied in their ecological context such as whitefish play an increasingly important role in ecological genomics [3]. Our work has confirmed that Lake Baikal is one more unique place to study genetic and phenotypic divergence among sympatric whitefish ecotypes. The comparative study of Baikal whitefish and omul brain transcriptomes revealed quantitative differences between species. Several genes involved with species diversity were identified and RT-PCR testified that differences in gene expression were not simple polymorphisms among fish within a species. Since the genomic sequence of both organisms studied remains unavailable, the generated data, albeit informative, remains speculative. Nevertheless, the presence of parallelism in differences of gene expression with similar sympatric whitefish pairs is evident. Thus, we hope this study will aid in future studies aimed at identifying the full genetic sequence of a number of expressed transcripts and will be a prelude to a more detailed analysis of adaptive variation and evolution of gene expression of Baikal whitefish and omul. The next step should be a comparative sequencing of transcriptomes. Undoubtedly, the questions that need to be further examined are the following. (1) Is parallelism in phenotypic adaptation of Lake Baikal and other whitefishes toward the use of the pelagic niche accompanied by parallelism in candidate gene transcription? (2) Is complete isolation between Lake Baikal whitefish ecotypes and multilevel pattern of intraspecific phenotypic divergence in pelagic ecotype accompanied by the extent of candidate gene transcription?

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank Dr. B.O. Glotov for critical reading of the paper. The work was financially supported by RFBR Grant no. 05-04-486880, no. 10-04-01583, no. 14-04-01242, Grants of the Program of Principal Scientific Schools of Russia (Project 2395.2008.4), the Program "Molecular and Cellular Biology" of the Presidium of the Russian Academy of Sciences, and the Program of the Siberian Brunch of the Russian Academy of Sciences "Vivaria, collections of cell cultures, unique bacterial strains, microorganisms and plants." DNA was sequenced in the Interinstitute "Genome" Center (http://www.genome-centre.narod.ru/) organized under the support of RFBR Grant no. 00-04-55000.

References

[1] A. R. Whiteley, N. Derome, S. M. Rogers et al., "The phenomics and expression quantitative trait locus mapping of brain transcriptomes regulating adaptive divergence in lake whitefish species pairs (Coregonus sp.)," *Genetics*, vol. 180, no. 1, pp. 147–164, 2008.

[2] J. Jeukens, D. Bittner, R. Knudsen, and L. Bernatchez, "Candidate genes and adaptive radiation: Insights from transcriptional adaptation to the limnetic niche among coregonine fishes (Coregonus spp., Salmonidae)," *Molecular Biology and Evolution*, vol. 26, no. 1, pp. 155–166, 2009.

[3] L. Bernatchez, S. Renaut, A. R. Whiteley et al., "On the origin of species: insights from the ecological genomics of lake whitefish," *Philosophical Transactions of the Royal Society B*, vol. 365, no. 1547, pp. 1783–1800, 2010.

[4] D. Pigeon, A. Chouinard, and L. Bernatchez, "Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of lake whitefish (Coregonus clupeaformis salmonidae)," *Evolution*, vol. 51, no. 1, pp. 196–205, 1997.

[5] K. Østbye, P.-A. Amundsen, L. Bernatchez et al., "Parallel evolution of ecomorphological traits in the European whitefish Coregonus lavaretus (L.) species complex during postglacial times," *Molecular Ecology*, vol. 15, no. 13, pp. 3983–4001, 2006.

[6] L. Bernatchez and J. J. Dodson, "Phylogeographic structure in mitochondrial DNA of the lake whitefish (Coregonus clupeaformis) and its relation to Pleistocene glaciations," *Evolution*, vol. 45, no. 4, pp. 1016–1035, 1991.

[7] D. Campbell and L. Bernatchez, "Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes," *Molecular Biology and Evolution*, vol. 21, no. 5, pp. 945–956, 2004.

[8] N. Derome, P. Duchesne, and L. Bernatchez, "Parallelism in gene transcription among sympatric lake whitefish (Coregonus clupeaformis Mitchill) ecotypes," *Molecular Ecology*, vol. 15, no. 5, pp. 1239–1249, 2006.

[9] D. V. Politov, N. Y. Gordon, K. I. Afanasiev, Y. P. Altukhov, and J. W. Bickham, "Identification of palearctic coregonid fish species using mtDNA and allozyme genetic markers," *Journal of Fish Biology A*, vol. 57, pp. 51–71, 2000.

[10] D. V. Politov, J. W. Bickham, and J. C. Patton, "Molecular phylogeography of Palearctic and Nearctic ciscoes," *Annales Zoologici Fennici*, vol. 41, no. 1, pp. 13–23, 2004.

[11] L. V. Sukhanova, V. V. Smirnov, N. S. Smirnova-Zalumi, S. V. Kiril'chik, D. Griffiths, and S. I. Belikov, "The taxonomic position of the Lake Baikal omul, Coregonus autumnalis migratorius (Georgi), as revealed by sequence analysis of the mtDNA cytochrome b gene and control region," *Advances in Limnology*, vol. 57, pp. 97–106, 2002.

[12] L. V. Sukhanova, V. V. Smirnov, N. S. Smirnova-Zalumi, S. V. Kirilchik, and I. Shimizu, "Grouping of Baikal omul Coregonus autumnalis migratorius Georgi within the C. lavaretus complex confirmed by using a nuclear DNA marker," *Annales Zoologici Fennici*, vol. 41, no. 1, pp. 41–49, 2004.

[13] V. V. Smirnov, N. S. Smirnova-Zalumi, and L. V. Sukhanova, *Microevolution of Baikal Omul*, Novosibirsk, Publishing House of Siberian Branch of Russian Academy of Sciences, 2009, (Russian).

[14] L. V. Sukhanova, V. V. Smirnov, N. S. Smirnova-Zalumi, T. V. Belomestnykh, and S. V. Kirilchik, *Molecular Phylogeography of Lake Baikal Coregonid Fishes*, Schweizerbart, Stuttgart, Germany, 2012.

[15] V. D. Mats, "The structure and development of the Baikal rift depression," *Earth Science Reviews*, vol. 34, no. 2, pp. 81–118, 1993.

[16] O. S. Bychenko, L. V. Sukhanova, S. S. Ukolova et al., "Genome similarity of Baikal omul and sig," *Russian Journal of Bioorganic Chemistry*, vol. 35, no. 1, pp. 86–93, 2009.

[17] M. D. Purugganan, "The molecular evolution of development," *Bioessays*, vol. 20, pp. 700–711, 1998.

[18] S. M. Rogers and L. Bernatchez, "Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (Coregonus clupeaformis)," *Molecular Ecology*, vol. 14, no. 2, pp. 351–361, 2005.

[19] G. Gibson and E. Honeycutt, "The evolution of developmental regulatory pathways," *Current Opinion in Genetics and Development*, vol. 12, no. 6, pp. 695–700, 2002.

[20] G. E. Robinson, C. M. Grozinger, and C. W. Whitfield, "Sociogenomics: social life in molecular terms," *Nature Reviews Genetics*, vol. 6, no. 4, pp. 257–270, 2005.

[21] N. Aubin-Horth, J. K. Desjardins, Y. M. Martei, S. Balshine, and H. A. Hofmann, "Masculinized dominant females in a cooperatively breeding species," *Molecular Ecology*, vol. 16, no. 7, pp. 1349–1358, 2007.

[22] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression," *Science*, vol. 270, no. 5235, pp. 484–487, 1995.

[23] J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2nd edition, 1989.

[24] J. M. Ruijter, C. Ramakers, W. M. H. Hoogaars et al., "Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data," *Nucleic Acids Research*, vol. 37, no. 6, article e45, 2009.

[25] M. W. Pfaffl, G. W. Horgan, and L. Dempfle, "Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR," *Nucleic Acids Research*, vol. 30, no. 9, p. e36, 2002.

[26] T. Azhikina, I. Gainetdinov, Y. Skvortsova, and E. Sverdlov, "Methylation-free site patterns along a 1-Mb locus on Chr19 in cancerous and normal cells are similar. A new fast approach for analyzing unmethylated CCGG sites distribution," *Molecular Genetics and Genomics*, vol. 275, no. 6, pp. 615–622, 2006.

[27] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[28] S. Audic and J. Claverie, "The significance of digital gene expression profiles," *Genome Research*, vol. 7, no. 10, pp. 986–995, 1997.

[29] M. L. Evans and L. Bernatchez, "Oxidative phosphorylation gene transcription in whitefish species pairs reveals patterns of parallel and nonparallel physiological divergence," *Journal of Evolutionary Biology*, vol. 25, pp. 1823–1834, 2012.

[30] M. L. Evans, L. J. Chapman, I. Mitrofanov, and L. Bernatchez, "Variable extent of parallelism in respiratory, circulatory, and neurological traits across lake whitefish species pairs," *Ecology and Evolution*, vol. 3, pp. 546–557, 2013.

[31] A. G. Skryabin, *Biology of Baikal Whitefish*, Nauka, Moskow, Russia, 1969.

[32] N. Aubin-Horth, C. R. Landry, B. H. Letcher, and H. A. Hofmann, "Alternative life histories shape brain gene expression profiles in males of the same population," *Proceedings of the Royal Society B*, vol. 272, no. 1573, pp. 1655–1662, 2005.

[33] N. Aubin-Horth, B. H. Letcher, and H. A. Hofmann, "Interaction of rearing environment and reproductive tactic on gene expression profiles in Atlantic salmon," *Journal of Heredity*, vol. 96, no. 3, pp. 261–278, 2005.

[34] E. C. Suarez-Castillo and J. E. Garcia-Arraras, "Molecular evolution of the ependymin protein family: a necessary update," *BMC Evolutionary Biology*, vol. 7, article 23, 2007.

[35] J. C. Lin, W. Ho, A. Gurney, and A. Rosenthal, "The netrin-G1 ligand NGL-1 promotes the outgrowth of thalamocortical axons," *Nature Neuroscience*, vol. 6, no. 12, pp. 1270–1276, 2003.

[36] S. Dziennis and N. J. Alkayed, "Role of Signal Transducer and activator of transcription 3 in neuronal survival and regeneration," *Reviews in the Neurosciences*, vol. 19, no. 4-5, pp. 341–361, 2008.

[37] A. Bonni, Y. Sun, M. Nadal-Vicens et al., "Regulation of gliogenesis in the central nervous system by the JAK-STAT signaling pathway," *Science*, vol. 278, no. 5337, pp. 477–483, 1997.

[38] V. Gongidi, C. Ring, M. Moody et al., "SPARC-like 1 regulates the terminal phase of radial gliaguided migration in the cerebral cortex," *Neuron*, vol. 41, no. 1, pp. 57–69, 2004.

[39] S. Lively, M. J. Ringuette, and I. R. Brown, "Localization of the extracellular matrix protein SC1 to synapses in the adult rat brain," *Neurochemical Research*, vol. 32, no. 1, pp. 65–71, 2007.

[40] F. J. Roisen, F. J. Wilson, and G. Yorke, "Immunohistochemical localization of troponin-C in cultured neurons," *Journal of Muscle Research and Cell Motility*, vol. 4, no. 2, pp. 163–175, 1983.

[41] G. Schevzov, P. Gunning, P. L. Jeffrey et al., "Tropomyosin localization reveals distinct populations of microfilaments in neurites and growth cones," *Molecular and Cellular Neurosciences*, vol. 8, no. 6, pp. 439–454, 1996.

[42] S. T. Sayers, T. Khan, R. Shahid, M. F. Dauzvardis, and G. J. Siegel, "Distribution of alpha 1 subunit isoform of (Na,K)-ATPase in the rat spinal cord," *Neurochemical Research*, vol. 19, no. 5, pp. 597–602, 1994.

[43] V. V. Smirnov and I. P. Shumilov, *Baikal Omul*, Nauka, Novosibirsk, Russia, 1974.

[44] J. St-Cyr, N. Derome, and L. Bernatchez, "The transcriptomics of life-history trade-offs in whitefish species pairs (Coregonus sp.)," *Molecular Ecology*, vol. 17, no. 7, pp. 1850–1870, 2008.

[45] M. Trudel, A. Tremblay, R. Schetagne, and J. B. Rasmussen, "Why are dwarf fish so small? An energetic analysis of polymorphism in lake whitefish (Coregonus clupeaformis)," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 58, no. 2, pp. 394–405, 2001.

[46] S. Renaut, A. W. Nolte, and L. Bernatchez, "Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae)," *Molecular Ecology*, vol. 19, supplement 1, pp. 115–131, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 269402, 13 pages
http://dx.doi.org/10.1155/2014/269402

Research Article

Changes in Bacterial Population of Gastrointestinal Tract of Weaned Pigs Fed with Different Additives

Mercè Roca,[1] Miquel Nofrarías,[1] Natàlia Majó,[1,2] Ana María Pérez de Rozas,[1,3] Joaquim Segalés,[1,2] Marisol Castillo,[4] Susana María Martín-Orúe,[4] Anna Espinal,[5] Joan Pujols,[1,3] and Ignacio Badiola[1,3]

[1] *Centre de Recerca en Sanitat Animal (CReSA), UAB-IRTA, Campus de la Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*
[2] *Departament de Sanitat i Anatomia Animals, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*
[3] *Institut de Recerca i Tecnologia Agroalimentaria (IRTA), Caldes de Montbui, 08140 Barcelona, Spain*
[4] *Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*
[5] *Servei d'Estadística (SEA), Edifici D, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*

Correspondence should be addressed to Miquel Nofrarías; miquel.nofrarias@cresa.uab.cat

Received 28 February 2013; Revised 15 November 2013; Accepted 27 November 2013; Published 19 January 2014

Academic Editor: William H. Piel

This study aimed to provide novel insights into the gastrointestinal microbial diversity from different gastrointestinal locations in weaning piglets using PCR-restriction fragment length polymorphism (PCR-RFLP). Additionally, the effect of different feed additives was analyzed. Thirty-two piglets were fed with four different diets: a control group and three enriched diets, with avilamycin, sodium butyrate, and a plant extract mixture. Digesta samples were collected from eight different gastrointestinal segments of each animal and the bacterial population was analysed by a PCR-RFLP technique that uses 16S rDNA gene sequences. Bacterial diversity was assessed by calculating the number of bands and the Shannon-Weaver index. Dendrograms were constructed to estimate the similarity of bacterial populations. A higher bacterial diversity was detected in large intestine compared to small intestine. Among diets, the most relevant microbial diversity differences were found between sodium butyrate and plant extract mixture. Proximal jejunum, ileum, and proximal colon were identified as those segments that could be representative of microbial diversity in pig gut. Results indicate that PCR-RFLP technique allowed detecting modifications on the gastrointestinal microbial ecology in pigs fed with different additives, such as increased biodiversity by sodium butyrate in feed.

1. Introduction

The weaning period is one of the most critical stages in the life cycle of pigs. This is due to the fact that, in this short period of time, the animal is yielding to many changes, such as social, environmental, and dietary changes. The combination of these news circumstances generally results in a distressing stage for the animals, favoring the occurrence of different bacterial and viral illnesses [1]. For many years, antimicrobial growth promoters (AGPs) have been used to improve the body weight gain and to control the overgrowth of opportunistic pathogens, especially during the weaning period [2, 3]. Nowadays, the increasing risk of antimicrobial resistance has led the European Union to ban all the AGPs used in animal nutrition. As a consequence, the interest in pig gut microbiota has increased considerably during the last few years. A better knowledge of the gastrointestinal bacterial composition would help to find new feed strategies to avoid pathologic infections and thus keep pigs healthier. This necessity to deepen in the knowledge of the ecosystem of the digestive tract has coincided with the appearance and implementation of molecular techniques in different fields of science and also in the microbiology that allows to have a much wider vision of these complex systems.

Currently, different additives such as plant extracts or acidifiers have been used as alternative to AGPs. The wide antimicrobial spectrum of some plant extracts such as thyme, oregano, and cinnamon has been clearly demonstrated in different *in vitro* studies [4, 5]. However, less consistent antimicrobial effects have been observed when these compounds are used *in vivo*. In chickens, a reduction of *Escherichia coli* and *Clostridium perfringens* in cecum was detected using a plant extract blend from capsicum, cinnamaldehyde, and carvacrol [6]. In weaned piglets, Manzanilla et al. and Castillo et al. [7, 8] using a similar composition mixture detected an increase of *Lactobacillus* spp. by culture and quantitative PCR, respectively. Even so, more information on the precise *in vivo* action of these compounds and their effect on microbial communities are needed to consider plant extracts as a real alternative to AGPs.

On the other hand, dietary acidification with organic acids or their salts constitutes another alternative to AGPs due to their beneficial effects on protein digestion and performance [9, 10]. It is generally accepted that the antimicrobial action of the organic acids is mainly due to the acidification of the gastric medium and to the ability of these acids to dissociate into the microbial cells, causing a shift to particular bacterial groups [11, 12]. It has been hypothesized that sodium butyrate could help to maintain the epithelium integrity, protecting the animals against pathogenic agents [13]. Moreover, Gálfi and Bokori [14] reported that sodium-butyrate supplementation in piglets reduced coliform bacteria and increased the number of *Lactobacillus* spp. in the ileum. In poultry, this supplementation also reduced coliform bacteria such as *E. coli* and *Salmonella* spp. [15]. However, few studies are available to elucidate the effect of acids on the bacterial ecosystem in the gastrointestinal tract.

Despite the general use of AGPs, their exact mode of action is not clear, and different mechanisms have been proposed. Most of them are based on the reduction of intestinal microbial mass, resulting in a decreased production of growth of depressing microbial metabolites and in the competition for nutrients with the host [16, 17]. However, other mechanisms related to the selection of a healthier microbial community could also be implicated.

The present work was designed to evaluate the existence of differences in the microbial diversity of different gastrointestinal locations by PCR-restriction fragment length polymorphism (PCR-RFLP) in weaning piglets. Moreover, a second objective was to elucidate the effect of different additives (acidifier, plant extract mixture, and antibiotic) on the gastrointestinal microbial populations.

Table 1: Control diet composition, as fed basis.

Ingredient	g/kg
Corn	276
Barley	300
Soybean meal, 44% CP	40
Full fat extruded soybeans	40
Soya protein concentrate	60
Fish meal LT[a]	50
Dried whey	40
Acid whey[b]	150
Wheat gluten	6.8
Sepiolite	10
Dicalcium phosphate	11
L-Lysine·HCl	4.4
DL-Methionine	2.7
L-threonine	1.9
L-tryptophan	0.4
Choline chloride, 50% choline	2.0
Chromic oxide	1.5
Vitamin and mineral premix[c]	3.0
Calculated nutrient composition[d]	
GE, Mcal/kg	4.75
Crude protein, g/kg	183.9
Ether extract, g/kg	51.1
Crude fiber, g/kg	27.8
Ca, g/kg	6.44
P total, g/kg	6.95
P available, g/kg	4.01
Lysine, g/kg	13.87

[a] Fish meal low temperature: product obtained by removing most of the water and some or all of the oil from fish by heating at low temperature (<70°C) and pressing.

[b] Acid whey: product obtained by drying fresh whey (derived during the manufacture of cheeses) that has been pasteurized.

[c] Provided the following per kilogram of diet: vitamin A, 13,500 IU; vitamin D3, 2000 IU; vitamin E, 80 mg; vitamin K3, 4 mg; thiamine, 3 mg; riboflavin, 8 mg; vitamin B6, 5 mg; vitamin B12, 40 μg; nicotinic acid, 40 mg; calcium pantothenate, 15 mg; folic acid, 1.3 mg; biotin, 150 μg; Fe, 120 mg as iron carbonate; Cu, 175 mg as copper sulfate $5H_2O$; Zn, 110 mg as zinc oxide; Mn, 65 mg as manganese sulfate; I, 1 mg as potassium iodate; selenium, 0.10 mg as sodium selenite.

[d] Based on composition values from NRC (1998).

2. Materials and Methods

2.1. Animals, Housing, and Management. The trial was performed at the Experimental Unit of the Universitat Autònoma de Barcelona and received prior approval from the University Ethical Committee for Animal Experimentation. All the procedures involving animals were conducted in accordance with the European Union Guidelines for Animal Welfare (Directiva 86/609/CEE).

Thirty-two crossbred (Pietrain × [Landrace × Large White]) mixed male and female weaned pigs were selected from ten different litters of the same farm. The pigs were weaned at 20 ± 2 days of age with an average initial body weight of 5.9 ± 0.7 kg. The piglets were distributed in 4 groups (8 animals per diet; 2 pens per diet) according to their initial weight and litter. All pigs were allocated in an environmentally controlled room and the temperature was gradually reduced from 29 to 25°C over a period of 3 weeks.

2.2. Diets and Experimental Design. The pigs were fed with four different experimental diets. One group (CT group) received a control diet, which was formulated, with 60% cereals, 20% milk by-products, 6% soy protein concentrate, 5% low temperature fish meal, 4% soy bean meal 44, and 4% full fat soy as the main ingredients (Table 1). The remaining three groups received the same diet where three different additives were added as follows: 0.04% of avilamycin (AB group), 0.3% of sodium butyrate (AC group) or 0.03% of a plant extract mixture standardized in 5% (wt/wt) carvacrol (*Origanum* spp.), 3% cinnamaldehyde (*Cinnamomum* spp.), and 2% capsicum oleoresin (*Capsicum annum*) (XT group). The animals were fed *ad libitum* during three weeks and had free access to water.

2.3. Controls and Sampling. Individual body weights (BW) were registered weekly and the average daily gain was calculated. After three weeks with the experimental diets, pigs were euthanized with an intravenous injection of sodium pentobarbital (Dolethal, Vetoquinol, S.A., Madrid, Spain) (200 mg/kg BW). The euthanasia was carried out in 2 days (day 19 and day 21 of the study) with 4 animals of each group per day.

The abdomen was immediately opened and the whole gastrointestinal tract was removed. Luminal content of all the animals was collected from stomach (ST), proximal jejunum (PJ), distal jejunum (DJ), ileum (I), cecum (C), proximal colon (PC), distal colon (DC), and rectum (R). One g of digest sample was placed in 3 mL of 98% grade ethanol and kept at 4°C until DNA extraction as described in Castillo et al. [18] and Murphy et al. [19].

2.4. Sample Processing

2.4.1. DNA Extraction and PCR. A sample of 400 mg from the ethanol-treated digest content was washed twice with sterile buffered peptone water. DNA was extracted using the QIAamp DNA Stool Minikit (Qiagen Inc., Chatsworth, California) following the manufacturer's instructions with a minor modification, which consisted in adding 140 μL of 10 mg/mL of lysozyme in Tris-EDTA buffer (pH 8.0) (Sigma Chemical Co., St. Louis, MO, USA) to the lysis buffer and incubating it at 37°C during 30 minutes to improve the DNA extraction of Gram-positive bacteria. After elution from the column, 2 μL of Ribonuclease-A (Sigma Chemical Co., St. Louis, MO, USA) was added to eliminate residual RNA. Also, 4 μL of 0.8 μg/mL BSA (Sigma Chemical Co., St. Louis, MO, USA) was added to each sample. The DNA was stored at −20°C until PCR amplification.

PCR was performed with AmpliTaq Gold PCR-Master mix (Applied Biosystems, CA, USA) in a total reaction volume of 50 μL. The PCR mix consisted of 0.05 U/μL of AmpliTaq Gold polymerase, 0.8 μM per each primer, 0.1% of tween 20, 5 μL of template DNA (~100 ng), and autoclaved nanopure water. The 16S rDNA was amplified using the eubacterial primers: 357fm (5′-CTACGGGAG-GCAGCAGT-3′) designed by Muyzer et al. [20] and 907 rm (5′-CCGTCWATTCMTTTGAGTTT-3′) by Muyzer et al. [21]. The optimized conditions for amplification were as follows: activation of TaqGold at 94°C (4 min), 35 cycles of denaturation at 94°C (1 min), annealing at 45°C (1 min) with an increment of 0.1°C per cycle, extension at 72°C (1 min 15 s), and a final extension at 72°C (15 min).

2.4.2. Restriction Fragment Length Polymorphism (RFLP). The PCR products were digested with four restriction enzymes (*AluI, RsaI, HpaII*, and *CfoI*) in four independent reactions (F. Hoffmann-LaRoche Ltd Group, Basel, Switzerland). A reaction mixture was made containing 8 μL of the PCR product, 1 μL SA buffer (F. Hoffmann-La Roche Ltd Group, Basel, Switzerland), and 1 μL (10 U) of each restriction enzyme. Samples were incubated for 3 hours at 37°C. The different fragments were separated using a 2% high resolution agarose gel (Sigma Chemical Co., St. Louis, MO, USA) and visualized by staining with 0.5 μg/mL ethidium bromide. Step ladder 50 bp (Sigma Chemical Co., St. Louis, MO, USA) was used as a DNA molecular weight marker and a mixture of PCR-RFLP from *Pasteurella multocida*, *Enterococcus faecalis,* and *Clostridium perfringens* digested with *RsaI* was processed and used as internal control. DNA bands were visualized in a UV Chemigenius Image System (SynGene, Cambrige, UK) using the GeneSnap software (SynGene Analysis Cambridge, UK, version 3.02.00), and the size of the restriction fragments was calculated with the Gene Tools software version 3.02.00 (SynGene Analysis Cambridge, UK). To reduce subjective variation during gel observation, peaks with an intensity under 60 units were discarded. Background noise was eliminated using the 30-radium rolling disk method. Finally, four band profiles for each sample were obtained which corresponded to the digesting with the four aforementioned restriction enzymes.

2.5. Analysis of Band Patterns. With the analysis of the restriction fragments, microbial diversity and the similarity degree were calculated. To calculate the microbial diversity two parameters were used: the total number of bands and the Shannon-Weaver H' index [22]. The total number of bands was calculated for each of the samples collected in the study as a total sum of the bands obtained in the four enzymatic reactions. The Shannon-Weaver index (H') was calculated using digesta samples from the distal jejunum and proximal colon by means of the following function: $H' = -\sum Pi \log Pi$, where Pi was the importance probability of finding a given band in a tract. Pi was itself calculated with the function $Pi = ni/nt$, where ni is the height in the densitometric curves (intensities of the bands) of a given peak and nt is the total sum of all the peaks of a densitometric curve. The final value of the Shannon-Weaver index was obtained as the average of the Shannon-Weaver calculated for each animal.

To compare the similarity in the bacterial composition of the different experimental diets and intestinal tracts, several dendrograms were built by calculating the Manhattan distance (MD) [23]:

$$\mathrm{MD}_{(\mathrm{SA,\,SB})} = 100 - \frac{\sum_{\mathrm{IP}}^{\mathrm{FP}} \sum_{\mathrm{IE}}^{\mathrm{FE}} \left|\mathrm{ISA}\,(\mathrm{nP},\mathrm{mE}) - \mathrm{ISB}\,(\mathrm{nP},\mathrm{mE})\right|}{(\mathrm{FP}-\mathrm{IP}) * (\mathrm{FE}-\mathrm{IE})}, \tag{1}$$

where MD (SA, SB) is the Manhattan distance between the electrophoretic profiles of sample A and sample B, IP is initial electrophoretic position, FP is final electrophoretic position, IE is initial restriction enzyme, FE is final restriction enzyme, ISA (nP, mE) is intensity of electrophoretic profile of

Table 2: Effect of the diet on the microbial diversity (Shannon-Weaver index) in the distal jejunum and proximal colon digesta. The pigs were fed with a control diet (CT) or with the same diet with avilamycin (AB), sodium butyrate (AC), or with a mixture of plant extracts (XT). Different letters (a, b, c) show significant differences among treatments for the same tract ($P < 0.05$).

Shannon-Weaver index	AC	AB	CT	XT	SEM*	P value diet	P value day	P value diet $*$ day
Distal jejunum	1.32	1.26	1.23	1.22	0.028	0.07	0.0004	0.77
Proximal colon	1.48^{a}	1.39^{bc}	1.42^{b}	1.37^{c}	0.002	0.002	0.098	0.34

*SEM: standard error of the mean.

sample A at the n position of m restriction enzyme, and ISB (nP, mE) is intensity of electrophoretic profile of sample B at the n position of m restriction enzyme.

The values of this coefficient range from 0 to 100. Both the presence and the absence of a band and its intensity (peak height of densitometric curves) were taken into account to calculate them. Using the distance matrix, different dendrograms were created using the neighbour-joining algorithm.

2.6. Statistical Analysis. The effect of the diet and the gastrointestinal tract on the number of bands was examined through nonparametric tests (Kruskal-Wallis) using the SAS statistical package (SAS Institute, INC. 8.2, Cary, NC). The statistically significant results ($P < 0.05$) were later administered by the Wilcoxon test (2 × 2 comparisons) applying the Bonferroni correction for multiple comparisons. The effects of the diet on the Shannon-Weaver index and on the productive parameters were analyzed by means of the SAS GLM procedure (general linear model). For all the analyses, statistical significance was determined for values of $P < 0.05$.

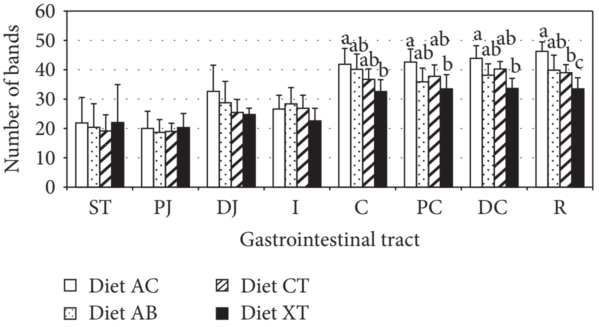


Figure 1: Microbial diversity in the different gastrointestinal segments. The pigs were fed with a control diet (CT) or with the same diet with avilamycin (AB), sodium butyrate (AC), or with a mixture of plant extracts (XT). The samples were collected from digestive contents in the stomach (ST), the proximal jejunum (PJ), the distal jejunum (DJ), the ileum (I), the proximal colon (PC), in the distal colon (DC), and the rectum (R). Error bars stand for standard deviation. Different letters (a, b, c) show significant differences among treatments for the same tract ($P < 0.05$).

3. Results

3.1. Clinical and Production Parameters. No clinical signs and no diarrhoea episodes were observed in any animal during the whole experimental period. There were no significant differences in growth performance ($P > 0.05$). However, the animals that received diets with some additive had a tendency to have a higher average daily weight gain than the CT group (124.7 ± 19.1, 177.4 ± 57.5, 177.6 ± 33.1, and 165.9 ± 37.4 g for CT, AB, AC, and XT, resp., $P = 0.069$).

3.2. Microbial Diversity

3.2.1. Number of Bands. The number of the bands from each gastrointestinal section and for each animal resulted from the total sum of the four enzymatic restrictions. The average value of the number of bands ranged from 18 to 46 per group (Figure 1). For all the diets, the number of bands was higher in the distal intestinal tracts (with values between 32.5 and 46.9) than in the proximal intestinal tracts (from 18.6 to 32.62) ($P < 0.05$). This effect was more remarkable in the AC and AB diets.

In some sections of the gastrointestinal tract (specifically in the distal jejunum and in the cecum) of the animals fed with the same diet, significant differences were observed among the four animals sacrificed on day 19 and the four sacrificed on day 21. This "sacrifice effect" was observed in the two aforementioned specific tracts in all the experimental diets. In the distal jejunum, animals from group AC euthanized on day 19 showed an average of 38 bands, whereas the animals euthanized on day 21 had an average of 27 bands ($P = 0.0029$). In the cecum, the four animals of group AC sacrificed on day 19 showed an average of 45 bands while the ones that were sacrificed on day 21 had an average of 38 bands ($P = 0.02$).

Statistically significant differences among the diets were basically observed in the large intestine (C, PC, DC, and R) (Figure 1). The most significant differences were observed between the AC and XT diets. A higher number of bands was observed in the animals fed with the AC diet and lower numbers in those fed with the XT diet (Figure 1).

3.2.2. Shannon-Weaver Index. The Shannon-Weaver index (H') was calculated exclusively using digesta samples from the distal jejunum and proximal colon. In all experimental diets, the degree of diversity measured by Shannon-Weaver index was lower in distal jejunum than in proximal colon (Table 2) ($P = 0.0001$). In the jejunum, numerical but not statistically significant differences were detected among experimental diets (Table 2). However, a decrease of the microbial diversity was observed from distal jejunum samples from the first day of sacrifice ($H' = 1.32$) to the second day of sacrifice ($H' = 1.20$), regardless of the experimental diet.

In proximal colon, the microbial diversity was affected by the composition of the diet (Table 2). Animals fed with the AC diet had a higher microbial diversity than others fed with other diets. In addition, animals fed with XT diets had a lower microbial diversity than those fed with CT diets too. No statistically significant differences were observed between the microbial diversity of the animals fed with AB diet and the animals fed with the CT diet.

3.3. Similarity in the Bacterial Composition. The similarity of bacterial populations between different intestinal tracts and different diets was evaluated with dendrograms which were built by calculating the Manhattan distance (Figures 2 to 7). The dendrogram formed with the band pattern of the different gastrointestinal tracts of pigs fed with a CT diet was shown in Figure 2. In this dendrogram, 2 clusters were observed depending on the day of sacrifice. The animals euthanized on day 19 (numeration from 1 to 4) are clustered in one branch, whereas the animals euthanized on day 21 (numeration from 5 to 8) are clustered in another branch. This sacrifice effect was not observed for the stomach samples. However, this effect was less clear in groups of animals fed with AC, AB, and XT diets (Figures 3, 4, and 5, resp.).

Analysing with more detail the different branches of the dendrogram of the animals fed with CT diets, it was possible to observe different clusters depending on the gastrointestinal section. Inside the group of animals sacrificed on day 19, it was possible to observe 2 subgroups. One subgroup was formed by distal intestinal segments (C, PC, DC, and R) and the other by the proximal tracts (PJ, DJ, and I). In the case of the animals euthanized on day 21, it was possible to observe 3 different clusters: one subgroup created by ST, PJ, and DJ samples, another subgroup constructed with different segments of the distal intestinal tract (C, PC, DC, and R), and a well-defined third subgroup created with the I samples. This separation between the proximal tracts and the posterior tracts was also observed in the other experimental diets (Figures 2 to 5).

To analyze the effect of the diets on a specific intestinal tract, different dendrograms were generated. We did not observe a clear cluster produced by diets on the microbial population from ST, PJ, DJ, I, C, DC, and R. Conversely, when pigs were fed with the AC diet, changes in the microbial population were detected in the proximal colon (Figure 6) and formed a distinctive cluster.

Because a sacrifice effect was previously observed in the biodiversity degree of the distal jejunum and in the cecum, dendrograms of these two intestinal tracts were generated with the animals sacrificed on day 19 and with the animals sacrificed on day 21 separately. Remarkably, in the distal jejunum, the animals euthanized on the first day showed a cluster depending on the diets, whereas the animals sacrificed on the second day did not show any specific association (Figure 7).

4. Discussion

The microbial population of the gastrointestinal tract of pigs has traditionally been studied by culture techniques [24, 25]. In recent years, molecular techniques have been introduced to improve the detection of bacteria, which are fastidious to culture [26, 27]. The use of these new techniques has allowed a better characterization of the composition of the intestinal microbiota.

In this work, the use of the PCR-RFLP technique permits a broader view of the composition of a complex bacterial ecosystem, such as the gastrointestinal tract of animals. In addition, the use of this technique allows us monitoring general variations that may occur in a bacterial population due to a change of the diet, time, and so forth. Also, with the PCR-RFLP technique we are exploring either unculturable and culturable bacteria or uncharacterized bacteria. It is estimated that the microbial community of the colon is composed of 400–500 different bacterial species [28]. A broader view of the ecosystem of the digestive tract can be obtained by PCR-RFLP. Although the number of resulting bands is relatively high (maximum 46), it cannot be ruled out that this technique might be underestimating the real degree of bacterial diversity. Such possibility may be due to a potential preferential amplification of certain bacterial species causing an insufficient amplification of the DNA of less prevalent bacteria. However, information provided by PCR-RFLP is always greater than that obtained from the cultivation of some specific bacterial groups.

In this study, using PCR-RFLP, it was possible to assess the diversity and similarity of the gastrointestinal microbial populations in relation to the different sections of the digestive tract and the different experimental diets. We can observe that microbial diversity increased significantly in more distal gastrointestinal segments than in the proximal sections. These results show a good concurrence with the results reported by Konstantinov et al. or Wang et al. [29, 30]. Several factors such as more neutral pH, slow intestinal transit, and/or low oxidation-reduction potential are associated with increased survival of bacteria in the hindgut [31]. In contrast, conditions applying to more proximal sections of the digestive tract (more acid environment, fast transit, and high bile acid concentration) make microbial diversity lower, with higher abundance of acid lactic bacteria [32]. Furthermore, when the band patterns are analyzed by creating dendrograms, proximal segments (stomach, proximal jejunum, and distal jejunum) and distal segments (cecum, proximal colon, distal colon, and rectum) show a tendency to be clustered. The ileum content samples formed a clearly distinct group, separated from both sections of the proximal bowel and distal intestinal segments. This result may be related to the fact that the ileum has some special features, such as pH being more similar to the caecum than to the contiguous small intestine. Besides, ileum has prominent Peyer's patches, a histological and immunological unique intestinal structure, which may affect its bacterial population.

Microbial diversity changed depending on the day of sacrifice in some parts of the digestive tract, especially in distal jejunum and cecum. The sacrifice effect was also observed when studying the degree of similarity in some of the dendrograms (CT diet). Besides, the animals euthanized on the first day showed a cluster of similarity depending on the diets, whereas this effect disappears on the second day of sacrifice. This effect could be related to the stress generated by the sacrifice of half of the experimental group (half of

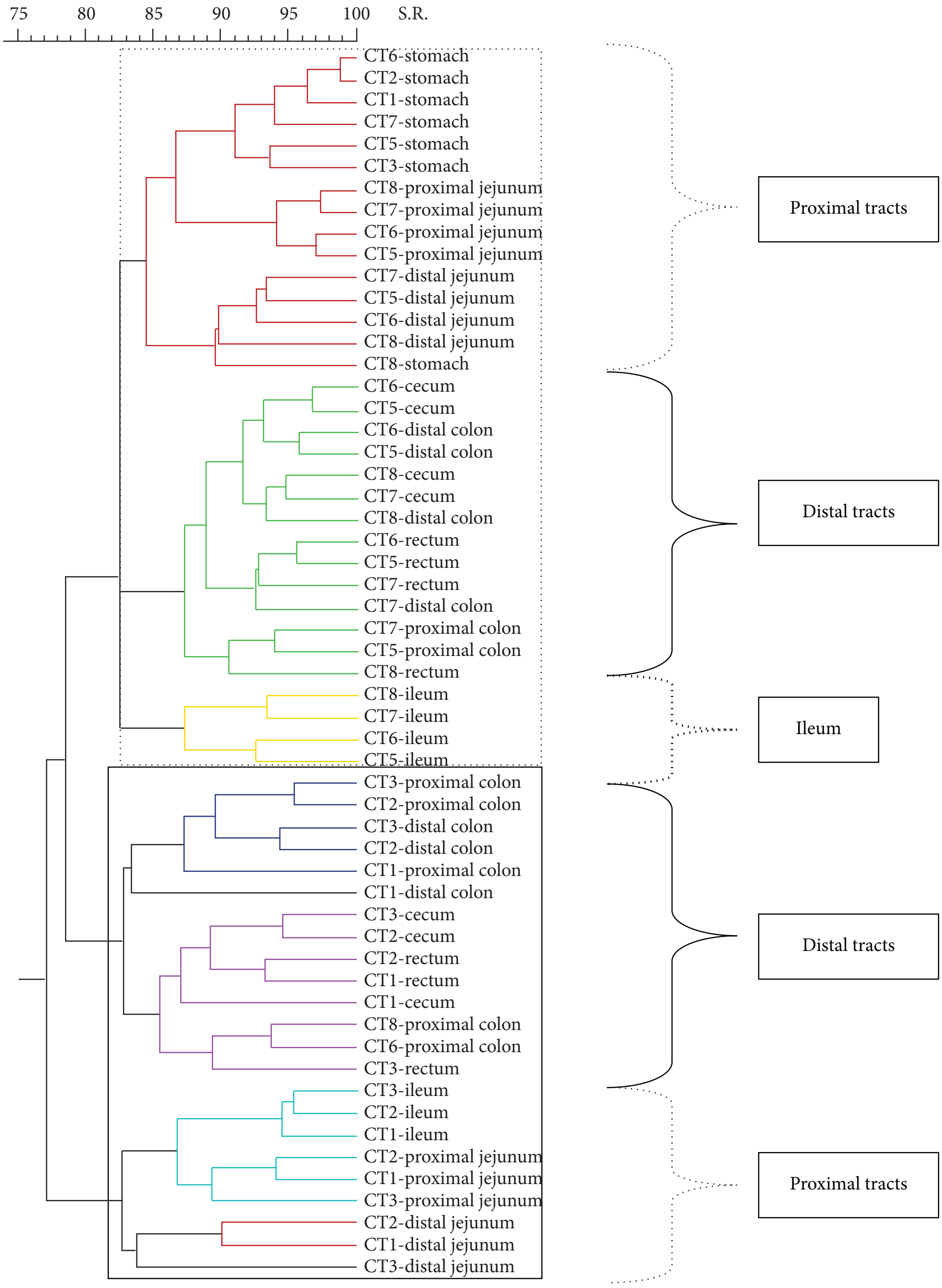


Figure 2: Dendrogram illustrating the similarity among band patterns obtained with PCR-RFLP among different gastrointestinal segments in pigs fed with a control diet (CT). Under the curly bracket represented with discontinuous line, the samples of proximal intestinal tracts are grouped; under the curly bracket with solid line the posterior intestinal tracts are clustered and under the curly bracket with thick discontinuous line the samples from the ileum tract are clustered. Inside the discontinuous line box the animals euthanized on day 21 are clustered and inside the continuous line box the animals euthanized on day 19 are clustered. The identification of each pig is shown in each sample.

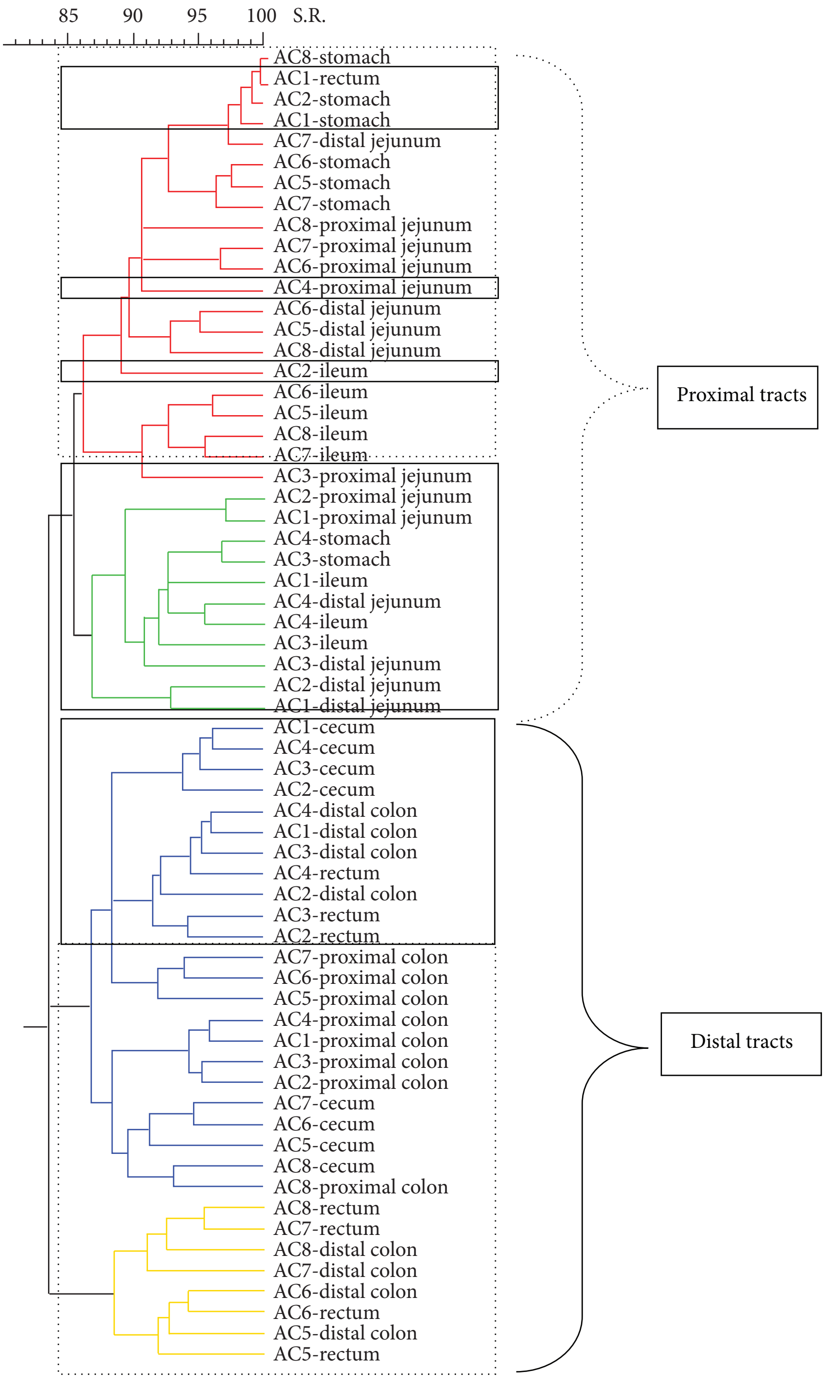


Figure 3: Dendrogram illustrating the similarity among band patterns obtained with PCR-RFLP among different gastrointestinal segments in pigs fed with a diet with sodium butyrate (AC). Under the curly bracket represented with discontinuous line, the samples of proximal intestinal tracts are grouped; under the curly bracket with solid line the posterior intestinal tracts are clustered. Inside the discontinuous line box the animals euthanized on day 21 are clustered and inside the continuous line box the animals euthanized on day 19 are clustered. The identification of each pig is shown in each sample.

the animals in a pen on first day) and to the subsequent modification of the hierarchy of this group. It has been shown that after acute stress, bacterial populations in the digestive tract can be altered quickly. Williams et al. [33] observed a decrease in the homogeneity of the band patterns after transporting the pigs. Consequently, the sacrifice effect and any factors that could alter the behavior of the animals may be important and should be taken into account in the design of experiments that analyze intestinal microbiota.

Apart from the ecological analysis of the microbiota throughout the gastrointestinal tract, this study also describes the effect of the incorporation of avilamycin, sodium

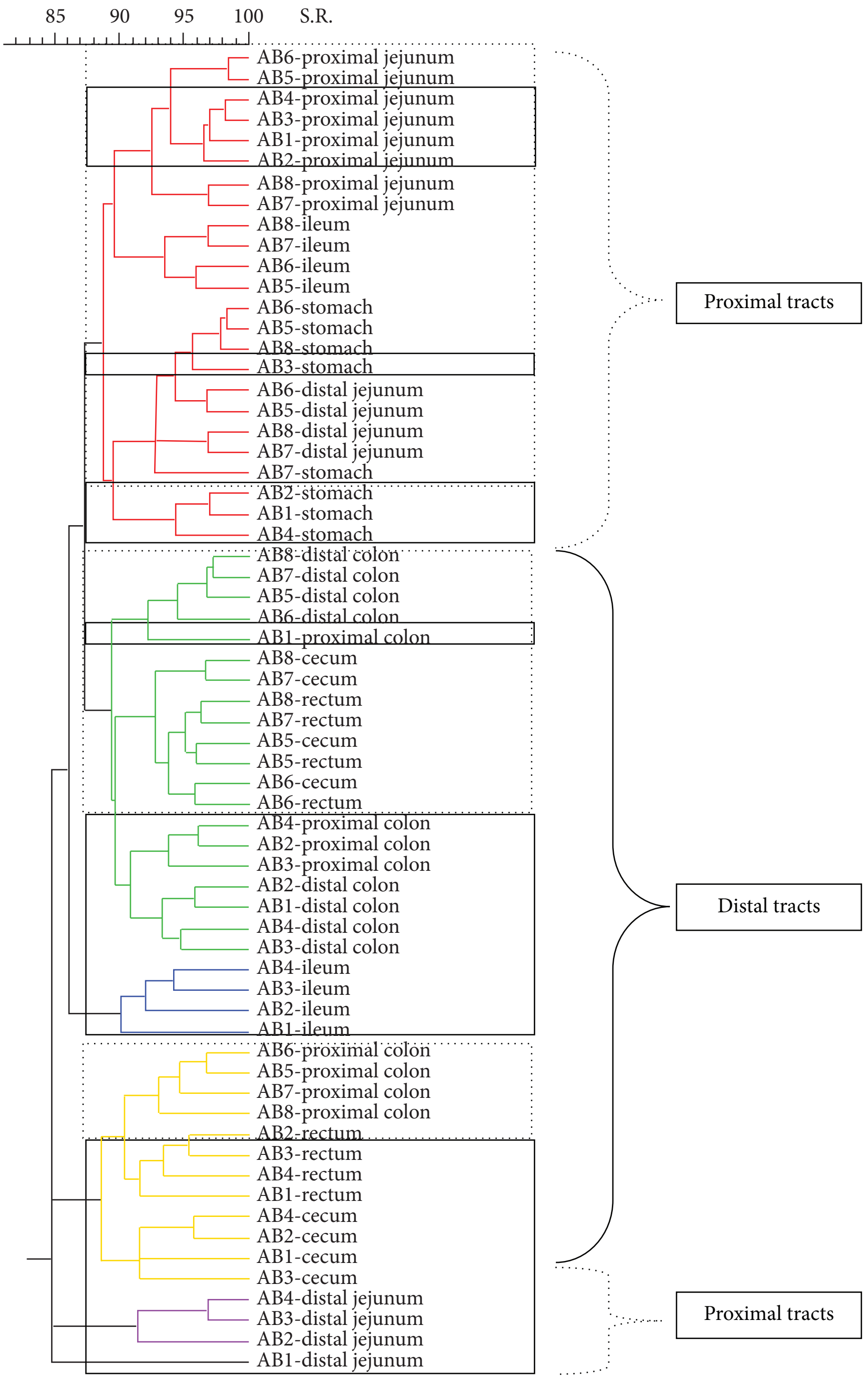


Figure 4: Dendrogram illustrating the similarity among band patterns obtained with PCR-RFLP among different gastrointestinal segments in pigs fed with a diet with avilamycin (AB). Under the curly bracket represented with discontinuous line, the samples of proximal intestinal tracts are grouped; under the curly bracket with solid line the posterior intestinal tracts are clustered. Inside the discontinuous line box the animals euthanized on day 21 are clustered and inside the continuous line box the animals euthanized on day 19 are clustered. The identification of each pig is shown in each sample.

butyrate, and a plant extract mixture on the gastrointestinal microbial populations. Adding these feed additives, the most significant differences are found in distal gastrointestinal segments, especially in the group of animals fed with the diet enriched with sodium butyrate or with a mixture of plant extracts. Thus, animals fed with the AC diet have a higher microbial diversity, while animals which were fed with plant extracts have a lower bacterial diversity. Gálfi and Bokori [14] observed changes in the ileal microbiota using 0.17% sodium n-butyrate in diets for weaned piglets. They detected a decrease in the proportion of coliform bacteria with a simultaneous increase in lactobacilli. Additionally,

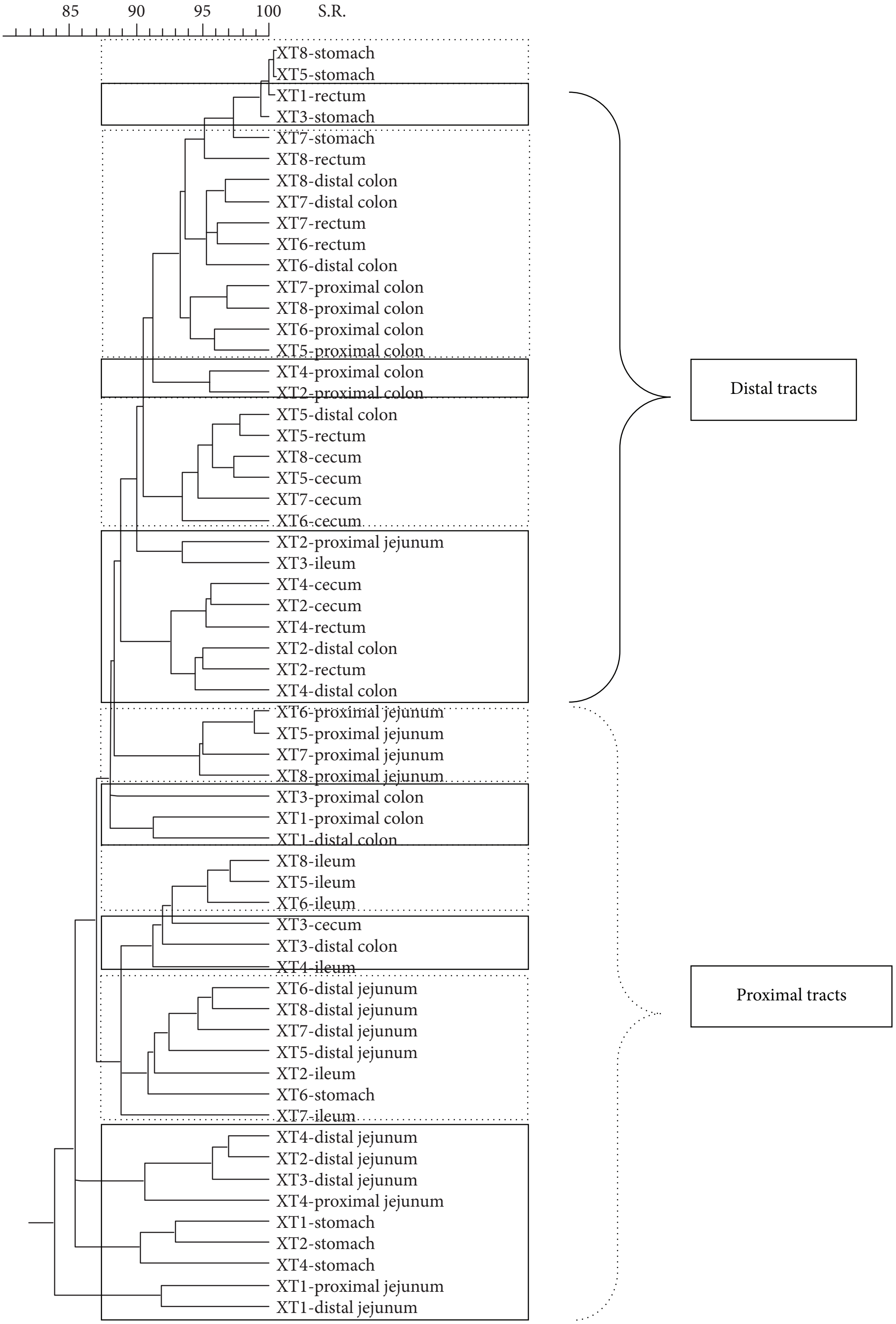


Figure 5: Dendrogram illustrating the similarity among band patterns obtained with PCR-RFLP among different gastrointestinal segments in pigs fed with a diet with a mixture of plant extracts (XT). Under the curly bracket represented with discontinuous line, the samples of proximal intestinal tracts are grouped; under the curly bracket with solid line the posterior intestinal tracts are clustered. Inside the discontinuous line box the animals euthanized on day 21 are clustered and inside the continuous line box the animals euthanized on day 19 are clustered. The identification of each pig is shown in each sample.

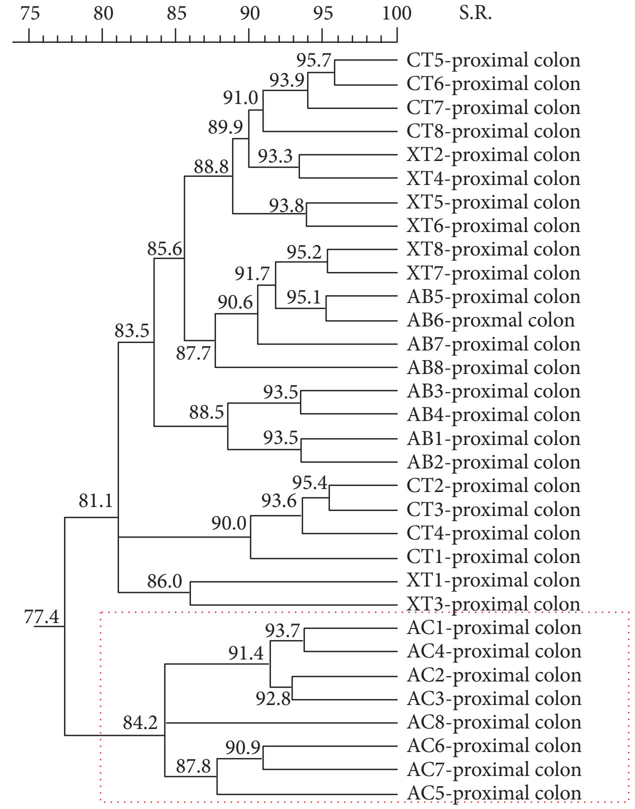


FIGURE 6: Ecological changes in microbial populations in proximal colon. Dendrogram illustrating the similarity among band patterns from proximal colon digest samples when comparing the four different experimental diets in weaning piglets. The experimental diets are indicated as control diet (CT) or the same diet with 0.04% avilamycin (AB), 0.3% sodium butyrate (AC), or 0.03% plant extract mixture (XT). The pig identification numbers are indicated for each sample.

Van Immerseel et al. [15], using microencapsulated butyric acid in young chickens, could also demonstrate a decrease colonization of *Salmonella* spp. in the caecum after an experimental infection. These two studies exhibit the evidence that butyrate acid exerts some effects on microbiota. However, there are no studies analysing the overall effect of butyrate acid on the microbial biodiversity along the gastrointestinal tracts. In the present study, an increase of microbial diversity was observed in proximal colon. The so called biodiversity has been proposed as an indicator of intestinal microbiota stability [34]. The increase of microbial diversity in proximal colon of AC pigs could be one of the factors that could explain the better numerical performance observed in these animals.

Moreover, other feed acidifiers have also demonstrated an effect on microbial diversity. Torrallardona et al. [35] detected an increased microbial diversity in the ileum using a diet with 0.5% of benzoic acid. On the other hand, Canibe et al. [36] detected a reduction of microbial diversity in both proximal intestinal segments and colon using a diet with formic acid. Factors such as the type of acid (a single acid or a mixture), dose, tolerance, and mode of action might explain these contradictory effects.

It is expected that the action of the organic acids will be higher in the proximal portions of the digestive tract (stomach and small intestine) [37]. However, in our study, the effect of sodium butyrate was observed on the bacterial ecosystem of the more distal intestinal segments rather than in the proximal portions. The increase in microbial diversity, found in animals fed with the AC diet, could be related to some direct effect (caused by a metabolite derived from the sodium butyrate) or an indirect effect (reduction of some bacterial species that, in turn, control the concentration of other bacterial species) of the sodium butyrate on bacterial populations of the proximal gastrointestinal tract. Thus, Van Winsen et al. [38] hypothesized that the number of bacteria in the Enterobacteriaceae family in the stomach determined the level of these bacteria in faeces. The authors attributed this result to an effect of the microbiota in proximal sections over the subsequent intestinal segments.

In a study performed in parallel and using the same samples as those used in this study [39], the authors found increased concentrations of butyric acid in the stomach due to its inclusion in the diet (CT = 4.87, AB = 5.11, and XT = 2.98 versus AC = 15.54, SEM = 0.970, P = 0.0001). However, the concentration of this acid showed no significant differences among experimental diets in other gastrointestinal segments. This could support the hypothesis that the effect of sodium butyrate on the hindgut microbiota might be an indirect effect.

The decrease in microbial diversity in animals fed with plant extracts has been previously reported in some studies [40]. This reduction in microbial diversity may be due to the antimicrobial effect of plant extracts, inhibiting some bacterial groups and promoting specific bacterial groups. In this case, in a parallel study using the same samples and the quantitative PCR technique, Castillo et al. [8] detected a significantly increased number of lactobacilli in the cecum in animals fed with the diet enriched with extracts of plants. However, in this same study, the authors found no significant differences in the quantity of total bacteria, either in animals fed with a diet enriched with sodium butyrate or in those fed with the diet enriched with plant extracts. These results indicate that these additives produce qualitative changes on the bacterial population of the gastrointestinal tract without affecting the total amount of bacteria. This fact has already been described in another study [36].

Our results show that animals fed with the control diet and animals fed with a diet enriched with avilamycin have similar microbial richness. This result is observed throughout the digestive tract. Collier et al. [41], using PCR-DGGE to study the bacterial composition in the ileum in pigs treated with tylosin for 21 days, observed a similar number of bands between the animals fed with the control diet and the animals receiving the food supplemented with the antibiotic. This outcome is attributed to an adaptation of the microbiota of antibiotic administration and may indicate a substitution of bacteria sensitive to antibiotics by bacteria resistant to them.

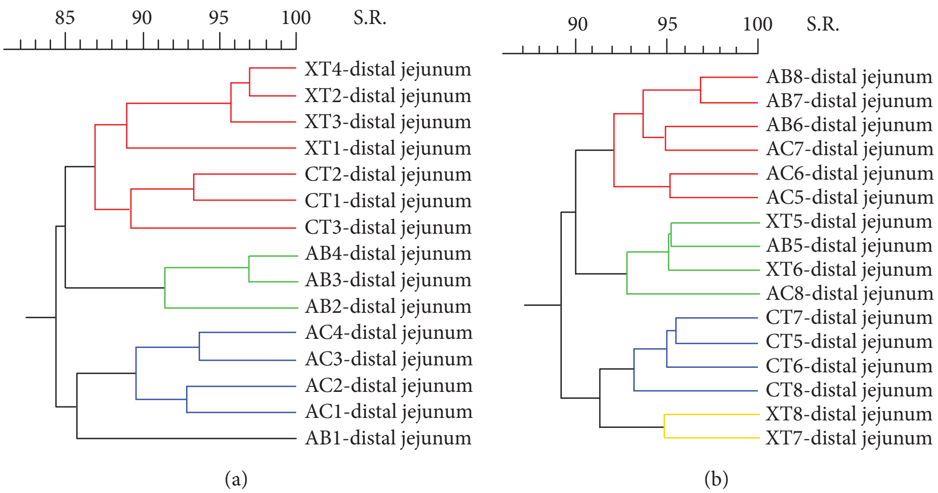


Figure 7: Ecological changes in microbial population in distal jejunum. Dendrograms show the percentage of similarity of the band patterns when comparing the four experimental diets for the animals sacrificed on day 19 (a) and for the animals sacrificed on day 21 (b). The experimental diets are indicated as control diet (CT) or the same diet with 0.04% avilamycin (AB), 0.3% sodium butyrate (AC), or 0.03% plant extract mixture (XT). The pig identification numbers are indicated for each sample.

5. Conclusion

In conclusion, changes in the complexity of the bacterial populations can be detected throughout the gut and with different additives by PCR-RFLP. Microbial diversity significantly increases from the small intestine to the large intestine. Additionally, obtained results suggest that the selection of proximal jejunum, ileum, and proximal colon are the most representative intestinal segments to study microbial diversity in pig gut. Besides, feed additives (such as sodium butyrate, a plant extract mixture, or avilamycin) are able to modify the gastrointestinal microbial ecology. The inclusion of sodium butyrate in weaned piglet diets increased the microbial biodiversity in distal intestinal segments, whereas the use of a mixture of plant extracts reduced it. More specific studies are required to clarify how these changes of the microbial diversity are significant factors to achieve a successful weaning process.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was financed by the Ministerio de Ciencia y Tecnología project (AGL2001-2621-C02-01). Mercè Roca received a grant from the Ministerio de Ciencia y Tecnología for this study. The authors thank Marta Muñoz and Noelia Benítez for accurate technical support and also thank Dr. Virgina Aragón, Dr. Marta Cerdà, Dr. Harry J. Flint, and Dr. Alex Olvera for helpful comments and Dr. Susagna Tubau for reviewing of the paper.

References

[1] P. Wallgren and L. Melin, "Weaning systems in the relation to disease," in *The Weaner Pig*, M. A. Varley and J. Wiseman, Eds., pp. 309–324, CABI Publishing, Nottingham, UK, 2001.

[2] G. L. Cromwell, "Why and how antibiotics are used in swine production," *Animal Biotechnology*, vol. 13, no. 1, pp. 7–27, 2002.

[3] B. B. Jensen, "The impact of feed additives on the microbiological ecology of the gut in young pigs," *Journal of Animal and Feed Sciences*, vol. 7, pp. 45–64, 1998.

[4] H. J. D. Dorman and S. G. Deans, "Antimicrobial agents from plants: antibacterial activity of plant volatile oils," *Journal of Applied Microbiology*, vol. 88, no. 2, pp. 308–316, 2000.

[5] A. Smith-Palmer, J. Stewart, and L. Fyfe, "Antimicrobial properties of plant essential oils and essences against five important food-borne pathogens," *Letters in Applied Microbiology*, vol. 26, no. 2, pp. 118–122, 1998.

[6] D. Jamroz and C. Kamel, "Plant extracs enhance broiler performance," *Journal of Animal Science*, vol. 80, p. 41, 2002.

[7] E. G. Manzanilla, J. F. Perez, M. Martin, C. Kamel, F. Baucells, and J. Gasa, "Effect of plant extracts and formic acid on the intestinal equilibrium of early-weaned pigs," *Journal of Animal Science*, vol. 82, no. 11, pp. 3210–3218, 2004.

[8] M. Castillo, S. M. Martín-Orúe, M. Roca et al., "The response of gastrointestinal microbiota to avilamycin, butyrate, and plant extracts in early-weaned pigs," *Journal of Animal Science*, vol. 84, no. 10, pp. 2725–2734, 2006.

[9] K. H. Partanen and Z. Mroz, "Organic acids for performance enhancement in pig diets," *Nutrition Research Reviews*, vol. 12, no. 1, pp. 117–145, 1999.

[10] B. R. Paulicks, F. X. Roth, and M. Kirchgessner, "Dose effects of potassium diformate (FormiŮ LHS) on the performance of growing piglets," *Agribiological Research*, vol. 49, no. 4, pp. 318–326, 1996.

[11] B. Gedek, M. Kirchgessner, U. Eidelsburger, S. Wiehler, A. Bott, and F. Roth, "Influence of formic acid on the microflora in different segments of the gastrointestinal tract 5. Nutritive value of organic acids in piglet rearing," *Journal of Animal Physiology and Animal Nutrition*, vol. 76, pp. 206–214, 1992.

[12] K. Partanen, *Organic Acids—Their Efficacy and Modes of Action in Pigs*, Nottingham University Press, Nottingham, UK, 2001.

[13] K. R. Gardiner, S. J. Kirk, and B. J. Rowlands, "Novel substrates to maintain gut integrity," *Nutrition Research Reviews*, vol. 8, pp. 43–66, 1995.

[14] P. Gálfi and J. Bokori, "Feeding trial in pigs with a diet containing sodium n-butyrate," *Acta Veterinaria Hungarica*, vol. 38, no. 1-2, pp. 3–17, 1990.

[15] F. van Immerseel, J. B. Russell, M. D. Flythe et al., "The use of organic acids to combat Salmonella in poultry: a mechanistic explanation of the efficacy," *Avian Pathology*, vol. 35, no. 3, pp. 182–188, 2006.

[16] D. B. Anderson, V. J. McCracken, R. I. Aminov et al., "Gut microbiology and growth promoting antibiotics in swine," *Nutrition Abstracts and Reviews B*, vol. 70, pp. 101–108, 1999.

[17] D. Hardy, D. Amsterdam, L. A. Mandell, and C. Rotstein, "Comparative in vitro activities of ciprofloxacin, gemifloxacin, grepafloxacin, moxifloxacin, ofloxacin, sparfloxacin, trovafloxacin, and other antimicrobial agents against bloodstream isolates of gram-positive cocci," *Antimicrobial Agents and Chemotherapy*, vol. 44, no. 3, pp. 802–805, 2000.

[18] M. Castillo, G. Skene, M. Roca et al., "Application of 16S rRNA gene-targetted fluorescence in situ hybridization and restriction fragment length polymorphism to study porcine microbiota along the gastrointestinal tract in response to different sources of dietary fibre," *FEMS Microbiology Ecology*, vol. 59, no. 1, pp. 138–146, 2007.

[19] M. A. Murphy, L. P. Waits, K. C. Kendall, S. K. Wasser, J. A. Higbee, and R. Bogden, "An evaluation of long-term preservation methods for brown bear (*Ursus arctos*) faecal DNA samples," *Conservation Genetics*, vol. 3, no. 4, pp. 435–440, 2002.

[20] G. Muyzer, E. C. De Waal, and A. G. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA," *Applied and Environmental Microbiology*, vol. 59, no. 3, pp. 695–700, 1993.

[21] G. Muyzer, A. Teske, C. O. Wirsen, and H. W. Jannasch, "Phylogenetic relationships of Thiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments," *Archives of Microbiology*, vol. 164, no. 3, pp. 165–172, 1995.

[22] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Chicago, Ill, USA, 1963.

[23] L. Kaufmann and P. J. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley and sons, New York, NY, USA, 1990.

[24] I. M. Robinson, M. J. Allison, and J. A. Bucklin, "Characterization of the cecal bacteria of normal pigs," *Applied and Environmental Microbiology*, vol. 41, no. 4, pp. 950–955, 1981.

[25] I. M. Robinson, S. C. Whipp, J. A. Bucklin, and M. J. Allison, "Characterization of predominant bacteria from the colons of normal and dysenteric pigs," *Applied and Environmental Microbiology*, vol. 48, no. 5, pp. 964–969, 1984.

[26] T. D. Leser, J. Z. Amenuvor, T. K. Jensen, R. H. Lindecrona, M. Boye, and K. Moøller, "Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited," *Applied and Environmental Microbiology*, vol. 68, no. 2, pp. 673–690, 2002.

[27] S. E. Pryde, A. J. Richardson, C. S. Stewart, and H. J. Flint, "Molecular analysis of the microbial diversity present in the colonic wall, colonic lumen, and cecal lumen of a pig," *Applied and Environmental Microbiology*, vol. 65, pp. 5372–5377, 1999.

[28] W. N. Ewing and D. J. A. Cole, "The microbiology of the gastrointestinal tract," in *The Living Gut. An Introduction to Microorganisms in Nutrition*, W. N. Ewing and D. J. A. Cole, Eds., pp. 45–65, Context Publications, Dungannon, Ireland, 1994.

[29] S. R. Konstantinov, A. Awati, H. Smidt, B. A. Williams, A. D. L. Akkermans, and W. M. De Vos, "Specific response of a novel and abundant Lactobacillus amylovorus-like phylotype to dietary prebiotics in the guts of weaning piglets," *Applied and Environmental Microbiology*, vol. 70, no. 7, pp. 3821–3830, 2004.

[30] X. Wang, S. P. Heazlewood, D. O. Krause, and T. H. J. Florin, "Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis," *Journal of Applied Microbiology*, vol. 95, no. 3, pp. 508–520, 2003.

[31] C. S. Stewart, "Microorganisms in hindgut fermentors," in *Gastrointestinal Microbiology*, R. I. Mackie and B. A. White, Eds., vol. 2, pp. 142–186, Chapman and Hall Microbiology Series, New York, NY, USA, 1997.

[32] J. H. Cummings and G. T. Macfarlane, "The control and consequences of bacterial fermentation in the human colon," *Journal of Applied Bacteriology*, vol. 70, no. 6, pp. 443–459, 1991.

[33] J. L. Williams, J. E. Minton, J. A. Patterson, J. Marchant Forde, and S. D. Eichert, "Lairage during transport of eighteen-kilogram pigs has an impact on innate immunity and commensal bacteria diversity in the intestines," *Journal of Animal Science*, vol. 86, no. 5, pp. 1232–1244, 2008.

[34] E. G. Zoetendal, C. T. Collier, S. Koike, R. I. Mackie, and H. R. Gaskins, "Molecular ecological analysis of the gastrointestinal microbiota: a review," *Journal of Nutrition*, vol. 134, no. 2, pp. 465–472, 2004.

[35] D. Torrallardona, I. Badiola, and J. Broz, "Effects of benzoic acid on performance and ecology of gastrointestinal microbiota in weanling piglets," *Livestock Science*, vol. 108, no. 1–3, pp. 210–213, 2007.

[36] N. Canibe, O. Højberg, S. Højsgaard, and B. B. Jensen, "Feed physical form and formic acid addition to the feed affect the gastrointestinal ecology and growth performance of growing pigs," *Journal of Animal Science*, vol. 83, no. 6, pp. 1287–1302, 2005.

[37] N. Canibe, S. H. Steien, M. Øverland, and B. B. Jensen, "Effect of K-diformate in starter diets on acidity, microbiota, and the amount of organic acids in the digestive tract of piglets, and on gastric alterations," *Journal of Animal Science*, vol. 79, no. 8, pp. 2123–2133, 2001.

[38] R. L. Van Winsen, B. A. P. Urlings, L. J. A. Lipman et al., "Effect of fermented feed on the microbial population of the gastrointestinal tracts of pigs," *Applied and Environmental Microbiology*, vol. 67, no. 7, pp. 3071–3076, 2001.

[39] E. G. Manzanilla, M. Nofrarías, M. Anguita et al., "Effects of butyrate, avilamycin, and a plant extract combination on the intestinal equilibrium of early-weaned pigs," *Journal of Animal Science*, vol. 84, no. 10, pp. 2743–2751, 2006.

[40] H. Namkung, M. Li, H. Yu, M. Cottrill, J. Gong, and C. F. M. deLange, "Impact of feeding blends of organic acids or

herbal extracts on growth performance, gut microflora and digestive funtion in newly weaned pigs," in *Proceedings of the 9th International Symposium on Digestive Physiology in Pigs*, pp. 93–95, Canada, 2003.

[41] C. T. Collier, M. R. Smiricky-Tjardes, D. M. Albin et al., "Molecular ecological analysis of porcine ileal microbiota responses to antimicrobial growth promoters," *Journal of Animal Science*, vol. 81, no. 12, pp. 3035–3045, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 587460, 8 pages
http://dx.doi.org/10.1155/2013/587460

Review Article

The Variety of Vertebrate Mechanisms of Sex Determination

Antonina V. Trukhina, Natalia A. Lukina,
Natalia D. Wackerow-Kouzova, and Alexander F. Smirnov

Department of Genetics and Biotechnology, Saint-Petersburg State University, Saint-Petersburg 199034, Russia

Correspondence should be addressed to Alexander F. Smirnov; afsmirnov@bio.pu.ru

Received 27 February 2013; Revised 26 September 2013; Accepted 28 October 2013

Academic Editor: Vassily Lyubetsky

The review deals with features of sex determination in vertebrates. The mechanisms of sex determination are compared between fishes, amphibians, reptilians, birds, and mammals. We focus on structural and functional differences in the role of sex-determining genes in different vertebrates. Special attention is paid to the role of estrogens in sex determination in nonmammalian vertebrates.

1. Peculiar Properties of Sex Determination among Different Vertebrates

One of the most fundamental and surprisingly diverse processes in the life history of an organism is the determination of sex. Its determination must be a very ancient process, with male and female sexes recognized in diverse organisms, from corals to worms, insects, fishes, birds, and mammals [1–4]. The transformation of three stages of sex formation (chromosomal predetermination, sex determination, and sex differentiation) was described in different groups of vertebrate. Sex determination at vertebrates has long been equated with gonadal differentiation into ovaries or testes. Consideration of different taxonomic animal groups allowed for establishing two general mechanisms of sex determination: genetic sex determination (GSD) and external sex determination (ESD). ESD at vertebrates is practically reduced to the temperature sex determination (TSD). Birds and mammals are characterized only by GSD, whereas crocodiles by TSD. Lizards, snakes, turtles, and bony fishes were described to have all possible mechanisms of sex determination [1]. There are two genetic sex determination systems: with heterogametic male—XY (mammals) and heterogametic females—ZW (birds). Note that both genetic systems are found in amphibians [1].

In organisms with heteromorphic sex chromosomes, such as birds and mammals, sex is set at fertilization by the differential inheritance of sex chromosomes [2–4]. The logical assumption is that sex-determining genes, inherited at fertilization, become active in the gonads during embryonic or larval life. However, the various reports of somatic sexual dimorphisms preceding the gonadal development call for a more considered definition of sex determination [4].

TSD was firstly discovered in reptiles: turtles, crocodiles, but not in snakes. The primary mechanism of sex determination is poorly understood. It obviously occurs in species with undifferentiated Y-chromosome. Higher temperature can produce either males or females, and the temperature ranges and lengths of exposure that influence TSD are remarkably variable among species. The classical view proposed that the developing gonads in vertebrate have the bipotential genital ridges: the cortex and the medulla. Thereafter, the process of sex differentiation depends on the alternative development of these two territories. Ovaries develop from the growing cortex, while testes develop from the medulla with an apparent antagonism between the two processes [5].

Although most genes involved in gonadal development are conserved at vertebrates, including species with TSD, the temporal and spatial gene expression patterns vary among species. Aromatase (CYP19), which regulates gonadal estrogen level, is proposed to be the main target of a putative thermosensitive factor for TSD. It is known that the estrogen levels may influence sex determination or gonad differentiation depending on the species. Yolk steroids of maternal origin and steroids produced by the embryonic nervous system should also be considered as sources of hormones that may play a role in TSD. It is an exception that different taxonomic groups of animals with TSD have different sex determination mechanisms. Moreover, there are thermosensitive genes: in *Emydidae*—*sox9*, in *Testudinae*—*sox9*, *sf1*, and *wt1*, and in *Emydidae*—*dax1*. It was proposed that in the case of temperature-dependent sex determination, a gene chain *amh-sox9* operates affecting the appearance of testes in contrast to the other chain genes in mammalian sex determination *sry-sox9-amh* [5, 6].

Teleost fishes (over 30,000 species) are the largest group of vertebrates which exhibit a remarkable variety of sexuality. Fish sexualities were categorized into gonochorism, synchronous/sequential hermaphrodite, or unisexual reproduction. Sex at fishes is determined genetically or by environmental factors [7]. The only known exception of unisexual species is the amazon mollies—*Poecilia formosa*. Sequential hermaphrodite (sex-changing) species have been recorded in 27 of 448 families across 7 orders of fishes, most of which have found a niche in coral reefs. In these fishes, the gonadal sex redifferentiation was observed during sex change in adulthood. Thus, the sex-changing fishes are ideal models to investigate gonadal differentiation in vertebrates. Wrasses (*Labridae*) are the major group of species that display protogynous sex change. Individuals of this species initially mature as either males or females. Protandrous and protogynous sex changes are generally irreversible. Thus, such sex changes happen once in a life cycle. There are fish species with XX/XY and the most common ZW/ZZ sex determination system; the unusual WXZ system is described in swordtail (platyfish) [7]. Not many fish species have morphologically different sex chromosomes (10%) and most of them are at an early stage of differentiation. Even in the case of animals where sex is determined by genetic factors, the molecular processes that lead to the formation of either testis or ovary are evolutionary labile. The sex determination in most therian mammals is triggered by the testis-determining gene *sry*. This role is played by *dmy*/*dmrt1bY* and *dmrt1* in medaka (*Oryzias latipes*) and chicken, respectively. Teleost fishes represent nearly half of all extant vertebrates and show a wide variety of sex determination mechanisms. Their sex can be determined by genetic factors or/and environmental factors. Recently, four novel sex determination (SD) genes or strong SD gene candidates in vertebrates were reported, all of them are in fishes: *amhy* in the Patagonian pejerrey *Odontesthes hatcheri*, *gsdf* in *Oryzias luzonensis* (a relative of medaka), *amhr2* in fugu *Takifugu rubripes*, and *sdY* in rainbow trout *Oncorhynchus mykiss* [8]. Complex epistatic sex system at fishes has been found consisting of a major female heterogametic ZW locus on chromosome 5, two separate male heterogametic XY loci on chromosome 7, and two additional interacting loci on chromosomes 3 and 20 [9].

Fishes have the most plastic system of germ and somatic cells in comparison with other animals. For them, the plasticity is maintained throughout the life cycle. It describes the impact on the process of such factors as temperature, pH, and population density. TSD at fish is less common than previously thought. The effect of estrogen acting through ER directly or indirectly regulates P450arom and AMH. The analysis of the differences between gonochoristic and hermaphroditic fish species will help to understand the mechanism of plasticity of sex determination in vertebrates [7].

Amphibians have two systems of sex chromosomes: one with heterogametic male (XX/XY) and another with the female (ZZ/ZW). Most urodeles (urodele salamanders) have XX/XY system. For 63 of 1500 species with determined sex, only 20 of the total number had differing sex chromosomes. There are heterogametic males in the ancestral species of toads *Leiopelma hamiltoni* and *L. hochstetteri*. Models of sex differentiation in amphibians can be divided into three types: (1) direct development of the undifferentiated gonad in testes or ovaries; (2) undifferentiated gonad development in the ovaries and testes through subsequent appearance of ovary; and (3) semidifferentiated called type-development phase of the testes of intersex. Most amphibians do not exhibit morphologically distinguishable sex chromosomes. In *Rana rugosa*, the X and W and also the Y and Z chromosomes are almost identical to each other based on the morphology and replication banding patterns, respectively. The Z shares its origin with the Y and the X—with the W. For a long time no sex-determining genes have been identified in amphibians. Recently [10, 11], a candidate for an ovary-determining gene, or *DM-W*, a W-linked DM-domain of gene was isolated from *Xenopus laevis*. However, the target gene downstream of *DM-W* is not known. To date, *DM-W* has not been found in any species of amphibian other than *X. laevis*. When *DM-W* is introduced into unfertilized eggs, ZZ transgenic male tadpoles form ovaries [10, 11].

Dmrt1 gene is an autosomal gene in *X. laevis*. Moreover, there is the fact that the phenotypic sex of *X. laevis* carrying a pair of Z chromosomes can be altered from male to female by estrogens. It suggests that the mechanism of sex determination is flexible in this species. Rigid sex determination would need a prolonged expression of sex-determining gene in *X. laevis*. In *R. rugosa* the expression of steroidogenic genes, such as *cyp11a1*, *star*, *hsd3b*, *cyp17*, *hsd17β*, and *cyp19* and gene encoding 5-α-reductase, starts in the indifferent gonads of male and female tadpoles prior to sex determination [11].

At present time, the sex-determining genes in some species at amphibians are not known. However, apart from the sex-determining genes, genes involved in gonadal differentiation may be conserved in all classes of vertebrates. This is supported by the fact that genes such as *foxl2*, *dmrt1*, *wt1*, *sox9*, *sf1*, *cyp19*, and *dax1* are evolutionary conserved genes from fish to mammals. The exogenous steroid hormones can determine the phenotypic sex of many species of amphibians. For example, XX female-to-male sex reversal in *R. rugosa* can be induced by testosterone although XX-females do not carry a male-determining gene on the X chromosome. Thus, a sex-determining gene is not actually necessary for sex determination in amphibians. In the other words, the steroid hormones could be the key factor for sex determination in amphibians. A male sex-determining gene, if it exists, probably supports steroid hormones to direct indifferent gonads to a testis by inhibiting *cyp19* transcription for ovarian formation. At present, factors for upregulation of *cyp17* in

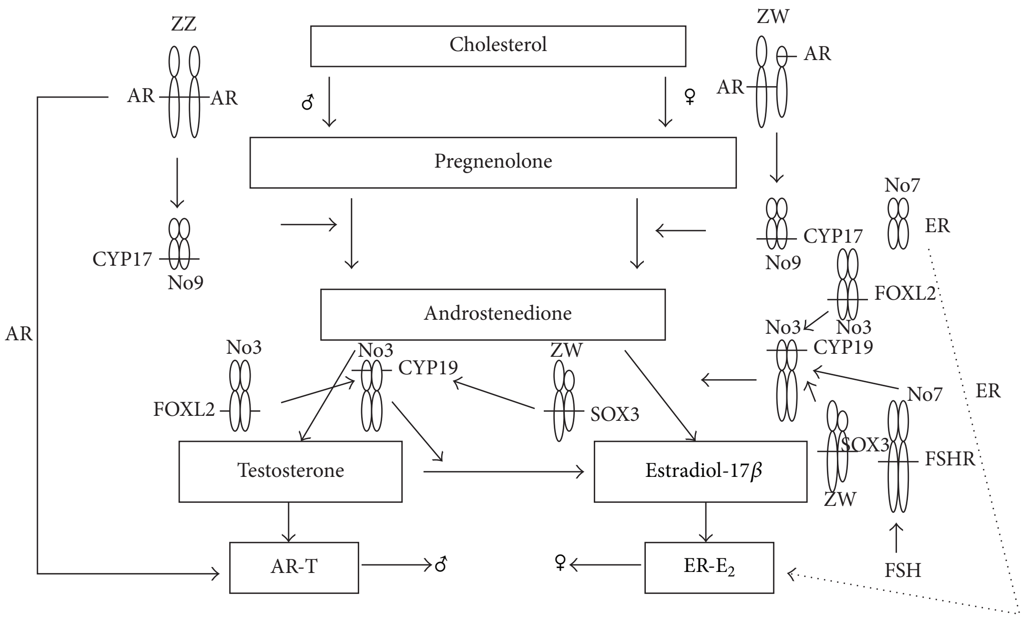


Figure 1: This schema presents sex determination by steroid hormones and some SD-genes in *Rana rugosa*. Testosterone and estradiol-17β are produced in the undifferentiated gonads of males and females, respectively. Sex chromosomes (Z, W) and autosomes (3, 9, 7) were denoted by letters and the numbers, respectively. AR-T and ER-E_2 indicate the complexes of steroid receptor. Localization of genes on chromosomes marked lateral line (adopted from [10, 11]).

Table 1: The role of estrogen in sex determination of vertebrates and phylogenetic distance (adopted from [12]).

Group of vertebrata	Distance from mammals (million years ago, MYR)	Influence of estrogen on sex determination
	Nonmammalian vertebrata	
Pisces	450–530	+
Amphibia	300	+
Reptilia	290	+
Aves	216–199	+
	Theria	
Monotremata	180	+
Marsupialia	160	+
Eutheria (Mammalia)	0	—

the indifferent gonad of *R. rugosa* remain to be identified (Figure 1) [10, 11].

2. Estrogens and Nonmammalian Vertebrate Sex Determination

Estrogen is both necessary and sufficient to drive ovarian development in many nonmammalian vertebrates (Table 1) [12]. Moreover there is the actual material about this hormone is able influence not only the differentiation of sex but also the appearance of specific sex gonads (sex determination) (Table 2) [13]. However, the role of estrogen in the mammalian gonad is less clear. Mouse ovarian development can proceed in the absence of estrogen signaling, but granulosa cell fate cannot be maintained. Estrogen receptor expression is conserved in the indifferent gonad of all mammals and many species also express the *cyp19* gene that encodes aromatase in the early ovary [13].

Furthermore, the estrogen is sufficient to drive ovarian development of the indifferent gonad in marsupial mammals. Estrogen treatment in alligators and turtles at the male-producing temperature induced ovarian transformation, including proliferation and entry of germ cells in the cortex into meiosis, and at the same time, degeneration of sex cords in the medullary region. Estrogen appears to promote the ovarian fate by stimulating the expansion of the cortex while inhibiting the maintenance of sex cords [13]. Animals utilizing TSD are susceptible to increases in temperature, as well as exposure to chemicals, such as synthetic estrogens. These factors have the potential to skew sex ratios. It is possible that controlling sex at the chromosomal level evolved as a protective mechanism in order to shield embryos from the changing environment [12, 13].

Birds and marsupials are unique because their sex is determined by classical GSD mechanism although the embryo remains sensitive to the effects of estrogen. It is

Table 2: Mechanisms of sex determination and experimental sex reversal in vertebrates (adopted from [13]).

Mechanism of sex determination	Species	Effectors	Experimental evidence of sex reversal
TSD (t°C)	*American alligator and Red-eared slider turtle*	Aromatase or Estradiol (E_2)	F-M: administration of an aromatase inhibitor introduced at F-producing temperatures; M-F: E_2 administration to eggs incubated at male-producing temperatures
GSD (XX/XY; ZZ/ZW)	*Oryzias latipes, Bufo bufo, and Xenopus laevis*	Androgene or E_2	—
GSD (ZZ/ZW)	*Gallus gallus*	Aromatase or Estradiol (E_2)	F-M: in ovo administration of an aromatase inhibitor to ZW animals; M-F: in ovo administration of E_2 to ZZ individuals
GSD (XX/XY)	*Macropus eugenii and Mus musculus*	*Sry*	F-M: treatment of XX individuals with Müllerian inhibiting substance (MIS) to cause germ loss, addition of *Sry* gene in XX embryos; loss of E_2 through ER$\alpha\beta$KO causes transdifferentiation of the ovary to testis-like structures in the adult; M-F: inactivation of *Sry* gene in XY embryos, E_2 administration to XY

Here: F: female, M: male, E_2: estrogene, X, Y, Z, W: sex chromosomes.

especially so for the chicken *Gallus gallus* and tammar wallaby *Macropus eugenii*, the most characterized models for avian and marsupial species, respectively. Although birds diverged from reptiles approximately 245 million years ago, the ovary-determining action of estrogen remains present in birds with the evolution of a genetic mechanism for sex determination. This phenomenon is also found in marsupials, the close relatives to the eutherian mammals. Most marsupials, such as the tammar wallaby, are born underdeveloped with a mixture of fetal and neonatal characteristics. Gonadal development in the marsupial is similar to that of mice and humans except that the development of marsupial fetuses occurs outside of the uterus. Marsupial gonads are sexually indifferent at birth, and gonadal differentiation commences immediately after birth. By the 7th postnatal day, the ovarian differentiation has progressed to the point where the gonad is morphologically distinguishable. In the tammar wallaby, unlike birds and reptiles, the ovary does not produce estrogen at the time of gonadal differentiation. In fact, the ovarian steroid production does not commence until almost 200 days after birth. Based on this fact, it appears that estrogen is not necessary for the differentiation of the ovary as it is in birds and reptiles. However, the administration of estrogen to wallaby male embryos immediately after birth can induce complete ovarian development similar to reptiles and birds. The ability of estrogen to feminize the male gonad in the marsupial suggests that estrogen can override the genetic components derived from the XY mechanism. It is possible to induce chicken sex reversal by using inhibitor of aromatase or the analogs of estrogen (Table 2) [13–16].

Mammals evolved from lower vertebrates approximately 80 million years ago. Accompanying that recent evolution was a new form of sex determination, GSD that utilizes only chromosomal composition to determine sex. Unlike birds and marsupials that use GSD but remain sensitive to steroids, the eutherian mammals evolved a mechanism whereby sex determination is completely resistant to steroids. While it appears that estrogen has no impact on primary sex determination in the developing eutherian embryo, a different situation arises after birth. In the adult mammalian female, estrogen plays a vital role in maintaining the ovary. In the aromatase knockout mouse, testicular cell types and structures arose in postpubertal ovaries. A similar phenomenon was observed in mice lacking both estrogen receptors α and β (ER$\alpha\beta$KO), where Sertoli cells and seminiferous-like structures appeared in the ovaries after puberty (Table 2). These observations bring about an intriguing hypothesis that the primitive estrogen-induced mechanism of ovary development remains present in eutherian females [13].

According to Pask [12], in the presence of estrogen, key male differentiation genes fail to be upregulated in the XY gonad and instead key ovary-promoting genes are upregulated leading to ovarian development. Estrogen appears to trigger sex reversal through the exclusion of *sox9* from entering the nucleus in the somatic cells of the developing gonad of nonmammalian vertebrate. In the absence of nuclear *sox9*, Sertoli cell development cannot be initiated and the somatic cells follow a granulosa cell fate. A conserved role for estrogen-mediating *sox9* action is consistent with several observations in mammals. In mice, *sox9* is able to autoregulate by binding to its own promoter. Activated estrogen receptor complexes can also move into the nucleus and along with *foxl2*, suppress *sox9* transcription by directly binding to the *sox9* enhancer, TESCO.

Table 3: The known sex-determining genes in vertebrate (adopted from [18]).

Species	The known sex-determing genes	The main peculiarities of the genes
	Mammals	
Homo sapiens, Mus musculus, and so forth.	*Sry*	Transcription factor, upregulator of *Sox9*; and testis-determining gene.
	Birds	
Gallus gallus domesticus, and so forth.	*Dmrt1*	Transcription factor, upregulator of *Sox9,* DM domain gene; and testis-determining gene.
	Amphibia	
Xenopus laevis	*DM-W*	Transcription factor, DM domain gene, and ovary-determining gene.
	Fishes	
Oryzias latipes	*Dmy*	Transcription factor, DM domain gene, and testis determining gene.
Oryzias luzonensis	*Gsdf*	Secretory protein belonging to the TGF-β superfamily.
Takifugu	*Amhr2*	Receptor for Amh.
Patagonian Pejerrey	*Amhy*	The Amh protein has been implicated in the regulation of germ cell proliferation and spermatogenesis.
Oncorhynchus mykiss	*SdY*	A novel protein that displays sequence homology with carboxy-terminal domain of interferon regulatory factor 9 (Irf9).

3. Sex-Determining Genes in Vertebrates

Although the molecular mechanisms underlying many developmental events are conserved across vertebrate taxa, the lability at the top of the sex-determining (SD) cascade has been evident from the fact that four master SD genes have been identified: *sry* (mammalian), *dmrt1* (chicken), *dmy* (medaka), and *DM-W* (*Xenopus laevis*) [8, 18]. Recently four novel candidates for vertebrate SD genes were reported, all of them are in fishes. These include *amhy* in the Patagonian pejerrey, *gsdf* in *Oryzias luzonensis, amhr2* in fugu, and *sdY* in rainbow trout (*Salmo gairdneri*) (Table 3) [18]. This large-scale natural experiment provides a resource that geneticists can use to search genetic variants of sex determination control. Accumulation of knowledge on such variants will allow us to distinguish conserved and diversified SD pathways among vertebrates and lead us into a deeper understanding of the vertebrate SD cascade. Given the evidence from all the fish species mentioned, together with previous studies of other nonmammalian species, it seems reasonable to imagine that many other teleosts, reptiles and amphibians also have experienced a turnover of sex chromosomes [8].

The comparison of new and old SD genes indicates that vertebrate sex-determining cascades are not as conserved as once thought. In eutherian mammals, *sry* is a recently evolved key Y-linked testis determinant. *Sry* is absent from the genome outside therian mammals (marsupials and placentals). In birds and lower vertebrates, there is a pervasive role for DM domain of genes in gonadal sex differentiation, and it is considered that these genes have an ancient association with sex. In mammals, the current evidence favors the idea that *sry* acts with the orphan nuclear receptor Sf1 to activate *sox9* expression in the developing XY gonad. *Sry* is turned off by *sox9* which then maintains its own expression. However, in the chicken embryo, *dmrt1* expression precedes that of *sox9* by at least 2 days (the 4th embryonic day versus the 6th day), implying that other intervening genes are involved. Interestingly, in the medaka fish, gene *sox9b* (the orthologue of tetrapod *sox9*) is not involved in testis determination, but has a function in germ cells. In lower vertebrates, *sry* is also absent and other triggers must exist. Given the different, independent origins of sex chromosomes among birds, reptiles, and amniotes, a variety of different testis-determining triggers are likely. In birds, testis development requires presumably the conserved Z-linked *dmrt1* gene. Yet, *dmrt1* is not sex-linked in various reptiles, pointing to another factor being involved. In mammals, *sox9* activates expression of the *fgf9* and genes encoding prostaglandin D synthase (*pdgs*) [7, 8, 17].

In the mammalian gonad, a key role for *sox9* during testicular development is the activation of Amh (anti-Müllerian hormone) (Figure 2, [17]).

In the chicken embryo, *amh* precedes expression of *sox9*, and the gene is expressed in both males and females. Hence, *sox9* does not activate *amh* in the avian system although it might upregulate it. The activator of expression of *amh* in the chicken and other vertebrates is unclear, but is likely to involve the orphan nuclear receptor SF1 which is expressed in male gonads (ZZ) compared to females (ZW) suggesting that a dosed Z-linked gene such as *dmrt1* may

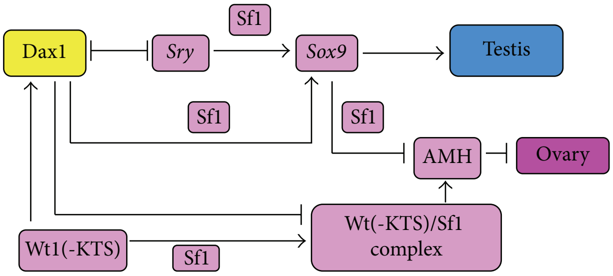


Figure 2: The role AMH in repressing of Müllerian duck differentiation. The Wolffian duck has to be maintained and stimulated to differentiate into the male tract and accessory organ. Then, the Müllerian duct system has to regress, due to action AMH secreted by Sertoli cells. *Sox9* and *Sf1* are both involved in the expression of the *AMH* gene as a result of their respective binding to the promoter and in part because of their ability to interact with each other (adopted from [17]).

activate its expression in avians. *Amh* may participate in testis determination by blocking estrogen synthesis, namely by repressing expression of the aromatase encoding gene *cyp19a1* (Figure 3) [4].

The key ovary determinant in mammals has not yet been defined, but the canonical β-catenin signaling pathway is required for ovarian morphogenesis [4]. The β-catenin signaling pathway appears to be conserved in the other vertebrates, including fishes, reptiles with TSD, and chickens. In these cases, R-Spo1, Wnt4, and/or β-catenin show female upregulation [4, 19]. This female pathway appears to be deeply conserved among vertebrates. A major difference between mammal and nonmammalian vertebrates is the requirement of estrogen for ovarian differentiation in the latter [4]. In eutherian mammals, the embryonic gonads are resistant to sex steroid effects although estrogen is required to maintain the postnatal ovary. The position of *foxl2* in the ovarian pathway appears to vary among the major groups. In the chicken embryo, *foxl2* expression from 5.5th embryonic day is one of the earliest known markers of ovarian development. *Foxl2* activates the *cyp19a1* 5′-regulatory region in the tilapia fish and in mammals. A major question relating to ovarian development is how the FOXL2 and R-SPO1/Wnt4 pathways interact to coordinate ovarian development. The two pathways appear to be independent in goat and in chicken embryos; FOXL2 and R-SPO1 localize to different ovarian compartments (medulla and cortex, resp.) (Figure 4) [4].

Accumulation of knowledge on recent variants of sex determination will allow us to distinguish conserved and diversified SD pathways among vertebrates and lead us into a deeper understanding of the vertebrate SD cascade. For example, it may be worth reconsidering the role of AMH signaling in gonadal sex determination even in mammals based on the results of previous fish studies [7, 8]. Until 2011, all four vertebrate master SD genes (or strong candidates) were known to code for transcription factors which could have been construed as evidence that gonadal sex determination in vertebrates is always triggered by transcription factors. However, the three novel candidates for the master SD genes in the Patagonian pejerrey, *Oryzias o. luzonensis*, and fugu code for growth factors or one of their receptors. Thus, these findings suggest alternative mechanisms of genotypic sex determination, in which the main trigger is not constrained to be a transcription factor [8, 20].

A growing body of evidence reveals the importance of epigenetic regulatory mechanisms, such as DNA-methylation, histone modifications, and the role of noncoding RNAs in controlling sex determination and gonadogenesis. Chicken Z-linking MHM region and specific sites of hypermethylated inside chicken *cyp19*, involvement of methylated lysine 9 of histone H3 (H3K9me) and heterochromatin protein 1 (HP1) in epigenetic modifications in the phenomenon of imprinting, and the presence conserved PHF7 protein from insects to mammals that exhibits male-specific expression in the germline can be mentioned as examples [14, 15, 21].

4. Summary

Different groups of vertebrates discovered the changes of all stages of the formation of sex characteristic of mammals: the chromosomal sex predetermination associated with the formation of the XX or XY zygote at fertilization, formation of testes or ovary (sex determination), and the full development of the respective gonads, associated features (sex differentiation). So much for the birds and Lepidoptera described heterogametic females and merging Z and W chromosomes at fertilization. For mammals, birds, amphibians, and fish, are revealed sex determining genes. Only some of them belong to the families of such as *SOX* genes and DM-containing genes, which are transcription factors similarly controlled the sex determination. Other fish genes were described (*Gsdf* in *Oryzias luzonensis*, *amhr2* in *Takifugu*, *amhy* in the Patagonian pejerrey, *sdY* in *Oncorhynchus mykiss*). The genes belong to completely different families encoding growth factors or receptors and they can influence directly on the proliferation of germ cells. It is possible to talk about specific plasticity of sex determination of fish. In amphibians, sex-determining genes virtually replaced steroids ones. The role of steroids in sex determination is clearly weaker in the evolutionary chain of amphibians, reptiles, birds, mammals, and marsupials. Around this scheme, the transition is implemented from pure TSD mechanism to pure GSD one.

A general feature of vertebrate gonadal sex-determining pathways is that master switches appear to have been added at the top of the hierarchy, with more conserved core genes appearing downstream. Genes *amh* and *sox9* in males are conserved across groups, but the upstream regulator differs (*sry* in mammals and DM domain genes in nonmammals). With respect to the ovarian development, the R-SPO1/Wnt4 and FOXL2 pathways are conserved from fishes to birds and mammals. However, the genetic networks are in fact more complex than this simple scenario. It would appear that the testis pathway is more changeable than the ovarian pathway although this cannot be confirmed in the absence of master ovary factors. For example, in male embryos, the Sertoli cell progenitors in the chicken apparently have

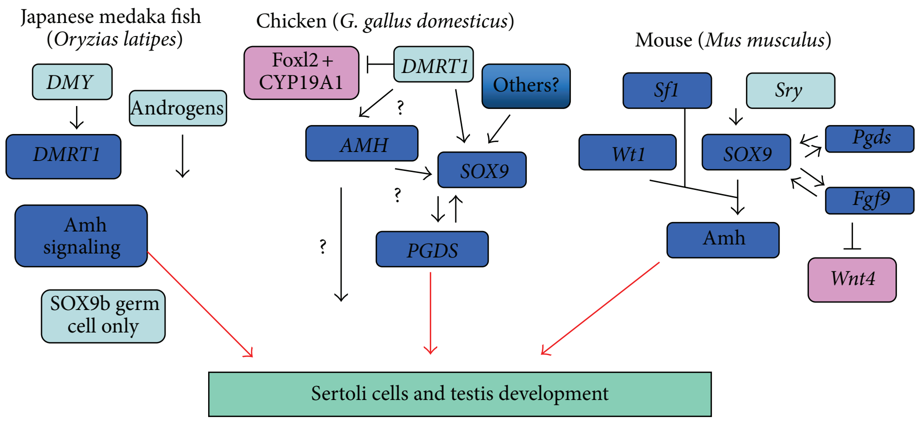


FIGURE 3: Genetic determination of testes in Japanese medaka fish, chicken, and mouse. Somatic gene *sox9* was involved in the control of testes from fish. AMH signaling pathway is a very conservative element of testes, along with other major regulators, appeared in other species (adopted from [4]).

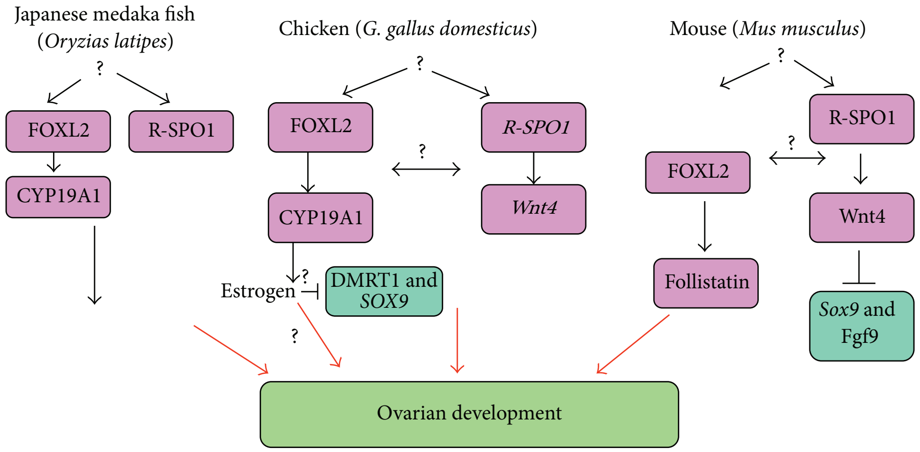


FIGURE 4: Genetic determination of the ovaries of Japanese medaka fish, chicken, and mouse. Genes such as *foxl2* and *r-spo1* are conserved. *Cyp19a1* plays a more important role in nonmammals (adopted from [4]).

a different developmental origin from those of the mouse. Differences between species must logically involve differences in gene expression, and this is reflected in the past—the embryonic gonads of vertebrates have been considered highly conserved in structure. While this is generally true, it is clear that the molecular pathways underlying this commonality of structure are actually quite plastic. Functional analyses are now possible in most groups, such as fishes and chicken, in addition to well-established strategies in mouse. Another area worthy of investigation is exactly how novel genes are incorporated to the top of the pathway and determining the evolutionary pressures that cause the relative shuffling of genes in the male and female pathways. These efforts will broaden our understanding of vertebrate sex determination and how it has evolved [4, 8, 16].

Abbreviations

GSD:	Genetic sex determination
ESD:	External sex determination
TSD:	Temperature sex determination
X, Y, Z, W:	Sex chromosomes
SD-gene:	Sex determination gene
ER:	Estrogen receptor
AR:	Androgen receptor
AMH:	Anti-Müllerian hormone
T:	Testosterone
E_2:	Estradiol
MYR:	Million years ago
F:	Female
M:	Male.

References

[1] T. Gamble and D. Zarkower, "Sex determination," *Current Biology*, vol. 22, no. 8, pp. R257–R262, 2012.

[2] A. P. Arnold, "The end of gonad-centric sex determination in mammals," *Trends in Genetics*, vol. 28, no. 2, pp. 55–61, 2011.

[3] A. P. Arnold, X. Chen, J. C. Link, Y. Itoh, and K. Reue, "Cell-autonomous sex determination outside of the gonad," *Developmental Dynamics*, vol. 242, no. 4, pp. 371–379, 2013.

[4] A. Cutting, J. Chue, and C. A. Smith, "Just how conserved is vertebrate sex determination?" *Developmental Dynamics*, vol. 242, pp. 380–387, 2013.

[5] H. Merchant-Larios and V. Diaz-Hernandez, "Environmental sex determination mechanisms in reptiles," *Sexual Development*, vol. 7, pp. 95–103, 2013.

[6] N. Valenzuela, J. L. Neuwald, and R. Literman, "Transcriptional evolution underlying vertebrate sexual development," *Developmental Dynamics*, vol. 242, pp. 307–319, 2013.

[7] Y. Kobayashi, Y. Nagahama, and M. Nakamura, "Diversity and plasticity of sex determination and differentiation in fishes," *Sexual Development*, vol. 7, pp. 115–125, 2013.

[8] K. Kikuchi and S. Hamaguchi, "Novel sex-determining genes in fish and sex chromosome evolution," *Developmental Dynamics*, vol. 242, pp. 339–353, 2013.

[9] N. F. Parnell and J. T. Streelman, "Genetic interactions controlling sex and color establish the potential for sexual conflict in Lake Malawi cichlid fishes," *Heredity*, vol. 110, no. 3, pp. 239–246, 2013.

[10] M. Nakamura, "The mechanism of sex determination in vertebrates—are sex steroids the key-factor?" *Journal of Experimental Zoology*, vol. 313, no. 7, pp. 381–398, 2010.

[11] M. I. Nakamura, "Is a sex-determining gene(s) necessary for sex-determination in amphibians? Steroid hormones may be key factor," *Sexual Development*, vol. 7, pp. 104–114, 2013.

[12] A. J. Pask, "A role for estrogen in somatic cell fate of the mammalian gonad," *Chromosome Research*, vol. 20, no. 1, pp. 239–245, 2012.

[13] A. C. Ditewig and H. H. Yao, "Organogenesis of the ovary: a comparative review on vertebrate ovary formation," *Organogenesis*, vol. 2, no. 2, pp. 36–41, 2005.

[14] X. Yang, J. Zheng, L. Qu et al., "Methylation status of cMHM and expression of sex-specific genes in adult sex-reversed female chickens," *Sexual Development*, vol. 5, no. 3, pp. 147–154, 2011.

[15] H. L. Ellis, K. Shioda, N. F. Rosenthal, K. R. Coser, and T. Shioda, "Masculine epigenetic sex marks of the CYP19A1/Aromatase promoter in genetically male chicken embryonic gonads are resistant to estrogen-induced phenotypic sex conversion," *Biology of Reproduction*, vol. 87, no. 1, article 23, pp. 1–12, 2012.

[16] L. Fang, R. Xin, Y. Che, and S. Xu, "Expression of sex-related genes in chicken embryos during male-to-female sex reversal exposure to diethylstilbestrol," *Integrative Agriculture*, vol. 12, no. 1, pp. 127–135, 2013.

[17] T. Jiang, C. C. Hou, Z. Y. She, and W. X. Yang, "The SOX gene family: function and regulation in testis determination and male fertility maintenance," *Molecular Biology Reports*, vol. 40, pp. 2187–2194, 2013.

[18] C. A. Smith, K. N. Roeszler, T. Ohnesorg et al., "The avian Z-linked gene DMRT1 is required for male sex determination in the chicken," *Nature*, vol. 461, no. 7261, pp. 267–271, 2009.

[19] K. L. Ayers, A. H. Sinclair, and C. A. Smith, "The molecular genetics of ovarian differentiation in the avian model," *Sexual Development*, vol. 7, no. 1–3, pp. 80–94, 2013.

[20] J. A. Graves, "How to evolve new vertebrate sex determining genes," *Developmental Dynamics*, vol. 242, pp. 354–359, 2013.

[21] F. Piferrer, "Epigenetics of sex determination and gonadogenesis," *Developmental Dynamics*, vol. 242, pp. 360–370, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 472163, 10 pages
http://dx.doi.org/10.1155/2013/472163

Research Article

Algorithms of Ancestral Gene Length Reconstruction

Alexander Bolshoy[1,2] **and Valery M. Kirzhner**[3]

[1] *Department of Evolutionary and Environmental Biology, Institute of Evolution, University of Haifa, 199 Aba-Hushi Avenue, Mount Carmel, Haifa 3498838, Israel*
[2] *Institute of Evolution, University of Haifa, Mount Carmel., Haifa 39105, Israel*
[3] *The Tauber Bioinformatics Center, University of Haifa, 199 Aba-Hushi Avenue, Mount Carmel, Haifa 3498838, Israel*

Correspondence should be addressed to Alexander Bolshoy; bolshoy@research.haifa.ac.il

Received 30 August 2013; Accepted 24 September 2013

Academic Editor: Vassily Lyubetsky

Ancestral sequence reconstruction is a well-known problem in molecular evolution. The problem presented in this study is inspired by sequence reconstruction, but instead of leaf-associated sequences we consider only their lengths. We call this problem ancestral gene length reconstruction. It is a problem of finding an optimal labeling which minimizes the total length's sum of the edges, where both a tree and nonnegative integers associated with corresponding leaves of the tree are the input. In this paper we give a linear algorithm to solve the problem on binary trees for the Manhattan cost function $s(v, w) = |\pi(v) - \pi(w)|$.

1. Introduction

Ancestral sequence reconstruction (ASR) is a well-recognized problem in molecular evolution [1]. Let **G** be a (phylogenetic) tree with **n** leaf nodes, and k strings over one alphabet (gene sequences) assigned to k leaves ($k \leq n$). ASR may be defined in the following way: assignment of strings to inner nodes "in the best possible way." There are two main paradigms for ASR: maximum parsimony (MP) and probabilistic-based reconstruction. The latter includes maximum likelihood (ML) and Bayesian reconstructions. MP reconstruction has a time complexity linear in the number of sequences analyzed. The problem of the parsimonious reconstruction of ancestral states for the given tree with the given states of its leaves (the most parsimonious assignment of the labels of internal nodes for a fixed tree topology) is a well-studied problem [2–4]. Efficient algorithms have also been developed for different types of ML-based reconstructions (reviewed in [5]). ASR methods require as input both a phylogenetic tree and a set of gene sequences associated with corresponding leaves of the tree [6].

ASR is related to gene sequence evolution while the problem presented in this paper, being inspired by ASR, deals with gene length variation. Instead of considering leaf-associated sequences we take into account only their lengths. Instead of the reconstruction of ancestral sequences, we search for the optimal reconstruction of ancestral gene lengths. The problem may be called ancestral gene length reconstruction (AGLR). AGLR is actually a problem of finding an optimal labeling which minimizes the total "length" sum of the edges, the minimum sum problem where both a tree and nonnegative integers associated with corresponding leaves of the tree are the input.

In the graph theory vertex labeling related problems were intensively studied [8]. Typically, the problems can be described as follows: for a given graph, find the optimal way of labeling the vertices with *distinct* integers. The problems and their solutions were described in [9–12]. In [13] we presented the algorithms to solve the minimum sum problem where both a tree and *positive* integers associated with *all leaves* of the tree are the input (finding the optimal way of labeling the vertices with *positive* integers). Here we would like to formulate the minimum sum problem where both a tree and *positive* integers associated with *some of the leaves* of the tree are the input (finding the optimal way of labeling the vertices with *nonnegative* integers). This problem reflects a situation in which the genome tree is constructed by one or another method for a set of genomes, the leaves of the tree are linked with the corresponding genomes of the set, and the leaves are labeled by integers designating lengths of genes of a chosen gene family. Some leaves would be labeled *zero* because corresponding genomes have no genes of the chosen gene family. Alternatively, it may be a case of a missing value but in this study we do not consider this case: in the problem definition that we bring here zero means "no value."

In this paper we provide a linear algorithm to solve max sum problem on binary trees for the Manhattan cost function $s(v, w) = |\pi(v) - \pi(w)|$. The algorithm uses dynamic programming technique and the properties of the Manhattan distance.

2. Preliminaries

Let **G** be a tree with **n** leaf nodes, vertex set **V**(**G**), and edge set **E**(**G**). $N = |V(G)|$. Let us number the leaf nodes of G:$1, 2, \ldots, n$. Let us number the root of $G : N$. An *integer labeling* $\boldsymbol{\pi}$ of G is a mapping π from G to a set of nonnegative integers, where label **0** is an out-of-the ordinary label meaning "absent value." Let us denote integer labeling of the leaf nodes of $G(\pi(1) = p_1, \ldots, \pi(n) = p_n)$. Let us denote by $g_{\min}$ and $g_{\max}$ minimum and maximum *positive* integers labeling leaf nodes: $g_{\min} = \min p_i : p_i > 0$; $g_{\max} = \max p_i$; $m = g_{\max} - g_{\min} + 1$.

Let us introduce a cost function φ of the edge $vw \in E(G)$:

$$\varphi(x, y) = \begin{cases} 0 & \text{if } x = y \text{ else} \\ C_1 & \text{if } x = 0 \text{ else} \\ C_2 & \text{if } y = 0 \text{ else} \\ \theta(x, y), \end{cases} \tag{1}$$

where the nonnegative cost function $\theta(x, y)$ has the following distance properties:

(i)

$$\begin{gathered} \theta(x, y) \geq 0 \quad x = y \longleftrightarrow \theta(x, y) = 0 / * \\ \text{function is equal to zero if} \\ \text{and only if its arguments are equal } * / \end{gathered} \tag{2}$$

(ii)

$$\theta(x, y) = \theta(y, x) / * \text{ symmetry } * / \tag{3}$$

(iii)

$$\begin{aligned} x > y \longrightarrow [&(\theta(x, y) < \theta(x, y - 1)), \\ &(\theta(x, y) < \theta(x + 1, y))]; \end{aligned} \tag{4}$$

(iv)

$$\begin{aligned} x < y \longrightarrow [&(\theta(x, y) < \theta(x - 1, y)), \\ &(\theta(x, y) < \theta(x, y + 1))]. \end{aligned} \tag{5}$$

$C_1 > C_2 > m = g_{\max} - g_{\min} + 1$. C_1 is a gain penalty, C_2 is a loss penalty, and θ is a length change penalty function. Since the likelihoods of loss and gain events are likely to differ, we may need to weight them differently. This is achieved by introducing different penalties $C_1 > C_2$; the loss penalty is normally assigned a value close to $g_{\max} - g_{\min}$, whereas the gain penalty should be larger due to biological considerations. They suggest that, on average, gene loss might be a more likely event than gene gain. Therefore, different gain penalties were used in our study similarly to as it was done in [14].

An example of a function $\theta(x, y)$ is $|\pi(v) - \pi(w)|^{\lambda}$. In case of $\lambda = 1$ we obtain an absolute value of the difference between labelings v and w: $|\pi(v) - \pi(w)|$. In case of $\lambda = 2$ we obtain a square of the difference between labelings v and w: $(\pi(v) - \pi(w))^2$.

2.1. An Arbitrary Tree and an Arbitrary Cost Function. Given a tree G, an integer labeling of the leaves of $G(p_1, \ldots, p_n) = 1$, the gain penalty C_1, the loss penalty C_2, and a cost function θ ((1)–(5)), the minimum sum problem is to find a labeling which minimizes the total cost:

$$S(G) = \min_{\pi} \sum_{\forall \{vw\} \in E(G)} \varphi(\pi(v), \pi(w)) \quad \text{over all } \pi. \tag{6}$$

2.2. A Binary Tree Problem . Given a binary tree G, an integer labeling of the leaves of $G(p_1, \ldots, p_n)$, the "gain" penalty C_1, and the "loss" penalty C_2, the *Manhattan* minimum sum problem is to find the labelings which minimize the sum S over all π

$$\begin{aligned} S(G) = &\sum_{\forall \{vw\} \in E(G) \& \pi(v) \neq 0 \& \pi(w) \neq 0} |\pi(v) - \pi(w)| \\ &+ k_1 \cdot C_1 + k_2 \cdot C_2, \end{aligned} \tag{7}$$

where k_1 is a number of edges of type $(\pi(v) = 0 \& \pi(w) > 0)$, and k_2 is a number of edges of type $(\pi(v) > 0 \& \pi(w) = 0)$.

3. Problem Solutions

3.1. DP Algorithm (for the Problem (1)*).* Due to the properties ((2)–(5)) of the cost function $\theta(x, y)$ all labels of the optimal labeling must be either equal to 0 or in the interval $[g_{\min}, g_{\max}]$. As a consequence of this, the dynamic programming (DP) method is applicable for the problem. It will be easier to explain the DP method on a binary tree using $\sigma_k(i)$ notation. The quantity $\sigma_k(i)$ will be interpreted as the minimal cost, given that node k is assigned integer i, to the subtree with the node k as a root of the subtree.

3.1.1. DP Algorithm for a Binary Tree

Up Phase. A procedure called DP_up calculates the costs $\sigma_k(i)$ of all nodes $V(G)$ of the tree G, given a cost function φ.

When we compute $\sigma_N(i)$ for the root node (the index of the root is N), then we simply choose the minimum of these values:

$$S(G) = \min_i \sigma_N(i) \tag{8}$$

Initiation. Given labeling of the leaf nodes of $G(p_1, \ldots, p_n\}$ at the tips of the tree the $\sigma_i(j)$ are easy to compute. The cost is 0 if the observed integer p_i is integer j, and infinite otherwise.

$$\sigma_i(j) = \left\{ \frac{0 \ \text{ if } j = p_i}{\infty \ \text{ otherwise}} \right\}. \tag{9}$$

Iteration. For the immediate common ancestor of the nodes l and r, node a, we have

$$\begin{aligned} \sigma_a(0) &= \min\left(\sigma_l(0), C_1 + \min_j \sigma_l(j)\right) \\ &\quad + \min\left(\sigma_r(0), C_1 + \min_k \sigma_r(k)\right), \\ \sigma_a(i) &= \min\left(\min_j \left[\theta(i,j) + \sigma_l(j)\right], C_2 + \sigma_l(0)\right) \\ &\quad + \min\left(\min_k \left[\theta(i,k) + \sigma_r(k)\right], C_2 + \sigma_r(0)\right), \\ &\qquad \forall i, j, k \in \left[g_{\min}, g_{\max}\right]. \end{aligned} \tag{10}$$

The interpretation of this equation is immediate. The smallest possible cost given that node a is assigned zero is either the cost $\sigma_l(0)$ or the "gain" penalty C_1 plus the minimum of $\sigma_l(j)$, the least of the two plus the minima of corresponding values associated with the right descendant tree. The smallest possible cost given that node a is assigned i is a sum of two values: the first one is either the cost $\theta(x,y)$ of the edge from node a to node l, plus the cost $\sigma_l(j)$ of the left descendant subtree given that node l is in state j, or the "loss" penalty C_2 plus $S_l(0)$; the second one is the cost $\theta(i,k)$ of the edge from the node a to the node r, plus the cost $\sigma_r(k)$ of the right descendant subtree given that node r is in state k. We select those values of j and k which minimize that sum. Equation (10) is applied successively to each inner node in the tree, doing a postorder tree traversal. Finally it computes all the $\sigma_N(i)$, and then (8) is used to find the minimum cost for the whole tree. The complexity of the Up_phase of the algorithm is $O(N^* m^* m)$.

Traceback. The procedure calculates the labels $\pi(p)$ of all nodes p of the tree G.

Choose any integer i which provides the minimum of the $\sigma_N(i)$—it is the root label. It may be either zero or a positive i. Doing a preorder tree traversal, successively label each inner node in the tree: for any inner node p, and given that a parent label i was reconstructed, the label $\pi(p) = j$ is easily reconstructed as well.

3.1.2. DP Algorithm for an Arbitrary Tree

Up-Phase. A procedure DP_up calculates the costs $\sigma_k(i)$ of all nodes $V(G)$ of the tree.

Suppose that the k_a descendant nodes of the node a are called b_j. The following equation will therefore be similar to (10) replacing the sum of σ_l and σ_r by the total sum of σ_{j_1}, while j_1 traverses all values of b_j:

$$\sigma_a(0) = \sum_{j_1}^{k_a} \min\left[\sigma_{j_1}(0), C_1 + \min_j \sigma_{j_1}(j)\right], \tag{11}$$

$$\begin{aligned} \sigma_a(i) &= \sum_{j_1}^{k_a} \min\left[\min_j\left(\theta(i,j) + \sigma_{j_1}(j)\right), C_2 + \sigma_{j_1}(0)\right], \\ &\qquad \forall i, j \in \left[g_{\min}, g_{\max}\right]. \end{aligned} \tag{12}$$

This equation is applied successively to each node in the tree, doing a postorder tree traversal. Finally it computes all the $\sigma_N(i)$, and then (8) is used to find the minimum cost for the whole tree.

Down Phase. As Traceback above: Consider the following.

3.2. DP Algorithm for a Manhattan Sum for a Binary Tree (Problem (2)).

Manhattan distance $\theta(\pi(v), \pi(w))$ is an absolute value of the difference between labelings v and $w : |\pi(v) - \pi(w)|$. This distance measure has the following property: if siblings have positive labels, then all integers that lie between these values may equally serve as optimal labels of a parent.

(i) If $(\pi(l) \le \pi(r))$, then for all $k\pi(l) \le k \le \pi(r)$ the score $\theta(k,\pi(l)) + \theta(k,\pi(r)) = k - \pi(l) + \pi(r) - k = \pi(r) - \pi(l)$.

(ii) If $(\pi(l) \le \pi(r))$, then for all $k < \pi(l) \le \pi(r)$ the score $\theta(k,\pi(l)) + \theta(k,\pi(r)) = \pi(l) - k + \pi(r) - k = \pi(r) - \pi(l) + 2(\pi(l) - k)$.

(iii) If $(\pi(l) \le \pi(r))$, then for all $\pi(l) \le \pi(r) < k$ the score $\theta(k,\pi(l)) + \theta(k,\pi(r)) = k - \pi(l) + k - \pi(r) = \pi(r) - \pi(l) + 2(k - \pi(r))$.

So, as it would be proven below, at the bottom-up stage of the DP algorithm it would be sufficient to assign to each node a in the tree G four values: left(a), right(a), $Z(a)$, and $X(a)$. The meanings of the values are as follows: left and right are bounds of an interval associated with the node a, Z is a cost value $\sigma_a(0)$, and X is a cost $\sigma_a(i)$ for any integer i from the interval: left $\le i \le$ right.

Initiation. Given labeling of the leaf nodes of $G(p_1, \ldots, p_n) = 1$ these four values are easy to compute for the leaf nodes:

> for $(i = 1;\ i \le n;\ i++)$ if $(p[i] == 0)$ $\{Z[i] = 0$; left$[i] = 0$; right$[i] = 0$; $X[i] = C_1 + C_2\}$ else $\{Z[i] = C_1 + C_2$; left$[i] = p[i]$; right$[i] = p[i]$; $X[i] = 0\}$.

3.2.1. Examples.

Let us consider the simplest trees with two, three, and four labeled leaves. The simplest tree configuration is presented in Figure 1. There is only one node to label—the root node.

(i) Figure 1(a): no genes are assigned to the leaves → no gene is assigned to the root.

(ii) Figure 1(b): the left leaf has no gene, and the right leaf has a gene with the length equal to 136 → the root

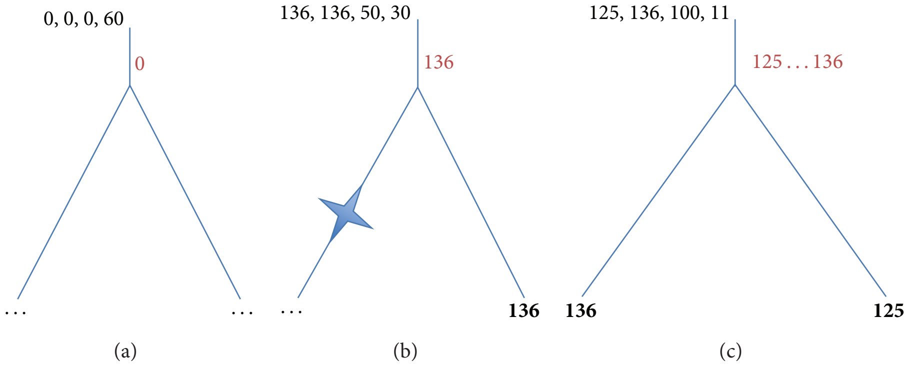


FIGURE 1: Assignment of bottom-up stage values (left, right, Z, and X) in 2-leaf trees. The "gain" penalty $C_1 = 50$; the "loss" penalty $C_2 = 30$. Optimal labels are in red.

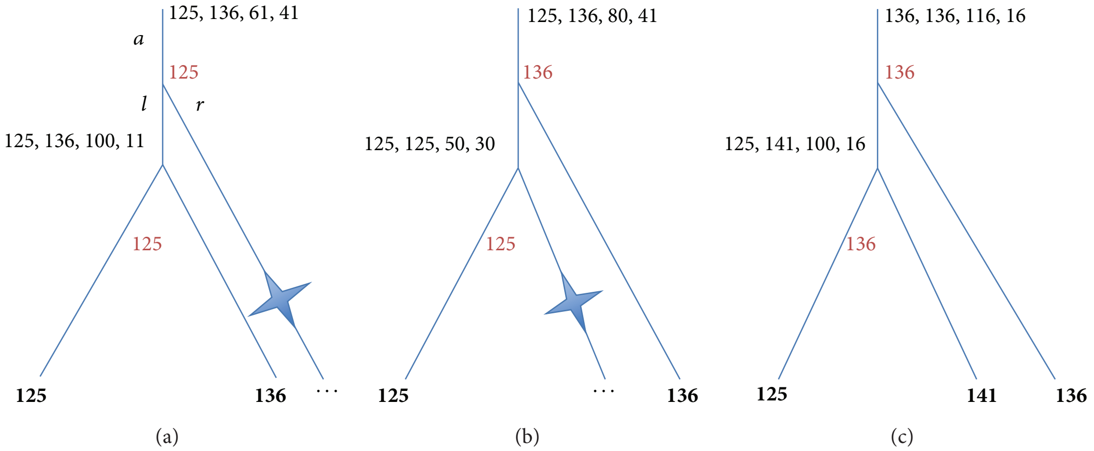


FIGURE 2: Assignment of bottom-up stage values (left, right, Z, and X) in 3-leaf trees. The "gain" penalty $C_1 = 50$; the "loss" penalty $C_2 = 30$. Optimal labels are in red.

is labeled by 136; the score is equal to the loss penalty $C_2 = 30$.

(iii) Figure 1(c): any label $125 \le k \le 136$ is good to label the root; the score is equal to $136 - 125 = 11$.

The next simplest tree topology—three-leaf trees—is presented in Figure 2. There are two nodes to label, the inner node and the root.

(i) Figure 2(a): the inner node is labeled analogically to the root in Figure 1(c): any k $125 \le k \le 136$ is equally good to label the inner node; the root node is labeled analogically to the root in Figure 1(b): $(Z(\text{root}) = C_1 + (136 - 125)) > (X(\text{root}) = C_2 + 11) \rightarrow$ the root is labeled by any k $125 \le k \le 136$, that is, by 125.

(ii) Figure 2(b): labeling is similar to that of Figure 1(a).

(iii) Figure 2(c): the inner node is labeled analogically to Figure 2(a): any label $125 \le k \le 136$ is good to label it; the score is equal to $136-125 = 11$. The root should be labeled by 136 because $125 < 136 < 141$.

Determination of the optimal labeling of the four-leaf trees is very similar to the examples described above. Figure 3 illustrates labeling of the tree where all four leaves have nonzero labels: ((125, 141), (136, 150)). Labeling of the inner nodes is as above (Figure 2(c)): [125, 141] and [136, 150]. All integers of the intersection between these two close intervals are optimal values to label the root: $[125, 141] \cap [136, 150] = [136, 141]$. In Figure 3 we present the value 136 as a chosen suitable label.

Examples of the trees with very distinct subtrees are presented in Figures 4 and 5. In Figure 4 we present a tree obtained by merging two very different subtrees. The left 4-leaf subtree has very obvious intuitive labeling of internal nodes: all nodes should be labeled by zero. The right subtree is identical to the tree presented in Figure 2(c). Merging of these two subtrees produces bottom-up stage values (left, right, Z, and X) to the new root equal to [125, 136, 111, 91]. In spite of assignins the interval [125, 136] to the root only the value 136 provides the optimal solution. (We would like to express our gratitude to the anonymous reviewer for bringing our attention to this situation.) We formulate this rule below describing traceback stage of the algorithm. Figure 4 is chosen to illustrate labeling of nodes similar to the root of the tree.

After considering these few simple examples, we describe the algorithm.

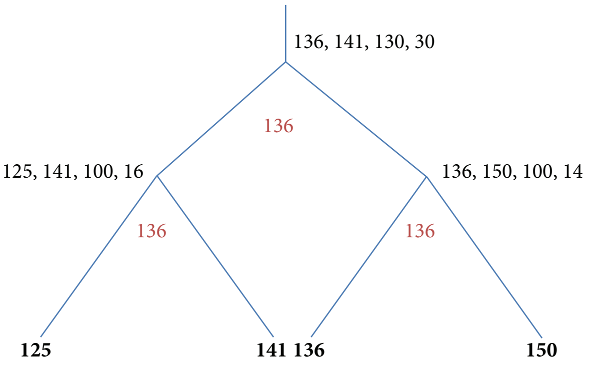


Figure 3: Assignment of bottom-up stage values (left, right, Z, and X) in a 4-leaf tree with all four leaves labeled by positive integers. The "gain" penalty $C_1 = 50$; the "loss" penalty $C_2 = 30$. Optimal labels are in red.

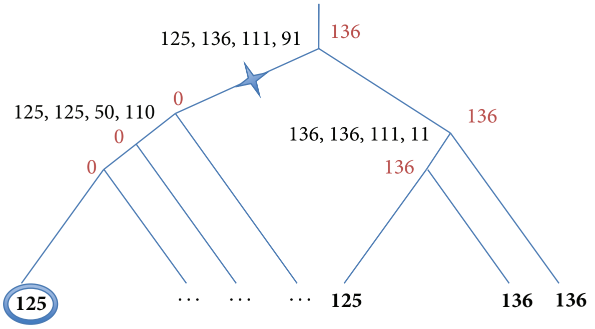


Figure 4: Labeling of a "peculiar" tree. The left subtree has three zero and one nonzero leaf, while the right subtree has three nonzero leaves. The "gain" penalty $C_1 = 50$; the "loss" penalty $C_2 = 30$. Optimal labels are in red.

3.2.2. Bottom-Up Stage

Initiation. Given labeling of the leaf nodes of $G(p_1, \ldots, p_n)$ at the tips of the tree the $\sigma_i(j)$ are easy to compute. The cost is 0 if the observed integer p_i is integer j, and

$$C_1 + C_2 \quad \text{otherwise.} \tag{13}$$

Iteration. Doing a postorder tree traversal assign successively to each node in the tree the abovementioned four values left(a), right(a), $Z(a)$, and $X(a)$. An interval [left(a), right(a)] is assigned according to the following rule: if anyone of two children intervals is not defined, then assign the interval of the other child; otherwise, a parent interval is either an intersection of the intervals of its children or an interval that lies between these intervals if their intersection is empty. Z is a cost value $\sigma_a(0)$, where for the Manhattan distance we can rewrite (10) as

$$\begin{aligned} Z(a) = \sigma_a(0) = {} & \min\left(\sigma_l(0), C_1 + \sigma_l(j)\right) \\ & + \min\left(\sigma_r(0), C_1 + \sigma_r(j)\right), \end{aligned}$$

$$\begin{aligned} X(a) = \sigma_a(i) = {} & \min\left(\min_j\left[\theta(i,j) + \sigma_l(j)\right], C_2 + \sigma_l(0)\right) \\ & + \min\left(\min_k\left[\theta(i,k) + \sigma_r(k)\right], C_2 + \sigma_r(0)\right) \\ = {} & \min\left(\min_j\left[|i-j| + \sigma_l(j)\right], C_2 + \sigma_l(0)\right) \\ & + \min\left(\min_j\left[|i-j| + \sigma_r(j)\right], C_2 + \sigma_r(0)\right). \end{aligned} \tag{14}$$

3.2.3. Pseudocode. For more details see Pseudocode 1.

3.2.4. Traceback Stage

Interval Correction Rule. Following the bottom-up stage four values left(a), right(a), $Z(a)$, and $X(a)$ are assigned to every internal node a of the tree. An interval (left(a), right(a)) should be diminished if one of the edges connecting the node a with its son becomes of type $(k, 0)$, $k > 0$. Let us denote sons of the node a by $l(a)$ and $r(a)$. Correction condition $\Omega(a)$ would be formulated as

$$\begin{aligned} \Omega(a) = (X(a) \leq Z(a)) \quad & \text{and} \\ & ((X(l(a)) > Z(l(a)))\, V\, (X(r(a)) > Z(r(a)))). \end{aligned} \tag{15}$$

If $\Omega(a)$ is TRUE, then the bounds of the corrected interval would be obtained by intersection of the interval associated with the node with the corrected interval associated with the corresponding son:

$$\begin{aligned} \text{if } X(l(a)) > Z(l(a)), \text{then } & \left[\text{left}'(a), \text{right}'(a)\right] \\ & = [\text{left}(a), \text{right}(a)] \\ & \cap [\text{left}(r(a)), \text{right}(r(a))] \\ & \text{else } \left[\text{left}'(a), \text{right}'(a)\right] \\ & = [\text{left}(a), \text{right}(a)] \\ & \cap [\text{left}(l(a)), \text{right}(l(a))]; \end{aligned} \tag{16}$$

Otherwise, the bounds of the corrected interval would not be changed from the original ones:

$$\left[\text{left}'(a), \text{right}'(a)\right] = [\text{left}(a), \text{right}(a)]. \tag{17}$$

Initiation. Labeling of the root: if $X(N) \leq Z(N)$, then correct the root interval according to (15)–(17), and then choose an integer from the corrected interval assigned to the root node otherwise choose 0—it is the root label $\pi(N)$.

Iteration. Doing a preorder tree traversal, successively label each node in the tree either by an integer from the corrected interval assigned to this node which is the nearest to its parent

```
/* Assignment values left[$a$] and right [$a$] */
FL = FALSE;
if (left[$l$] == 0) { left[$a$] = left[$r$]; right[$a$] = right[$r$]} else
if (left[$r$] == 0) { left[$a$] = left[$l$]; right[$a$] = right[$l$]} else
if (left[$l$] $\le$ left[$r$]) {
  if (right[$l$] < left[$r$]) {
        left[$a$] = right[$l$]; right[$a$] = left[$r$]; FL = TRUE;
  } else {
        left[$a$] = left[$r$]; right[$a$] = min(right[$l$], right[$r$]);
  }
} else {
  if (right[$r$] < left[$l$]) {
        left[$a$] = right[$r$]; right[$a$] = left[$l$]; FL = TRUE;
  } else {
        left[$a$] = left[$l$]; right[$a$] = min(right[$l$], right[$r$]);
  }
}
/* Assignment values $Z[a]$ and $X[a]$: $Z$ is a cost of $\sigma_a$ (0), $X$ is a cost of $\sigma_a$ ($i$) for all $i$:
left $\le i \le$ right */
if (FL) diff = right[$a$] − left[$a$]; else diff = 0;
$Z(a) = \min(Z(l), C_1+X(l)) + \min(Z(r),C_1+X(r))$
$X(a) = \min(\text{diff} + X(l) + X(r), X(r) + C_2+ Z(l)), X(l) + C_2+ Z(r), 2C_2 + Z(l) + Z(r))$
```

Pseudocode 1

Table 1: List of archaeal genomes for Figure 4.

No.	Name	Kingdom	Group
0	*Aeropyrum pernix* K1	A	C
1	*Archaeoglobus fulgidus* DSM 4304	A	E
8	*Caldivirga maquilingensis* IC-167	A	C
29	*Haloarcula marismortui* ATCC 43049	A	E
30	*Halobacterium salinarum* R1	A	E
31	*Halobacterium* sp. NRC-1	A	E
32	*Haloquadratum walsbyi* DSM 16790	A	E
35	*Hyperthermus butylicus* DSM 5456	A	C
36	*Ignicoccus hospitalis* KIN4/I	A	C
37	*Metallosphaera sedula* DSM 5348	A	C
38	*Methanobrevibacter smithii* ATCC 35061	A	E
39	*Methanococcoides burtonii* DSM 6242	A	E
40	*Methanococcus aeolicus* Nankai-3	A	E
41	*Methanococcus maripaludis* C5	A	E
42	*Methanococcus maripaludis* C6	A	E
43	*Methanococcus maripaludis* C7	A	E
44	*Methanococcus maripaludis* S2	A	E
45	*Methanosaeta thermophila* PT	A	E
46	*Methanosarcina acetivorans* C2A	A	E
47	*Methanosarcina barkeri* str. fusaro	A	E
48	*Methanosarcina mazei* Go1	A	E
49	*Methanosphaera stadtmanae* DSM 3091	A	E
50	*Methanospirillum hungatei* JF-1	A	E

Notations of the groups: E: *Euryarchaeota*, C: *Crenarchaeota*.

label (it may be either the value equal to the parent label or the boundary value of the interval assigned with the node) or by 0.

The proof of the correctness of a simpler algorithm (without zero-labeled leaves) is published in [15]. In the Appendix there are several lemmas, from which the correctness of the algorithm presented here results.

3.2.5. Example. In Figure 5 of [7] the consensus trees obtained from 100 genome trees were presented. The trees were produced on the basis of 80% randomly chosen COGs, and the right tree was produced on the basis of 15%-jackknifing (the explanations in the text of [7]). This tree possesses phylogenetic reasonableness.

(a) The representatives of both prokaryotic Kingdoms: Eubacteria and Archaea are clustered separately. In other words, Archaeal organisms (genomes 0, 1, 8, 29–32, and 35–50) form a monophyletic group.

(b) Euryarchaeota and Crenarchaeota form monophyletic groups.

A part of this tree was selected to illustrate the algorithm. We took the upper part of the tree related exclusively to Archaea (see A/B marked arrow in Figure 4(b) from [7]) and placed the root at the point dividing all Archaeal genomes into Euryarchaeota and Crenarchaeota (see E/C marked arrow in Figure 4(b) from [7]). Thus, Figure 5 is a part of Figure 4(b) from [7] labeled according to COG0835. This COG was randomly selected as suitable for purposes of illustration. Table 1 presents a list of Archaeal genomes from the whole set of genomes that were used for a genome tree construction (Figure 4(b) from [7]). Table 2 presents the lengths of the Archaeal proteins of this COG.

To assign labels to the leaves of the tree of Figure 5 two preprocessing steps were done: (1) taking off outliers, the lengths 328 of the *H. marismortui* protein and 344 of the *M.*

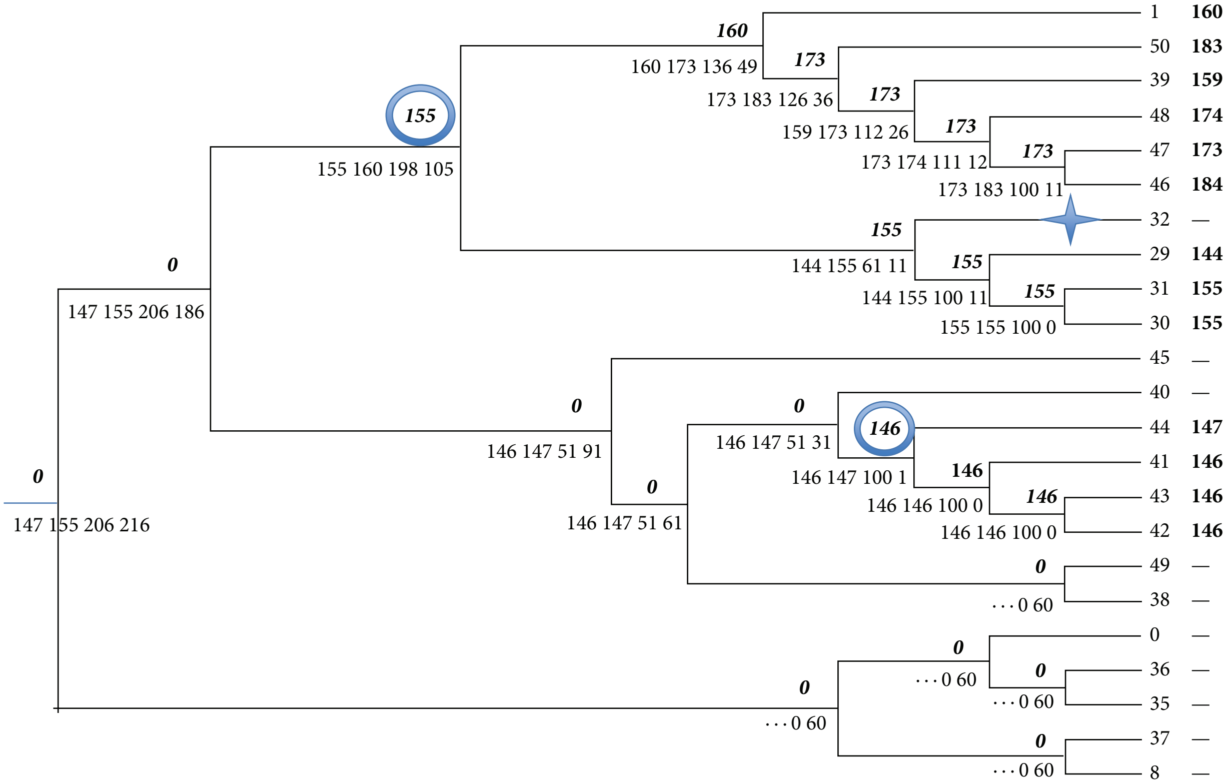


Figure 5: Archaeal part of Figure 4(b) from [7] labeled accordingly to COG0835.

hungatei protein are obvious outliers; (2) taking the median value of paralog's lengths of the genomes 30, 31, 46, 48, and 50. Figure 5 presents results of application of the bottom-up and traceback stages of the algorithm to this tree: a quartet that was assigned to a node a at the bottom upstage is shown under the edge linking the node a and its parent node, a label that was assigned to the node a at the traceback stage, is shown over the same edge.

As we can see the root is labeled by zero. There are two gene-birth events and one gene-death event. One gene was born with the length of 155 and another gene birth is labeled by 146. Genome number 32 (*Haloquadratum walsbyi*) has no protein from COG0835, while other *Haloarchaea* (genomes 29–31) do have. Thus, the edge connecting with leaf labeled by 32 is marked with a gene-loss symbol.

4. Discussion

In [15] the algorithms to find the optimal labeling of the vertices of the tree under Wagner parsimony were presented. A simple extension of the problem could be finding the optimal labeling of the vertices of the tree with nonnegative integers. This more realistic approach requests special consideration of zero labeling. Wedges of type $(k, 0)$, $k > 0$, should be scored differently from wedges of type $(0, k)$, $k > 0$, because the $(k, 0)$ notes gene loss, while $(0, k)$ notes gene gain. These events should be scored differently. Interestingly, this differentiated scoring in addition to tree labeling resulted in reconstruction of "parsimonious" evolutionary scenario. Reconstruction of a gene evolution along a species tree is an interesting and principal problem. Lyubetsky and his coworkers contributed a lot to formulation and solving this problem. In their studies [16–22] the authors tackled mainly two important and sophisticated phylogenetic problems. The obtained results are partially reviewed in the first section of [22] which also provides an extended biological background and relevant references. Reconstruction of a gene evolution along a species tree (to build the evolutionary scenario), following the approach of Lyubetsky et al., is to find an optimal mapping of a gene tree into a species tree. (An example of a different approach was presented in [14].) The second problem is to construct a supertree from the given set of gene trees.

As it was mentioned in [22], the first problem, stated as a tree-into-tree mapping, is solved in polynomial (often linear, and at maximum cubic) time even for the case of time slices and horizontal gene transfers. The algorithms presented in our study are polynomial as well.

Choosing C_1 (a gain penalty), C_2 (a loss penalty), and θ (a label change penalty) is crucial for reconstruction of trustworthy evolutionary scenario. However, it is very difficult task and we cannot claim categorically that choosing "correct" parameters of the model will result in truly reliable reconstruction. We do plan to make a comparison between results obtained by abovementioned methods of Lyubetsky and ours (work in progress).

Table 2: Protein lengths of the chemotaxis signal transduction proteins. Archaeal part of COG0835.

Number	COG	Length	Genome name
1	835	160	*Archaeoglobus fulgidus* DSM 4304
29	835	144	*Haloarcula marismortui* ATCC 43049
29	835	328	*Haloarcula marismortui* ATCC 43049
30	835	132	*Halobacterium salinarum* R1
30	835	178	*Halobacterium salinarum* R1
31	835	132	*Halobacterium* sp. NRC-1
31	835	178	*Halobacterium* sp. NRC-1
39	835	159	*Methanococcoides burtonii* DSM 6242
41	835	146	*Methanococcus maripaludis* C5
42	835	146	*Methanococcus maripaludis* C6
43	835	146	*Methanococcus maripaludis* C7
44	835	147	*Methanococcus maripaludis* S2
46	835	182	*Methanosarcina acetivorans* C2A
46	835	184	*Methanosarcina acetivorans* C2A
47	835	173	*Methanosarcina barkeri* str. fusaro
48	835	159	*Methanosarcina mazei* Go1
48	835	189	*Methanosarcina mazei* Go1
50	835	124	*Methanospirillum hungatei* JF-1
50	835	167	*Methanospirillum hungatei* JF-1
50	835	169	*Methanospirillumhungatei* JF-1
50	835	169	*Methanospirillum hungatei* JF-1
50	835	174	*Methanospirillumhungatei* JF-1
50	835	176	*Methanospirillum hungatei* JF-1
50	835	183	*Methanospirillum hungatei* JF-1
50	835	187	*Methanospirillum hungatei* JF-1
50	835	189	*Methanospirillum hungatei* JF-1
50	835	190	*Methanospirillum hungatei* JF-1
50	835	198	*Methanospirillum hungatei* JF-1
50	835	200	*Methanospirillum hungatei* JF-1
50	835	344	*Methanospirillum hungatei* JF-1
50	835	779	*Methanospirillum hungatei* JF-1

To prepare input for the algorithm, as it was done above for 3.2.5, the original data is to be transformed to the following format: to each (genome, COG) pair one standardized protein length should be assigned (as we described in [7]). For a given COG, each organism is represented by a calculated length—a median length of all paralogous proteins. A natural extension would be to formulate the labeling problem taking into account existence of paralogs.

We may define a *k-tuple* integer labeling Π of G as a mapping Π from G to a set of k-tuples composed of integers $\Pi(v) = \{\pi_1(v), \pi_2(v), \dots, \pi_{k(v)}(v)\}$, where $\pi_i(v) \le \pi_{i+1}(v)$ for all $1 \le i < k(v)$. The simplest extension would be to introduce the case with *identical* sizes of k-tuples composed of *nonnegative* integers. A *uniform* k-tuple integer labeling Π_c of G is characterized by a constant $k(v)$ for all v. The stretch of the edge vw in a $\Pi_c(G)$ is a simple sum $c_{vw} = \sum_{i=1}^{k} \varphi(\pi_i(v), \pi_i(w)) \cdot \varphi(x, y)$ is defined as in (1). Given a uniform k-tuple integer labeling of the leaves of G the minimum sum problem is to find a labeling which minimizes the total sum of the stretches of the edges. Some $\boldsymbol{\pi_i(v) = 0}$. The minimum sum problem is that of minimizing $\mathbf{s(G)} = \sum_{\forall\{vw\}\in E(G)} c_{vw}$ over all Π_c for given k. By some modifications of the algorithms presented in this paper the minimizing k-tuple labeling can be found. This model again is a gain-loss model. More sophisticated extension must provide more realistic definition of distance between two k-tuples composed of positive integers by introducing duplication events.

Appendix

Lemma A.1 (root optimal label). *Suppose that the root node is called rt and suppose that its children are called l and r. The claim is*

(1) *if* $(\pi(l) = \pi(r) = 0)$, *then* $\pi(rt) = 0$ *else*

(2) *if* $(\pi(l) = 0)$, *then* $\pi(rt) = \pi(r)$ *else*

(3) *if* $(\pi(r) = 0)$, *then* $\pi(rt) = \pi(l)$ *else*

(4) *if* $(\pi(l) \le \pi(r))$, *then* $\pi(l) \le \pi(rt) \le \pi(r)$ *else*

(5) $\pi(l) \ge \pi(rt) \ge \pi(r)$.

Proof. If we consider a subtree with the node k as a root then $\sigma_k(i)$ designates the minimal cost, given that node k has a label i:

$$\begin{aligned} S(G) &= \min_i \sigma_N(i) \\ &= \min_i \left[\varphi(\pi(rt), \pi(l)) \right. \\ &\qquad \left. + \varphi(\pi(rt), \pi(r)) + \sigma_l(\pi(l)) + \sigma_r(\pi(r))\right]. \end{aligned} \tag{A.1}$$

Proof of the subclaims (1)–(3) is trivial. Case (4) is

$$\begin{aligned} S(G) &= \min_i \sigma_N(i) \\ &= \min_i \left[|\pi(rt) - \pi(l)| + |\pi(rt) - \pi(r)| \right. \\ &\qquad \left. + \sigma_l(\pi(l)) + \sigma_r(\pi(r))\right]. \end{aligned} \tag{A.2}$$

Let us introduce a new numbering π' by changing only the root label: $\pi'(rt) = \pi(l)$. It is easy to see that $S_{\pi'}(G) < S_{\pi}(G)$. It means that for optimal integer labeling π the following is correct: $\pi(rt) \ge \pi(l)$. Likewise, we prove that for optimal integer labeling $\pi(rt) \le \pi(r)$. Let us denote $k = \pi(rt) : \pi(l) \le k \le \pi(r)$:

$$\begin{aligned} &\forall k\,(\pi(l) \le k \le \pi(r))\, S_{\pi}(G) \\ &\quad = k - \pi(l) + S_{\pi}(l) + \pi(r) - k + S_{\pi}(r), \quad \text{q.e.d.} \end{aligned} \tag{A.3}$$

□

Lemma A.2 (leaf parent optimal interval). *Every node of the optimal integer labeling that all its children are leaf nodes has either a zero label or a label between labels of its children.*

Suppose that the root node is called a and suppose that its children are called l and r. The claim is

(1) *if* $(\pi(l) = \pi(r) = 0)$, *then* $\pi(a) = 0$ *else*

(2) *if* $(\pi(l) = 0)$, *then* $\pi(a) = \pi(r)$ *else*

(3) *if* $(\pi(r) = 0)$, *then* $\pi(a) = \pi(l)$ *else*

(4) *if* $(\pi(l) \leq \pi(r))$, *then* $\pi(l) \leq \pi(a) \leq \pi(r)$ *else*

(5) $\pi(l) \geq \pi(a) \geq \pi(r)$.

Proof. Figure 1 illustrates this lemma. Proof of the subclaims (1)–(3) is trivial. In case of condition (4) let us prove that for optimal integer labeling $\pi(a) \geq \pi(l)$. Suppose $\pi(a) < \pi(l)$. Let us denote $(\pi(l) - \pi(a)) = \delta$; $\pi(r) - \pi(l) = \gamma$. Let us introduce a new numbering π' by changing only the label of the node a: $\pi'(a) = \pi(l)$. It is to show that $S_{\pi'}(G) < S_{\pi}(G)$. Indeed,

$$\begin{aligned}
S_{\pi'}(G) = S_{\pi}(G) &- \left|\pi(k) - \pi(j)\right| \\
&- (p_i - \pi(k)) - (p_{i+1} - \pi(k)) \\
&+ \left|\pi'(k) - \pi(j)\right| + \left(p_i - \pi'(k)\right) + \left(p_{i+1} - \pi'(k)\right) \\
= S_{\pi}(G) &+ \left(\left|p_i - \delta - \pi(j)\right| - \left|p_i - \pi(j)\right|\right) \\
&- (p_i - \pi(k)) - (p_{i+1} - \pi(k)) + (p_i - p_i) \\
&+ (p_{i+1} - p_i) = S_{\pi}(G) \\
&+ \left(\left|p_i - \delta - \pi(j)\right| - \left|p_i - \pi(j)\right|\right) \\
&- \delta - (\delta + \gamma) - \gamma = S_{\pi}(G) - \delta.
\end{aligned} \tag{A.4}$$

Likewise, we prove that for optimal integer labeling $\pi(k) \leq p_{i+1}$. Q.e.d. $p_i \leq \pi(k) \leq p_{i+1}$. □

Lemma A.3 (parent optimal interval—(I)). *An optimal label of a parent either is equal to zero or lies between extreme values of optimal labels of its children. If an optimal integer labeling π provides the labels of two siblings i_1 and i_2 satisfying the conditions $a_1 \leq \pi(i_1) \leq b_1$ & $a_2 \leq \pi(i_2) \leq b_2$, then the label of their parent k satisfies* $\min(a_1, b_1, a_2, b_2) \leq \pi(k) \leq \max(a_1, b_1, a_2, b_2)$. *Proof is as for Lemma A.1.*

Lemma A.4 (parent optimal interval—(II)). *An optimal label of a parent in case of the empty intersection of the optimal intervals of its children lies between these intervals.*

If an optimal integer labeling π provides the labels of two siblings i_1 and i_2 satisfying the conditions $(a_1 \leq \pi(i_1) \leq b_1)$ & $(a_2 \leq \pi(i_2) \leq b_2)$ then if $(b_1 \leq a_2)$ then $b_1 \leq \pi(k) \leq a_2$ else if $(b_2 \leq a_1)$ then $b_2 \leq \pi(k) \leq a_1$.

Proof. (1) $b_1 \leq a_2$. Let us assume $\pi(k) < b_1$; $\pi(k) = b_1 - \alpha$. Then we introduce a new labeling π' by changing labels for three nodes: $\pi'(k) = b_1$; $\pi'(i_1) = b_1$; $\pi'(i_2) = a_2$:

$$\begin{aligned}
&S_{\pi}(G) - S_{\pi'}(G) \\
&\quad = \left(\left|\pi(j) - \pi(k)\right| - \left|\pi(j) - \pi'(k)\right|\right) \\
&\quad\quad + \left(\left|\pi(k) - \pi(i_1)\right| - \left|\pi'(k) - \pi'(i_1)\right|\right) \\
&\quad\quad + \left(\left|\pi(k) - \pi(i_2)\right| - \left|\pi'(k) - \pi'(i_2)\right|\right), \\
&\left(\left|\pi(j) - \pi(k)\right| - \left|\pi(j) - \pi'(k)\right|\right) \\
&\quad = \left|\pi(j) - b_1 + \alpha\right| - \left|\pi(j) - b_1\right| = \alpha, \\
&\left(\left|\pi(k) - \pi(i_1)\right| - \left|\pi'(k) - \pi'(i_1)\right|\right) \\
&\quad = \left((\pi(k) - \pi(i_1)) - \left(\pi'(k) - \pi'(i_1)\right)\right) \\
&\quad = b_1 - \alpha - \pi(i_1) - b_1 + b_1 \\
&\quad = b_1 - \alpha - \pi(i_1), \\
&\left(\left|\pi(k) - \pi(i_2)\right| - \left|\pi'(k) - \pi'(i_2)\right|\right) \\
&\quad = \pi(i_2) - b_1 + \alpha - a_2 + b_1, \\
&S_{\pi}(G) - S_{\pi'}(G) = \alpha + b_1 - \alpha - \pi(i_1) \\
&\quad\quad + \pi(i_2) + \alpha - a_2 = (b_1 - \pi(i_1)) \\
&\quad\quad + \alpha + (\pi(i_2) - a_2) > 0.
\end{aligned} \tag{A.5}$$

From assumption $\pi(k) < b_1$ follows that π is not an optimal labeling. Similarly, we can prove that from assumption $\pi(k) > a_2$ follows that π is not an optimal labeling.

(2) $b_2 \leq a_1$. Similarly to (1) let us assume $\pi(k) < b_2$; $\pi(k) = b_2 - \alpha$. Then we introduce a new labeling π' by changing labels for three nodes: $\pi'(k) = b_2$; $\pi'(i_1) = a_1$; $\pi'(i_2) = b_2$. $S_{\pi}(G) - S_{\pi'}(G) > 0$, so we have a contradiction with the statement that $\pi(G)$ is an optimal labeling. □

Lemma A.5 (parent optimal interval—(III)). *An optimal interval of a parent is either an intersection of the optimal intervals of its children or an interval that lies between these intervals in case that their intersection is empty.*

If an optimal integer labeling π provides the labels of two siblings i_1 and i_2 satisfying the conditions $a_1 \leq \pi(i_1) \leq b_1$ & $a_2 \leq \pi(i_2) \leq b_2$, then the label of their parent k satisfies the following condition:

if $([a_1, b_1] \cap [a_2, b_2]) \neq \emptyset$ *then* $\pi(k) \in [a_1, b_1] \cap [a_2, b_2])$ *else*

if $b_1 < a_2$ *then* $\pi(k) \in [b_1, a_2]$ *else* $\pi(k) \in [b_2, a_1]$.

For example, if $a_1 \leq a_2 \leq b_1 \leq b_2$, then **Lemma A.5** *states that $\pi(k)$ satisfies the following condition $a_2 \leq \pi(k) \leq b_1$. Proof is similar to proof in Lemma A.4.*

References

[1] D. A. Liberles, Ed., *Ancestral Sequence Reconstruction*, Oxford University Press, Oxford, UK, 2007.

[2] W. M. Fitch, "Towards defining the course of evolution: minimum change for a specific tree topology," *Systematic Zoology*, vol. 20, pp. 406–416, 1971.

[3] D. Sankoff and P. Rousseau, "Locating the vertices of a steiner tree in an arbitrary metric space," *Mathematical Programming*, vol. 9, no. 1, pp. 240–246, 1975.

[4] D. Sankoff, "Minimal mutation trees of sequences," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 35–42, 1975.

[5] T. Pupko, A. Doron-Faigenboim, D. A. Liberles, and G. M. Cannarozzi, "Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences," in *Ancestral Sequence Reconsruction*, D. A. Liberles, Ed., Oxford Oxford University Press, 2007.

[6] H. Ashkenazy, O. Penn, A. Doron-Faigenboim et al., "FastML: a web server for probabilistic reconstruction of ancestral sequences," *Nucleic Acids Research*, vol. 40, pp. W580–W584, 2012.

[7] A. Bolshoy and Z. Volkovich, "Whole-genome prokaryotic clustering based on gene lengths," *Discrete Applied Mathematics*, vol. 157, no. 10, pp. 2370–2377, 2009.

[8] F. R. K. Chung, "Some problems and results in labelings of graphs," in *The Theory and Applications of Graphs*, G. Chartland, Ed., pp. 255–263, John Wiley & Sons, New York, NY, USA, 1981.

[9] M. A. Iordanskii, "Minimal numberings of the vertices of trees," *Soviet Mathematics Doklady*, vol. 15, pp. 1311–1315, 1974.

[10] M. K. Goldberg and I. Klipker, "An algorithm for a minimal placement of a tree on a line," *Sakartvelos Mecnierebata Akademiis Moambe*, vol. 83, pp. 553–556, 1976 (Russian).

[11] F. R. K. Chung, "On optimal linear arrangements of trees," *Computers and Mathematics with Applications*, vol. 10, no. 1, pp. 43–60, 1984.

[12] F. R. K. Chung, "On the cutwidth and the topological bandwidth of a tree," *SIAM Journal on Algebraic and Discrete Methods*, vol. 6, pp. 268–277, 1985.

[13] A. Bolshoy and V. Kirzhner, "Algorithms of an optimal integer tree labeling," http://arxiv.org/abs/1305.5551.

[14] B. G. Mirkin, T. I. Fenner, M. Y. Galperin, and E. V. Koonin, "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes," *BMC Evolutionary Biology*, vol. 3, no. 1, article 2, 2003.

[15] J. S. Farris, "Methods for computing Wagner trees," *Systematic Zoology*, vol. 19, pp. 83–92, 1970.

[16] K. Y. Gorbunov and V. A. Lyubetsky, "Reconstructing the evolution of genes along the species tree," *Molecular Biology*, vol. 43, no. 5, pp. 881–893, 2009.

[17] K. Y. Gorbunov and V. A. Lyubetsky, "An algorithm of reconciliation of gene and species trees and inferring gene duplications, losses and horizontal transfers," *Information Processes*, vol. 10, pp. 140–144, 2010 (Russian).

[18] J.-P Doyon, C. Scornavacca, K. Y. Gorbunov, G. J. Szeollosi, V. Ranwez, and V. Berry, "An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers," in *Lecture Notes in Bioinformatics*, S. Istrail, P. Pevzner, and M. Waterman, Eds., vol. 6398 of *Subseries of Lecture Notes in Computer Science*, pp. 93–108, Springer, Berlin, Germany, 2010.

[19] K. Y. Gorbunov and V. A. Lyubetsky, "The tree nearest on average to a given set of trees," *Problems of Information Transmission*, vol. 47, no. 3, pp. 274–288, 2011.

[20] K. Y. Gorbunov and V. A. Lyubetsky, "Fast algorithm to reconstruct a species supertree from a set of protein trees," *Molecular Biology*, vol. 46, no. 1, pp. 161–167, 2012.

[21] V. A. Lyubetsky, L. I. Rubanov, L. Y. Rusin, and K. Y. Gorbunov, "Cubic time algorithms of amalgamating gene trees and building evolutionary scenarios," *Biology Direct*, vol. 7, pp. 1–20, 2012.

[22] L. Y. Rusin, E. V. Lyubetskaya, K. Y. Gorbunov, and V. A. Lyubetsky, "Reconciliation of gene and species trees," *BioMed Research International*. In press.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 407072, 18 pages
http://dx.doi.org/10.1155/2013/407072

Research Article

The Continuing Debate on Deep Molluscan Phylogeny: Evidence for Serialia (Mollusca, Monoplacophora + Polyplacophora)

I. Stöger,[1,2] J. D. Sigwart,[3] Y. Kano,[4] T. Knebelsberger,[5] B. A. Marshall,[6] E. Schwabe,[1,2] and M. Schrödl[1,2]

[1] *SNSB-Bavarian State Collection of Zoology, Münchhausenstraße 21, 81247 Munich, Germany*
[2] *Faculty of Biology, Department II, Ludwig-Maximilians-Universität München, Großhaderner Straße 2-4, 82152 Planegg-Martinsried, Germany*
[3] *Queen's University Belfast, School of Biological Sciences, Marine Laboratory, 12-13 The Strand, Portaferry BT22 1PF, UK*
[4] *Department of Marine Ecosystems Dynamics, Atmosphere and Ocean Research Institute, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8564, Japan*
[5] *Senckenberg Research Institute, German Centre for Marine Biodiversity Research (DZMB), Südstrand 44, 26382 Wilhelmshaven, Germany*
[6] *Museum of New Zealand Te Papa Tongarewa, P.O. Box 467, Wellington, New Zealand*

Correspondence should be addressed to M. Schrödl; michael.schroedl@zsm.mwn.de

Received 1 March 2013; Revised 8 August 2013; Accepted 23 August 2013

Academic Editor: Dietmar Quandt

Molluscs are a diverse animal phylum with a formidable fossil record. Although there is little doubt about the monophyly of the eight extant classes, relationships between these groups are controversial. We analysed a comprehensive multilocus molecular data set for molluscs, the first to include multiple species from all classes, including five monoplacophorans in both extant families. Our analyses of five markers resolve two major clades: the first includes gastropods and bivalves sister to Serialia (monoplacophorans and chitons), and the second comprises scaphopods sister to aplacophorans and cephalopods. Traditional groupings such as Testaria, Aculifera, and Conchifera are rejected by our data with significant Approximately Unbiased (AU) test values. A new molecular clock indicates that molluscs had a terminal Precambrian origin with rapid divergence of all eight extant classes in the Cambrian. The recovery of Serialia as a derived, Late Cambrian clade is potentially in line with the stratigraphic chronology of morphologically heterogeneous early mollusc fossils. Serialia is in conflict with traditional molluscan classifications and recent phylogenomic data. Yet our hypothesis, as others from molecular data, implies frequent molluscan shell and body transformations by heterochronic shifts in development and multiple convergent adaptations, leading to the variable shells and body plans in extant lineages.

1. Introduction

Molluscs are a morphologically megadiverse group of animals with expansive body plan modifications. There is no doubt about the monophyly of Mollusca as a whole or of any of the eight extant molluscan classes, based on strong morphoanatomical evidence and the consensus of molecular studies [1]. Despite a number of important recent studies, resolving ingroup molluscan topology remains contentious (Figure 1(a)) and a major challenge of invertebrate evolution [2].

Other studies have not had access to suitable material for broad taxon sampling, in particular for monoplacophorans, a class of small deep-sea molluscs that still remain rare and largely inaccessible [3, 4]. Among several recent studies on molluscan phylogeny, most use a subset of classes [5–7]; only one phylogenomic study so far has included all eight classes [8].

Multigene studies on ribosomal proteins [6] and housekeeping genes [7] and two broad phylogenomic (EST-based) data sets [5, 8] supported a monophyletic clade Aculifera. This clade comprises those molluscs with a partial or entire body covered by a cuticle with calcareous spicules or scales and is composed of shell-less vermiform molluscs (aplacophoran) and shell-plate bearing Polyplacophora (chitons). The opposing clade Conchifera (incorporating the five classes

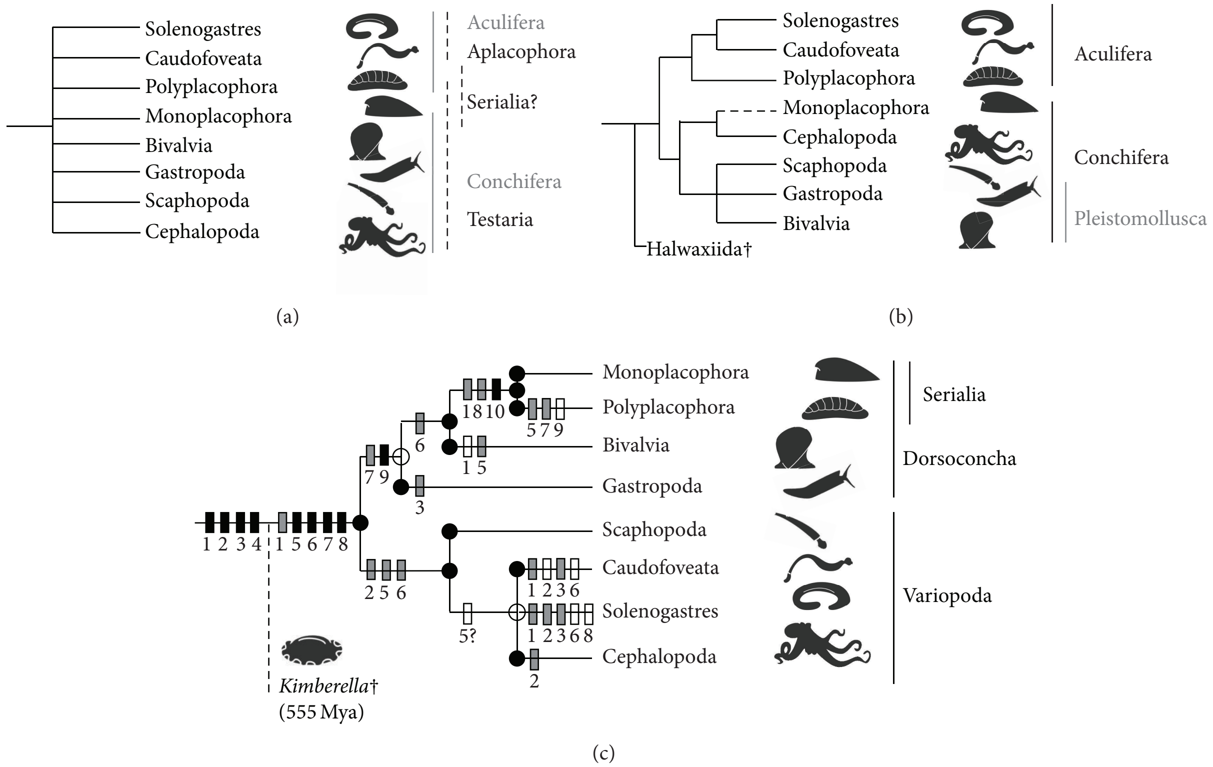


Figure 1: Schematic trees of molluscan relationships. (a) showing traditional proposed subdivisions. (b) consensus tree of two recent molluscan phylogenies inferred from large-scale genomic data by Kocot et al. [5] and Smith et al. [8]. The traditional concepts of Aculifera and Conchifera are supported but with differing positions of scaphopods. Monoplacophora is missing in the data set of Kocot et al. [5] (dotted line reflects the position of Monoplacophora in Smith et al. [8]). (c) the preferred multilocus tree with morphological features indicated numerically on branches. Unfilled dots indicate maximum Bayesian node support, filled dots additional high (>75%) bootstrap support in ML analyses. The Ediacaran fossil genus *Kimberella* corresponds to the description of molluscan stem-group features (1–4, below); crown group taxa originating in the Cambrian and later are united by additional features. Black boxes indicate first appearance of features; grey boxes indicate significant adaptive change; unfilled boxes indicate trait reversals: (1) radula: bipartite in stem molluscs and paedomorphic aplacophorans; broadened, on cartilages and specialised in crown molluscs, stereoglossate-like in Serialia; lost in Bivalvia (and several gastropods); (2), foot with broad gliding sole: transformed into digging foot in variopods (and derived bivalves), narrowed and reduced in aplacophorans, and forming the funnel in cephalopods; (3) circumpedal mantle cavity, miniaturised and anteriorly dislocated in torted gastropods while placed posteriorly in vermiform molluscs; (4) separate mantle covered with cuticula (with calcareous spicules in chitons, aplacophorans, and probably *Kimberella*); (5) dorsal shell: duplicated/fragmented in bivalves and chitons, lost in aplacophorans (and members of most other classes); (6) head with paired appendages: multiplied into feeding tentacles in variopods; trait for head reduction in bivalves plus Serialia and aplacophorans; (7) pericardium: heart fused around intestine in Dorsoconcha; (8) paired ctenidia: expanded to serially repeating gills in Serialia (and nautiloid cephalopods) and reduced in Solenogastres and some gastropod lineages; (9) complex stomach with style (reduced in carnivorous subgroups and chitons; convergently (?) present in a caudofoveate family); (10) paired eightfold dorsoventral muscles; (11) (not shown) statocysts (lost convergently in chitons and aplacophorans); (12) (not shown) suprarectal visceral commissure (subrectal convergently in chitons and aplacophorans).

with a "true" shell) remains controversial; phylogenomic studies recovered a monophyletic clade Conchifera [5, 8], but ribosomal protein multigene and housekeeping gene analyses showed paraphyletic Conchifera [6, 7].

A contradictory alternative hypothesis was proposed by earlier ribosomal RNA-dominated multilocus studies that included Monoplacophora and recovered this class as the sister to Polyplacophora [4, 9, 10]. This clade "Serialia" combines conchiferan and aculiferan members and is thus incompatible with results of recent molecular studies or the morphological Testaria (i.e., Conchifera + Polyplacophora) hypothesis (Figure 1). This result was widely criticised in the literature (e.g., [11]). Yet initial deficiencies [12] of the study by Giribet et al. [9] were addressed by Wilson et al. [10] and Serialia recovered again in a partially overlapping data set by Meyer et al. [13] and independently by Kano et al. [4].

The single phylogenomic data set with a monoplacophoran species also indicated some signal for Serialia, though weaker than that supporting a relationship of cephalopods and monoplacophorans within Conchifera [8]. Phylogenomic data sets cannot yet cover the same density of taxon sampling relative to targeted gene approaches, and while systematic errors of phylogenomic analyses have been explored recently (e.g., [14–16]), there is already a suite of tools available for addressing well-known pitfalls of ribosomal RNA-based sequences (e.g., [17–20]). All data sets may still contribute to ongoing investigations of phylogeny if used and interpreted with care.

Where published topologies differ radically from concepts born from morphoanatomical hypotheses, these results have often been dismissed as artefacts even by the studies' own authors. In addition to the "Serialia" concept, several studies over the last decade have repeatedly recovered Caudofoveata sister to Cephalopoda (e.g., [6, 9, 10, 21–23]). But this pattern has low support values [6, 12]. The position of scaphopods is also highly variable, sometimes in a clade with gastropods and bivalves [5, 7, 8] or sister to aplacophorans and cephalopods [9, 10, 21]. With only eight major clades to rearrange, it could be a serious handicap that many studies exploring molluscan topology have had to exclude one (e.g., [5, 7, 21]) to three (e.g., [4, 6]) classes, and all but one previous study [10] used single-taxon exemplars for at least one [9] to as many as three [7, 8] of those clades. More and better quality data from the monoplacophorans are necessary to resolve molluscan relationships and particularly the two mutually exclusive hypotheses Serialia and Aculifera. We assembled a large multilocus data set for molluscs, including novel sequences of three monoplacophoran species (added to previously published data for only two species, *Veleropilina seisuimaruae* and *Laevipilina hyalina*). To determine the plausibility of this new topology, we applied several tests for phylogenetic informativity, saturation of sites, and compositional heterogeneity within the molecular data sets and have also considered our results against other molecular, morphological, and fossil evidence. Finally we calculated a new time tree via a relaxed molecular clock approach, using multiple sets of fossil calibration points.

Applying carefully calibrated molecular clocks on broad extant taxon sets and reconstructing characters on dated ancient lineages are indispensable for interpretation of enigmatic key fossils such as *Halkieria* or *Nectocaris* that may form part of the early evolutionary history of the group (e.g., [24–27]). We present an alternative view on molluscan evolution that supports the Serialia hypothesis and demonstrates that the debate on pan-molluscan relationships is still in progress.

2. Material & Methods

2.1. DNA Extraction, PCR, and Sequencing. DNA from 12 molluscan taxa, including 3 previously unsampled monoplacophoran species, was extracted using the Qiagen Blood and Tissue Kit (Qiagen, Hilden) by following the manufacturer's instructions. Amplifications of the four standard marker fragments, partial 16S, partial 18S, partial 28S, and complete H3, were carried out under PCR conditions and with primer pairs shown (see Supplementary Material available online at http://dx.doi.org/10.1155/2013/407072). Sequencing reactions were operated on an ABI 3730 48 capillary sequencer of the sequencing service of the Department of Biology of the LMU Munich by using the amplification primers. Newly generated sequences were edited in Sequencher version 4.7 (Gene Codes Inc., Ann Arbor, MI, USA).

2.2. Taxon and Gene Sampling. To compile a comprehensive and dense taxon sampling for resolving deep molluscan relationships, we expanded earlier published data sets [9, 10] by our own and archived (Genbank) data, including a broad selection of outgroups and initially including any molluscs with substantial sequence information available for five standard marker fragments (partial 16S rRNA, partial or complete 18S rRNA, partial 28S rRNA, complete H3, and partial COI). In some poorly sampled but significant ingroup clades we also included species with fragmentary sequence data. Previously unpublished, partial 16S, complete 18S and 28S, complete H3, and partial COI sequences of *Veleropilina seisuimaruae* were provided separately by one of the authors (YK). The total initial data set comprised 158 taxa (141 molluscan and 17 outgroup taxa; Suppl. Table 2).

2.3. Data Cleaning and Alignment. All the downloaded and new single sequences, including all 28S sequences, and all individual amplicons for 18S sequences in Solenogastres, were cross-checked against the nucleotide database of BLAST [29] by using the blastn algorithm. Potentially aberrant or problematic fragments were removed from the data sets (Suppl. Table 3A).

In some bivalve 28S sequences a dubious part of ca. 500 bp was detected in an otherwise homogeneous molluscan alignment. This portion differed substantially in most bivalve taxa but not in all and was highly heterogeneous also in closely related species. No pattern could be observed, so we removed the dubious region (Suppl. Table 3B).

The 18S sequences of Solenogastres were partially excluded due to contamination. Retained sequences of *Epimenia* species (*E.* sp., *E. australis,* and *E. babai*) were aligned separately with the first uncontaminated sequences of Meyer et al. [13], and resulting large gaps were cut by hand according to the template sequences of *Micromenia fodiens, Simrothiella margaritacea* and *Wirenia argentea* (Meyer sequences in [13]).

Patellogastropoda has aberrant 18S and 28S sequences with many indels causing highly incongruent alignments (own observations), leading to long branches and attraction artefacts in previous [13] and our own analyses. Patellogastropoda clustered with long branched Cephalopoda and Solenogastres under different regimes (Table 1). To verify the correct position of Patellogastropoda within or outside other Gastropoda a more focused data set was generated comprising only gastropod taxa plus some selected, short-branched outgroup taxa, that is, two bivalves, two polyplacophorans, one annelid, and one kamptozoan. This alignment is more homogeneous, and patellogastropods appear as a moderately long branch in a rather derived position within the Gastropoda (Suppl. Figure 2). So we confirm that patellogastropods show aberrant evolution leading to long branch attraction artefacts in broader data sets [13]; therefore we excluded this clade from the main analyses.

Single alignments (per fragment) were created with Mafft version 6.847b [30] with the implemented E-INS-i algorithm. Alignments of 16S, 18S, and 28S rRNA were masked with Aliscore version 5.1 [17, 31] by running 10,000,000,000 replicates.

Table 1: Preanalyses comparing different taxon sampling and masking strategies; Mafft [30] and RNAsalsa [18] are alignment methods; Aliscore [17, 31] and Gblocks [35] are masking methods.

Dataset	Alignment treatment	Alignment length (bp)	Major changes in tree topology, compared to main topology (Figure 2, Supplementary Figure 1)
Total set (158 taxa)	Mafft-cut and paste inconsistent blocks in 18S and 28S fragments-Aliscore	10318	Annelida *s.l.* sister to Mollusca; Aplacophora monophyletic (Caudofoveata sister to Solenogastres); Patellogastropoda clusters with Cephalopoda
Total set (158 taxa)	Mafft-RNAsalsa-Aliscore	7597	Mollusca non-monophyletic; Caudofoveata, Solenogastres, Cephalopoda, and Scaphopoda cluster with Annelida *s.l.*; Neritimorpha basal sister to remaining Gastropoda; Patellogastropoda sister to partial Vetigastropoda (Lepetelloida + Vetigastropoda *s.s.*)
Total set (158 taxa)	Mafft-RNAsalsa-Gblocks	4083	Nemertea + Entoprocta + Cycliophora is sister to Mollusca; Heterobranchia is sister to remaining Mollusca; Patellogastropoda clusters with Solenogastres and Cephalopoda
Large set (142 taxa, excluding Patellogastropoda)	Mafft-Gblocks	5550	Annelida *s.l.* + Entoprocta + Cycliophora is sister to Mollusca
Large set (142 taxa, excluding Patellogastropoda)	Mafft-Aliscore	8721	Main analyses (Figure 2, Supplementary Figure 1)

All ambiguous positions were automatically cut with Alicut version 2.0 [17, 31] to remove highly variable positions that could lead to aberrant phylogenetic signals. The alignments of protein coding genes H3 and COI were manually checked for stop codons using MEGA5 [32]. The single data sets were concatenated automatically using FASconCAT version 1.0 [33]. This procedure resulted in a total alignment of 142 taxa with 8721 bp in length and a proportion of 60% gaps (Suppl. Table 5). Where taxon sampling had to be modified, for example, removing taxa or dubious gene fragments, this was done in the initial single data sets and the complete procedure of alignment, masking and concatenation was carried out again.

Final analyses were computed with the large data set excluding Patellogastropoda (142-taxon set), a targeted taxon subset (81-taxon set, alignment length 8367 bp, proportion of gaps 57%) after pruning fast-evolving species or derived members of densely sampled undisputed clades, and the gastropod data set (all gastropods including Patellogastropoda plus selected slowly evolving outgroups). Moreover, we generated and analysed diverse data sets for control reasons to test interclass topologies: the 142- and 81-taxon sets without Aplacophora, the 142-taxon set without long-branched Cephalopoda and Solenogastres, the 142-taxon set with COI and H3 coded as amino acids (142-taxon set amino acid), and one data set that comprises only 18S, 28S, and H3 fragments of the 142-taxon set (Suppl. Table 5). The concatenated sequence matrices of the two main analyses (142-taxon set and 81-taxon set) were deposited at TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S14594). New sequences generated herein were deposited at Genbank (Suppl. Table 2).

2.4. Preanalyses of the Data. Since saturated sequences have minimal or no phylogenetic signal and could even lead to anomalous results, we measured substitution saturation of the protein coding genes, namely, H3 and COI, with Xia's method implemented in DAMBE version 5.2.31 [37]. We used default parameters, and the proportion of invariable sites was specified. The method was executed for all three codon positions together, for combined first and second codon positions, and for third codon position separately. In both cases, H3 and COI, the index of substitution saturation (Iss) values of all three codon positions in combination were significantly smaller than critical index of substitution saturation (Iss.c) values. This was also true for the alignments of first and second codon positions. This assumes that those positions conserve phylogenetic signal and are useful for further analyses. In the case of third codon positions only, substantial saturation could be observed (Iss significantly higher than Iss.c). All results are shown in Supplementary Table 6. Although substitution saturation was observed in third codon positions of H3 and COI, we ran additional analyses with the complete sequence information (1st, 2nd, and 3rd codon positions) to implement potential phylogenetic signal for lower taxonomic levels.

To crosscheck the phylogenetic results of the data sets with and without excluded third codon positions of protein coding genes we conducted the same analyses with all three codon positions included, using distinct models of evolution for the three different codon positions and without third codon positions of H3 and COI.

Testing the evolutionary models for all genes and in case of COI and H3 for every single codon position and for codon positions one and two versus position three was carried out with the programs Modeltest version 3.7 [38] (for complete alignments) and MrModeltest version 2.3 [39] (for codon positions) by the help of PAUP* version 4b10 for Windows [40]. With the amino acid alignments of H3 and COI we additionally tested for the best fitting amino acid model of evolution using ProtTest version 2.4 [41]. As RAxML provides only a part of the models that can potentially be tested by ProtTest we only selected those models in our ProtTest analysis (DAYHOFF, DCMUT, JTT, MTREV, WAG, RTREV, CPREV, VT, BLOSUM62, and MTMAM). The resulting best models for all genes (16S, 18S, 28S, H3, and COI), distinct codon positions of H3 and COI, and amino acid alignments of H3 and COI as well as the corresponding proportions of invariant sites and the gamma distribution shape parameters are shown in Supplementary Table 4.

2.5. Phylogenetic Analyses. Maximum Likelihood (ML) analyses for all data sets were executed using RAxML-HPC for Windows [28] and RAxML version 7.2.6 [28] on the Linux cluster of the Leibniz Computer Centre. Parameters for the initial rearrangement settings and the rate categories were optimised under the GTRCAT model of evolution and a partition by genes (16S, 18S, and 28S) and codon positions (COI, H3) by conducting the hardway analysis described by Stamatakis [42].

First, a set of 10 randomised Maximum Parsimony (MP) starting trees was generated. Second, based on this set of starting trees, the ML trees with a specified setting of initial rearrangements (−i 10) and with an automatically determined initial rearrangement setting had to be inferred. Third, the number of rate categories was adjusted. Initial setting −c 10 was augmented by increments of 10 up to −c 50 for all MP starting trees. The fourth step was to execute 200 inferences on the original alignments. Finally, values of 1000 bootstrap topologies were mapped on the best-scoring ML tree.

Bayesian analyses for selected data sets were conducted with MrBayes v. 3.1.2 [43]. Partitioning with corresponding models of evolution, substitution rates and nucleotide frequencies were applied according to the results of Modeltest [38], MrModeltest [39], and ProtTest [41]. One tree was sampled every 1000 generations. If the average standard deviation of split frequencies declined 0.01 after 5 million generations the analysis was stopped. If not, analysis was continued with another 5 million generations. If the average standard deviation of split frequencies still did not decrease, the log likelihood values were examined with Tracer version 1.5 [44]. If the run reached stationarity, the analysis was stopped. Burn-in was set to 2500 after 5 million generations and to 5000 after 10 million generations.

2.6. Molecular Clock Analyses. Time estimations were performed with the software package BEAST version 1.6.1 [34]. The program is based on the Bayesian Markov Chain Monte Carlo (MCMC) method and therefore can take into account prior knowledge of the data. That is used when nodes in the topology are calibrated and the rate of molecular evolution along the branches is estimated.

We used nine fossil calibration points (Suppl. Table 7) with their corresponding prior distributions and assumed a relaxed clock with a lognormal distribution [45] of the rates for each branch (Suppl. Table 7). This setting is recommended because it additionally gives an indication of how clock-like the data are [46]. Calibration points were set with a minimum bound according to Jörger et al. [47]. To reduce computing time we used the targeted (81-taxa) data set for time estimations. The topology was constrained according to the resulting tree of the phylogenetic analyses.

An Xml-file with all information on data, calibration points, priors and the settings for the MCMC options was created with BEAUti version 1.4.7 [34]. Gamma-shaped priors for all nine calibration points were used (Suppl. Table 7). We assumed that the lower bound of each calibration point is not more than 10% of its maximum age. In case that the next older fossil is within these 10% boundary we used the maximum age of that fossil as lower bound for the younger fossil [48].

Detailed partitioning of genes (16S, 18S, and 28S) and codon positions of COI and H3 and the constraint tree topology were added by hand to the Xml-file. The analysis was executed for 30 million generations, sampling one tree every 1000 generations on the Linux cluster of the Leibniz Computer Centre. The implemented program Tracer version 1.5 [44] was used to confirm that posterior probabilities had reached stationarity. Burn-in was set to 25% (7500), so 22,500 trees were effectively analysed with TreeAnnotator version 1.6.1 [34] to form the summary tree. Further, to check the reliability of our fossils, we repeated the same analysis several times and always omitted one calibration point (Table 2; Suppl. Table 7).

2.7. Testing Hypotheses. Several existing hypotheses about the molluscan interrelationships (Table 3) were tested by executing Approximately Unbiased tests (AU tests) implemented in Treefinder version of October 2008 [36]. Therefore the input constraint trees were computed with RAxML-HPC [28] by using the −g-option and the associated partition by genes and codon positions. Those input tree topologies were tested in Treefinder with maximum number of replicates under the GTR model.

3. Results and Discussion

3.1. Analyses. Analysing traditional multilocus markers for several large taxon sets with Maximum Likelihood and Bayesian methods under different alignment and masking regimes (Table 1, Suppl. Table 5), we recovered consistent phylogenetic trees (Figure 1(c)) with monophyletic Mollusca in contrast to other studies with similar markers [9, 10, 19, 21] and strong support for the monophyly of all molluscan classes, including Bivalvia (also in contrast to some earlier studies [9, 19, 21]).

Our approach included rigorously testing of all amplicons before and after alignment, which led to the exclusion of aberrant or problematic, previously published sequences from the data set (Suppl. Table 3). Criticism of previous accounts using the same set of markers has included the incomplete representation of taxa and the varying extant of missing data [12, 49]. Missing data is a common burden of multilocus studies and will be more severe for phylogenomic approaches [14, 15]. Our preanalyses showed that dubious sequences or ambiguous parts of alignments had much greater effect on the outcome than selecting taxa with the highest amount of data available. Rather than maximizing sequences per species, we concentrated on increasing taxon sampling to minimise potential branch lengths. Our quality controlled 158-taxon set includes 17 lophotrochozoan outgroups. Analytical trials on different subsets of nonmolluscan outgroups altered outgroup topology and support values of some basal ingroup nodes but did not change the ingroup topology (Figure 1).

Alignment issues involved in ribosomal RNA data were addressed by an array of measures proven to be beneficial ([20]; see Section 2). Potential homoplasy in protein coding genes (especially the third codon positions) in our preferred multilocus analysis was addressed by additionally running the analysis with those fragments (COI and H3) encoded as amino acids. This had little effect on the topology but supported monophyletic Aplacophora. We applied a variety of alignment tools, including masking (Aliscore [17]) and refinement algorithms based on secondary structures (RNAsalsa [18]) and applied compartmentalised analyses of taxon clusters causing obvious alignment problems. Excluding patellogastropods (142-taxon set, Suppl. Figure 1; see Section 2) did not change our molluscan backbone topology (Figure 1(c)) but improved alignments. Separately analysing gastropods plus some slowly evolving outgroup taxa shows patellogastropods cluster with vetigastropods (Suppl. Figure 2). Our main aim was to elucidate molluscan relationships at the class level; thus we further pruned outgroups and fast-evolving members from more densely sampled ingroups (such as heterobranch gastropods) and used an 81-taxon set presented here in our main analysis (Figure 2).

3.2. The Basal Molluscan Dichotomy. In our new tree, the phylum Mollusca is divided into two clades (Figure 1(c), Figure 2, Suppl. Figure 1). The first clade is composed of Gastropoda sister to a clade of Bivalvia and Serialia (Monoplacophora + Polyplacophora). For convenience we will refer to this clade as “Dorsoconcha”; the name refers to the (plesiomorphic) presence of a dorsal shell for members of this clade, though modified to two lateral valves in bivalves and to (7-)8 dorsal plates in chitons, and the shell internalised or lost multiple times especially among gastropods.

Gastropods, bivalves, and monoplacophorans are commonly considered to be united by their single shell (secondarily split in bivalves) built by a shell gland at the mantle border (and by the entire mantle roof secreting organic matrix and calcareous layers letting the shell grow thicker, or repair damage). Chitons are traditionally excluded from the hypothetical clade “Conchifera” on the basis of their eight shell plates. The chiton girdle is also covered by a cuticle with embedded calcareous and organic sclerites, similar to the body cuticle of the shell-less aplacophorans, but according to our results, this is convergent and may reflect the different, single versus multicellular spicule formation in these taxa [50]. That chitons cluster with monoplacophorans rather than aplacophorans is congruent to previous molecular approaches that included monoplacophoran exemplars [4, 9, 10, 13]. The exception is the phylogenomic study by Smith et al. [8], in which a single monoplacophoran, *Laevipilina hyalina*, robustly clustered with cephalopods in the main analyses, though parts of the genes used also showed signal supporting an association with chitons.

In the second major molluscan clade, Scaphopoda are sister to a clade of vermiform Caudofoveata and Solenogastres, plus Cephalopoda. Herein we will call this clade “Variopoda,” referring to the various derived foot attributes of its members: the digging foot in Scaphopoda, reduced narrow gliding sole or completely lost in (adult) aplacophorans, and transformed in cephalopods possibly building parts of tentacles and funnel. Dorsoconcha appears as a monophyletic group although bootstrap support is low (60%), and the Variopoda is strongly supported in all Maximum Likelihood analyses; Bayesian posterior probabilities are high for both nodes (Figure 2, Suppl. Figure 1).

The placement of aplacophorans within Variopoda is unconventional, but a sister relationship between Scaphopoda and Cephalopoda has been previously put forward [51, 52]. Previous multilocus approaches with broad taxon sampling (i.e., more than one exemplar of each aplacophoran class) are actually not in general disagreement with Variopoda, since contaminated aplacophoran sequences may account for occasionally aberrant topologies [9, 10, 13]. Inner scaphopod topology resolves the two currently recognised groups Dentaliida and Gadilida, as does Cephalopoda splitting into modern Nautilida and Coleoidea, and is congruent with previous classifications [53].

We calculated time trees with a Bayesian molecular clock approach (Figure 3) using a mix of younger and older calibration points (Suppl. Table 7). We also tested sets of calibrations successively excluding each single calibration point used (Table 2) to minimise circularity involved by calculating individual node times [54]. All our time trees confirm a Precambrian origin of Mollusca (Table 2, Suppl. Table 8) in agreement with previous studies [7], and 95% confidence time bars of all our time trees allow for a Cambrian origin of those classes with a reliable fossil record (Figure 3). As a further sensitivity test we also calculated a time tree from a data set excluding aplacophorans; the topologies are congruent and node ages almost identical, confirming general time estimates (not shown).

Molluscan diversification occurred at an extremely rapid pace after the initial origination of the shell (Figure 3). Short branches at the base of the ingroup can be artefacts of signal erosion in deep nodes [55], but as we discuss below, the rapid early evolution of Mollusca is also supported by the fossil record. Our molecular clock indicates a potential time frame of only around 20–40 million years from the first shelled molluscs (ca. 560–540 Ma) to the presence of differentiated variopod, dorsoconch, gastropod, bivalve, and serialian stem lineages (ca. 520 Ma). The shell was central for rapid evolutionary success of molluscs, and shell modification and divergence are correlated with adaptive radiations during this early period.

3.3. Evaluating Molecular Data Sets. All recent multigene and phylogenomic studies [5–8] have tested the effects of gene sampling, analytical methods, and inference programs; like our results, their topologies were more or less robust, also

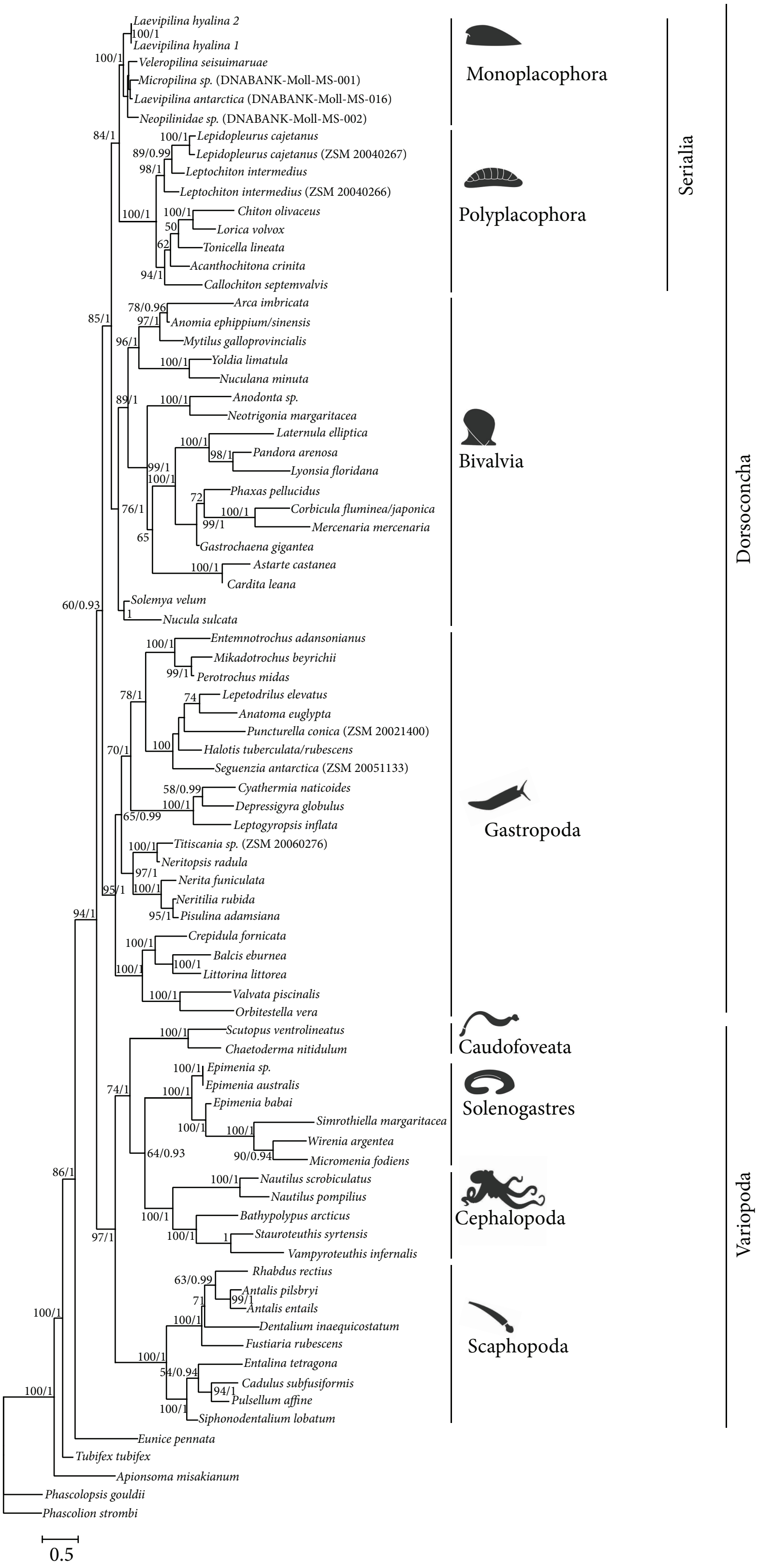


Figure 2: Preferred molluscan tree. Maximum Likelihood analysis (RAxML [28], hardway) of pruned 81-taxon set; values at nodes refer to bootstrap support (1000 pseudoreplicates, first value) and posterior probabilities obtained from the Bayesian analysis (second value).

Table 2: Sensitivity tests of individual calibration nodes used for relaxed molecular clock time estimates of major molluscan groups. Table shows influence of single calibration points on node ages of all other calibration points.

		Excluded calibration point									
		None	Diversification of Mollusca	Split Serialia/ Bivalvia	Origin of Cephalopoda	Split Polyplacophora/ Monoplacophora	Origin of Pteriomorpha	Origin of Caenogastropoda	Diversification of Scaphopoda	Split *Astarte/Cardita*	Diversification of Polyplacophora
Calibrated nodes	Diversification of Mollusca	551.02	**683.50***	550.58	549.76	551.52	551.55	551.59	551.68	552.10	551.72
	Split Serialia/Bivalvia	530.93	533.98	**523.58***	530.80	530.88	530.36	530.82	531.01	530.82	531.01
	Origin of Cephalopoda	504.92	511.40	504.40	**431.05***	504.26	504.85	504.75	503.96	504.16	504.35
	Split Polyplacophora/ Monoplacophora	493.06	493.44	491.79	493.68	**431.68***	493.47	493.13	493.47	493.79	493.14
	Origin of Pteriomorpha	475.06	474.81	474.20	474.43	474.97	**376.65***	474.56	475.20	474.16	474.82
	Origin of Caenogastropoda	421.49	422.82	421.77	421.91	422.61	421.54	**326.50***	421.95	420.85	421.61
	Diversification of Scaphopoda	359.95	360.02	359.97	359.55	360.12	359.59	359.37	**382.50***	359.19	359.60
	Split *Astarte/Cardita*	325.43	325.40	325.20	324.98	325.45	324.55	325.37	325.09	**44.34***	325.63
	Diversification of Polyplacophora	233.44	233.49	233.37	233.18	233.75	233.42	233.63	233.57	232.71	**243.67***

Bold ages marked with an asterisk (∗) indicate time estimations without calibration of this node.

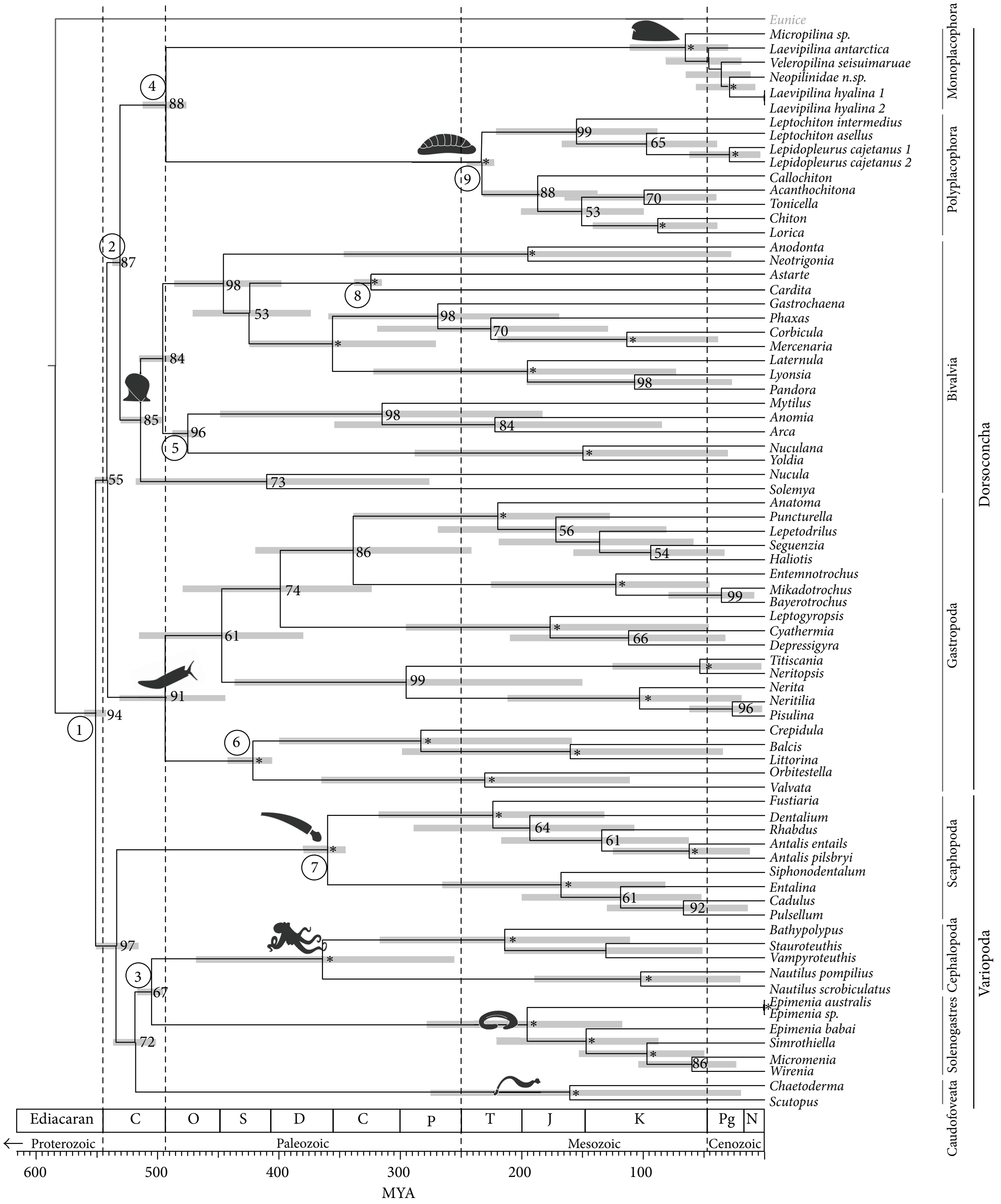


FIGURE 3: Chronogram of molluscan evolution. Divergence times (million years before present, Ma) estimated from BEAST version 1.6.1 [34] under an uncorrelated lognormal relaxed clock model; bars refer to the 95% highest posterior density. All nodes show maximum posterior probabilities (1.0, not indicated) from a run with 10^8 generations (25% burn-in). Numbers at nodes refer to bootstrap support values (>50%; asterisks are 100%) obtained from separate Maximum Likelihood analysis (RAxML [28], hardway, 1000 pseudoreplicates) of the same data set. Circled digits indicate calibrated nodes. Details of calibration can be found in Supplementary Table 7. Omitting Cambrian calibrations shifts molluscan diversification deeper into the Precambrian (for sensitivity analyses see Table 2).

against varying outgroup selection. Sensitivity analyses do not attribute the major split into Variopoda (or parts thereof) and Dorsoconcha or the recovery of Serialia to LBA effects. Yet our multilocus study uses fewer markers and nucleotides than "next-generation sequencing" studies [5–8], so it may be more prone to inadequate signal of certain markers or stochastic errors.

Split decomposition analyses of an earlier multilocus set [9] usually recovered the single monoplacophoran species among bivalves [12], consistent with a Dorsoconcha clade. Splitstree analyses (not shown) of our improved data set still show overall polytomy and some individual taxa are clearly misplaced in the network (e.g., the gastropod *Crepidula* clusters with cephalopods). Overall, most dorsoconch terminals are separated from variopods. Within Dorsoconcha, monoplacophorans cluster with chitons and bivalves. A lack of tree-like structure and *a priori* split support, especially in a large and heterogeneous taxon set, may not necessarily mean that there is too little signal for phylogenetic analyses; it just means that there is conflict that may or may not be resolved applying current models of sequence evolution.

Nuclear ribosomal RNA genes were shown to be informative even on deeper levels than basal molluscs, if treated adequately [20]. Other, supposedly faster-evolving mitochondrial markers (partial COI, 16S) were stringently masked herein, partitioned when necessary or excluded when saturated (Suppl. Tables 4–6). Combined analysis incorporates multiple tempos of evolution experienced by the different loci and is therefore more representative of deep evolutionary patterns. Our backbone topology is robust against varying the taxon and marker sets, masking and partitioning regimes, models of evolution, and methods of analyses (Table 1, Suppl. Tables 4 and 5).

3.4. Evaluating Alternative Morphological and Molecular Concepts. We directly evaluated the statistical fit of major competing morphology- or molecular-based concepts constraining our topologies and calculating their likeliness according to our data set. Using our preferred 81-taxon set with all markers, but also under most other schemes, the AU test rejects all the higher molluscan textbook concepts [1]: the Testaria, Aculifera, Conchifera, Cyrtosoma, and Diasoma hypotheses, with the highest possible statistical support; the same AU tests do not reject Dorsoconcha nor Variopoda (Table 3). We also tested our data against three new molecular concepts (Figure 1(b)): Pleistomollusca (Bivalvia + Gastropoda) established by Kocot et al. [5] and the clades of Monoplacophora and Cephalopoda [8] versus other conchiferans (Scaphopoda, Gastropoda, and Bivalvia) [5, 8]. Only the clade of Monoplacophora and Cephalopoda was not rejected with significant support in any of the main analyses, but all these groups received much lower AU values than our unconstrained topology.

While several recent phylogenomic studies recover Aculifera [5, 7, 8], the Serialia concept has been tested only by Smith et al. [8], by inclusion of a single monoplacophoran species. Though association with cephalopods is preferred, there is a weaker signal also for Serialia [8]. Kano et al. [4] recovered Serialia but did not include any aplacophoran taxa in their data set. The Serialia as a concept cannot be dismissed yet, and our dense taxon sampling herein, though based on far fewer sequences than recent phylogenomic approaches [5, 8], still may allow for a more differentiated and perhaps more correct view on molluscan interclass relationships.

The association of cephalopods and aplacophorans has been recovered previously but dismissed as an artefact of high substitution rates in rRNA genes [6, 13, 21]. But our results cannot easily be explained by long branch attraction (LBA) effects (*contra* [13]). Branch lengths of scaphopods and caudofoveates are moderate, and the variopod node is stable against removal of putative long branched taxa showing accelerated evolutionary rates or biased base compositions [13], such as the branches of Solenogastres or Cephalopoda or both (trees not shown).

Molluscan evolution, whatever the underlying tree, is known to be laden with convergence at all taxon levels, including morphological features previously suspected to be informative for deep phylogeny (e.g., [56–58]). Conclusions derived from single organ systems, or the shell alone, are not able to exclude alternative interpretations. Coding hypothetical bauplans rather than existing representatives has been criticised [59, 60] and may lead to erroneous assumptions especially in groups with uncertain internal topology such as gastropods or aplacophorans. Morphocladistic approaches to date (e.g., [61–63]) all recovered Testaria, but this hypothesis is not supported by any molecular approaches.

Our proposed topology and any other nontestarian hypothesis imply that ancestral molluscs were complex rather than simple. This means that many anatomical characters inherited by descendants may be plesiomorphic and thus not informative, or could have been reduced or lost repeatedly, implying a high level of homoplasy. In fact, early molluscan phylogeny may have been shaped by habitat-induced selective pressure combined with heterochronic processes (e.g., [64]). This combination may lead to concerted morphological parallelisms powerful enough to obfuscate any phylogenetic signal, which has been found to be the case in heterobranch gastropods (e.g., [47, 65, 66]). It is possible to disentangle even highly homoplastic and heterochronic groups (e.g., [67–69]) if detailed and reliable microanatomical data are available on a dense ingroup taxon sampling, which is, however, not yet available for most molluscs. Unfortunately none of the many competing morphology-based hypotheses on molluscan class interrelationships available at present appears to represent a reliable benchmark for evaluating molecular topologies.

3.5. Topologies Tested against the Fossil Record. Molluscan diversification has been widely assumed to originate from a basal "monoplacophoran" bauplan [59], although early single shelled molluscs cannot be reliable separated from gastropods or any nonmonoplacophoran univalve [70]. The earliest calcareous molluscan-like shells, including undisputed molluscs, appear in the uppermost Precambrian, in the late Nemakit-Daldynian ca. 543 Ma [70]. Polyplacophoran shell plates first appear in the Late Cambrian, almost 50 My

Table 3: Testing alternative topologies against various data sets. Results of Approximately Unbiased (AU) tests with Treefinder [36], various schemes. P-values of AU Test executed on selected taxon and data sets. Tested tree topologies were constrained in RAxML [28]. Only meaningful tests have been executed. P-value > 0.05: constrained topology is not rejected; P-value < 0.05: constrained topology is rejected significantly; P-value = 0: constrained topology is rejected with high significance.

Constrained topology	142-taxon set. all markers	81-taxon set. all markers	142-taxon set. 18S + 28S + H3	Aplacophora removed from 142-taxon set. all markers
Sinusoida	0.4244	Not tested	0.2652	0.0383
Mollusca + Kamptozoa	0.0	Not tested	0.0	0.0
Mollusca + Annelida	0.7421	0.7097	0.4090	0.3876
Testaria	0.0	0.0	0.0	Not tested
Aculifera	0.0	0.0	0.0	Not tested
Aplacophora	0.6908	0.3651	0.7730	Not tested
Conchifera	0.0	0.0	0.0	0.0333
Pleistomollusca	0.6665	0.0	0.0863	0.1927
Monoplacophora + Cephalopoda	0.1389	0.0632	0.0	0.2779
Scaphopoda + Gastropoda + Bivalvia	0.0154	0.0	0.0	0.1065
Scaphopoda + Cephalopoda	0.1913	0.0	0.2527	0.6914
Scaphopoda + Cephalopoda + Gastropoda	0.0	0.0	0.0	0.7232
Scaphopoda + Gastropoda	0.8850	0.9452	0.0573	0.8271
Diasoma (Scaphopoda + Bivalvia)	0.0	0.0	0.0	0.0
Monophyletic Protobranchia	0.0219	0.0	0.1085	0.0188
Dorsoconcha	0.6830	0.1097	0.3503	0.4048
Variopoda	0.3170	0.8903	0.6497	0.5952

later [7, 71]. This does not support the Testaria hypothesis that would suggest that chitons evolved before the invention of a true "conchiferan" shell. There are dubious disarticulated microscopic chiton-like plates [72] from the early Meishuchunian (likely Early Tommotian) of China, but these still appeared later rather than earlier than the very first undisputed conchiferan shells. The Aculifera concept with monoplacophorans sister to other members of Conchifera or our molecular basal dichotomy are both fully compatible with the origin of molluscan shells latest at the Precambrian/Cambrian boundary.

The earliest tryblidian monoplacophorans are recorded from the Late Cambrian [73]. Older, nontryblidian "monoplacophorans" do not show serialised muscle scars and thus cannot be considered part of the crown-group. Yet the earliest reliable bivalves with elaborated hinge and ligament (*Fordilla*, *Pojetaia*) appear earlier, in the Early Tommotian ([74]; ca. 535 Ma). Both Aculifera and our basal dichotomy are not contradicted by the early appearance of bivalves. Under an Aculifera topology, chiton-like stem members could appear soon after a terminal Precambrian split separating Aculifera and Conchifera. Interpreting Early to Middle Cambrian sachitids (halwaxiids) as stem aculiferans would help fill this gap [7], but these taxa show a chronological sequence of shell plate loss rather than acquisition, which may be contrary to a progressive transition to chitons. The mosaic taxon *Phthipodochiton*, which has been proposed as a stem aplacophoran, does not appear until the Ordovician [75, 76]; other fossils from the Silurian, combining aplacophoran and polyplacophoran features with some soft tissue preservation, have also been used to support the Aculifera hypothesis [77]. These could also simply represent further disparity in extinct Polyplacophora. Regardless, there is compelling evidence from molecular systematics as well as fossil evidence that aplacophorans lost their ancestral shell (or shell plates) secondarily, and many other groups show repeated shell-loss or evolution to a vermiform body plan.

The topologies recovered by Vinther et al. [7] and Kocot et al. [5] support Aculifera but also imply that cephalopods are sister to Aculifera [7] or represent the earliest-diverging conchiferans [5] (excluding monoplacophorans from the analysis). However, there is no evidence for cephalopod-like fossils appearing earlier than, for example, bivalves. Similarly, bivalves are derived within Conchifera in the topology of Smith et al. [8], which is contradicted by the early fossil record of bivalves. In contrast, our basal dichotomy could fit with the many univalve small shelly fossils occurring earlier in the fossil record than bivalves, and both monoplacophorans and polyplacophorans appear later, actually at a similar time in the Latest Cambrian, and as predicted by a split of Serialia into Monoplacophora and Polyplacophora.

3.6. The Timing of Early Molluscan Evolution. The molluscan stem is Precambrian according to all our molecular time trees. The Vendian (555 Ma) body fossil *Kimberella* was discussed as a mollusc [78], but not widely accepted as such, and rather treated as lophotrochozoan stem member or "no more specifically than as a bilaterian" [79]. According to previous constrained (e.g., [7]) and our less constrained time trees (Table 2, Suppl. Table 8), however, *Kimberella* appears late enough in the fossil record to be considered as a potential stem mollusc. The other recent molecular clock for Mollusca puts the stem Mollusca even deeper [4], but *Kimberella* is within the 95% HPD interval for the split of the basal dichotomy also recovered herein. Having confirmed the conceptual basis of our proposed topology is not rejected by evidence in the fossil record, we further consider the timing of the radiation of specific clades proposed by our molecular clock analyses (Figure 3).

Cap-shaped Helcionellidae from the terminal Precambrian (e.g., *Latouchella*) are putative monoplacophorans according to the seminal study by Runnegar and Pojeta [80] or a separate molluscan class [81] or, based on nonserial muscle scars, gastropods [70]. Our time tree suggests that Nemakit-Daldynian and Earliest Tommotian molluscs with symmetrical cap-shaped shells with large openings are stem molluscs (or in the stem of one part of the basal dichotomy). In contrast, helicoid shells from the same period such as Aldanellidae (e.g., [82]) could well be gastropods, whether or not the animal was torted [70, 82].

Early Tommotian *Watsonella*, formerly known as *Heraultipegma* (the putatively earliest rostroconch), is a laterally compressed, bivalve-like univalve [70], possibly with dorsomedially decalcified or even bivalved shell [83]. This and other laterally compressed Watsonellidae may pre-date the first reliable Bivalvia (Early to Middle Tommotian *Fordilla*; [74] versus [70]) by some million years and thus could well be stem bivalves (or offshoots of the dorsoconch stem) according to our time tree (Figure 3).

It is important to note that neither reliable Monoplacophora (*sensu* Tryblidia) nor reliable Polyplacophora (i.e., Paleoloricata) are known before Late Cambrian, and this is confirmed in our chronograms (Figure 3). Yu [84] interpreted the Early Cambrian Merismoconchia as having eight pairs of muscles on a pseudometameric shell, linking 8-plated chitons with single shelled monoplacophorans in a transitional row of shell fusion. The similarity of merismoconchs with both serialian classes is curious, and their early occurrence in the pretrilobite Meishucun Stage suggests they could be early stem Serialia. The microscopic merismoconchs with their ventrally still connected shell segments and seven observed pairs of muscle scars may have been a transitional stage in how to make a foot efficient for sucking and a shell more flexible to adapt to uneven hard substrates. According to our time tree (Figure 3), chiton-like shell "fragmentation" into fully separated plates occurred much later, after splitting from single-shelled monoplacophoran-like ancestors.

The Cambrian (Atdabanian) *Halkieria* and related Middle Cambrian halwaxiids could also be interpreted as stem Serialia (Figure 3). A role as ancestral lophotrochozoans for halwaxiids as suggested by Edgecombe et al. [79] is not supported by our analysis.

According to our time tree (Figure 3), Yochelcionellidae, conspicuous Tommotian to Middle Cambrian shells that have a "snorkel," could be part of the gastropod radiation as suggested by Parkhaev [70], or members of the dorsoconch stem lineage, or variopod stem members. The latter possibility is especially intriguing, since Yochelcionellidae evolved a "flow-through" water system with two shell openings; a dorsal shell elongates laterally and fuses ventrally, and the body axis shifts towards anterior growth extending head and foot out of a now tube-like shell. This condition is displayed by living and fossil variopods (i.e., scaphopods, cephalopods, and nonwatsonellid Rostroconchia).

Our results show that scaphopods could have split off from the variopod stem earlier, that is, in the Early Cambrian, but the oldest potential scaphopods in the familiar modern tusk-like shape are from the Ordovician [85] or even post-Devonian [86]. There is a vast record of Middle Cambrian tube-like shells that may be unrecognised parts of the early scaphopod diversification that started much earlier and morphologically less constrained than previously expected [87].

Knightoconus, a Middle to Late Cambrian large "monoplacophoran" conical shell with internal septa but no siphuncle [88], was described as a stem cephalopod [80] but subsequently questioned (e.g., [89]) and ultimately suspected to be a brachiopod [90]. *Knightoconus* could fit stratigraphically with stem cephalopods based on our evidence (Figure 3), but its morphological interpretation remains in doubt. The earliest reliable cephalopod fossils are the small bodied, septate, and siphuncle-bearing *Plectronoceras* from the Late Cambrian. Some versions of our analysis used *Plectronoceras* as a soft bound calibration point; by not using *Plectronoceras*, the origin of cephalopods shifts considerably towards the Silurian (Table 2).

Recently, shell-less and coleoid-shaped Lower Cambrian *Nectocaris pteryx* was regarded as a cephalopod [24], but this was immediately rejected on several lines of argument [91, 92]. Other putative Early Cambrian nectocaridids such as *Vetustovermis* [93] are superficially similar to *Nectocaris* in having a pair of long cephalic tentacles and stalked eyes but show a ventral foot separated from the supposedly wing-like mantle. Interpreting *Nectocaris* as having an axial cavity with gills and a funnel would provide synapomorphies for interpreting Nectocarididae as stem cephalopods [24, 94]. Molecular clock estimates can provide further insight to such contentious interpretations; according to our time estimates (which excluded nectocaridids as potential calibration points), *Nectocaris* is too ancient to be a cephalopod (Figure 3). If *Nectocaris* could be accepted as molluscan based on its contentious morphological interpretation, our time trees would be compatible with the idea that nectocaridids are stem variopods or within the stem of an aplacophoran/cephalopod or aplacophoran clade. Nectocaridid features with superficial similarities to coleoid cephalopods [24, 94] instead could be ancestral attributes of variopods: an anteriorly elongated body with head, long and flexible head tentacles, putative preoral hood, and a more or less reduced foot.

The fossil record offers shells and body fossils which, by their occurrence and morphology, at least hypothetically fill our time tree with life. The topology and timing of our hypothesis of early molluscan evolution is not rejected by fossil evidence.

3.7. Dorsoconcha. Molecular, morphological, and palaeontological evidence support (or fail to reject) our basal molluscan dichotomy. The clade Dorsoconcha includes most shelled molluscs and 98% of living species in four classes: Gastropoda, Bivalvia, Polyplacophora, and Monoplacophora.

We note two inferred potential morphological synapomorphies of Dorsoconcha, both relating to the digestive system and both somewhat ambiguous: the intestine is surrounded by the pericardium in basal lineages of gastropods, bivalves, and in monoplacophorans and may be positionally homologous in chitons (Figure 1(c) character 7) and a rotating enzymatic crystalline style (or protostyle; Figure 1(c) character 9). Many basal, noncarnivorous molluscs have a more or less well-developed stomach separated into sorting zones, but only dorsoconchs (and a family of caudofoveates [61]) have the complex style; this was secondarily lost in chitons, which have a derived position in our proposed topology.

Most previous studies on the phylogeny of molluscs have been driven by the Conchifera concept [1, 95] and emphasised the opinion that Serialia violates putative conchiferan synapomorphies [12]. Such features all are plesiomorphic for dorsoconchs in our topology (Figure 1(c)). We note several potential apomorphies for Serialia (Figure 1(c)): the serial (seven or) eightfold (octoserial) dorsoventral pairs of muscle bundles, with two pairs of intertwined muscle bundles in chitons and also partly present in large *Neopilina* [95, 96]; serial gills in a circumpedal mantle cavity; a highly similar cerebral nerve cord; and a longitudinal elongation of the dorsoventrally flattened body, to mention just some (Figure 1(c)). The most prominent feature of Serialia, serial paired foot retractors, is also present in bivalves, but octoserial retractors appear in Ordovician *Babinka* and not in the earliest known bivalves in the Cambrian [97] (Figure 1(c) character 10). While head and buccal apparatus are reduced almost completely in bivalves, Serialia elaborated the buccal mass evolving highly similar radulae and the radula bolster. Similar foot and radula structures in patellogastropod limpets [61] could be either plesiomorphic or convergent, because Patellogastropoda are either an isolated early-diverging gastropod group or relatively recently derived within Vetigastropoda [98, 99].

From this topological result and the available fossil evidence, we propose that the last common ancestor of monoplacophorans and chitons was cap-shelled and adapted to epibenthic life in shallow waters, rasping algae or other microorganisms from rocky substrates (Figure 1). In this scenario, chitons are not primitive molluscs but rather a derived group, potentially adapted to high-energy marine shores. Monoplacophorans initially also were shallow water dwellers [73] but could have colonised deeper waters during the Palaeozoic, where modern monoplacophorans still occur [100]. The Cenozoic or Late Cretaceous molecular dating of the diversification of living monoplacophorans and their short inner branches ([4], Figure 3) are compatible with earlier assumptions of pronounced anagenetic changes in the long stem line of these so-called “living fossils” [4, 100].

3.8. Variopoda. The clade Variopoda (Figure 1(c)) groups the scaphopods, aplacophorans, and cephalopods together in all our analyses, and it is very well supported. We infer several features of variopods, including an apparent propensity for habitat-induced transformations (noted in the taxon epithet; Figure 1(c) character 2). Some other roughly hypothesised apomorphies may refer to a clade of scaphopods and cephalopods only, that is, to variopods only under the assumption that aplacophorans represent highly paedomorphic and thus aberrant offshoots (see below): lateral extension of a primitively dorsal cap-like shell forming a tube; twisting the growth axis during ontogeny from initial dorsoventral to an anterior body extension, translocating head foot and mantle cavity with anal opening anteriorly; formation of a ring-like dorsoventral muscle insertion; multiplication of cephalic tentacles into prey-capturing feeding tentacles; and at least partly using muscle antagonist rather than merely hydrostatic systems in these tentacles (convergently in gastropod cephalic sensory tentacles); a hood is formed anterior to the mouth; and muscular retraction of the foot is used to pump water, waste, and gametes through/out the mantle cavity.

A clade of scaphopods and cephalopods repeatedly has been proposed based on morphological data, sometimes with one or the other or both together allied with gastropods [1], and was recovered by molecular data [52] and broadly within some pan-molluscan molecular phylogenies [10, 21]. In contrast, morphocladistic neontological [101] and palaeontological studies (e.g., [80, 102]) advocated the Diasoma concept suggesting scaphopods as sister to bivalves with a rostroconch ancestor. Developmental data showing different ontogeny of shells have not supported the latter opinion [103]. Diasoma has been equivocally recovered within one mitogenomic analysis ([104], but see [105] for limitations of protein coding mitochondrial genes), and in one supplementary analysis of transcriptome data [8]. Similar features such as a digging foot could be interpreted as convergent adaptations to infaunal life.

The two aplacophoran classes Caudofoveata and Solenogastres have never been associated with either scaphopods or cephalopods in morphological studies. In our analyses aplacophorans are usually paraphyletic, but some permutations, in particular when excluding (the faster-evolving, but stringently masked) COI and 16S markers, recover a clade Aplacophora sister to Cephalopoda. Aplacophora as a clade is not rejected by AU analyses of the combined 5-marker set either (Table 3). A single origin of vermiform body plans in the cephalopod stem lineage could arguably be more parsimonious than arising twice independently. Monophyly of Aplacophora is indicated by all recent studies using multiple nuclear protein coding genes and phylogenomic data sets ([5, 7, 8]; Figure 1(b)) but not neuroanatomy [106].

Aplacophorans may share an inferred tendency of modifying the ancestral foot, they have an elongated body with a foot (or head) shield with strong retractor muscle in caudofoveates, and the atrial cavity especially in Solenogastres could be interpreted as a modified preoral hood, as remnants of a hypothesised variopod body plan. Yet there is no morphological indication for a specifically aplacophoran-cephalopod clade. Interpretation of the vermiform molluscan morphology as progenetically derived rather than reflecting a basal molluscan condition (also assumed under the Aculifera concept) actually allows for hypotheses that resolve them at any position in the molluscan tree (or makes their position impossible to recover using currently available anatomical data). Assuming that aplacophorans (once or twice independently) initially evolved into interstitial secondary worms could be correlated with precerebral ganglia present in caudofoveates [106]; these transformations have evolved many times independently in interstitial worm-like gastropod groups, which are likely progenetic [47]. Calcareous spicules also evolved many times convergently in different interstitial shell-less gastropod lineages [47] and a protective dorsal cuticle covering the body evolved within progenetic corambid sea slugs [67, 68]. "Regressive" [*sensu* [107]] traits in aplacophorans such as miniaturisation, losses of shell, tentacles, and cephalisation have been attributed to progenesis [64]. The serial dorsoventral muscle grid of aplacophorans resembles early ontogenetic stages observed in other molluscs [108] and could be paedomorphic, but it is still an adaptive innovation for nonlarval stages. The narrow bipartite radulae of aplacophorans are specialised tools for microcarnivory but also resemble some stem molluscan radula types [56]; evidence from Cambrian fossils is more congruent with an ancestral unipartite radula [109].

Our topology places aplacophorans in an unconvential position; however, there is consensus among all recent molecular studies that aplacophorans represent derived rather than plesiomorphic members of Mollusca (Figure 1). These notes on the specific feature of aplacophorans therefore are of general interest to resolving the pattern and tempo of molluscan evolution, regardless of differences between our new topology and other studies.

3.9. Molluscan Ancestors. The origin of molluscs is a long-standing question, and speculations on the "hypothetical ancestral mollusk" depend on character-polarity and even topological assumptions [1, 59]. Broad genomic analyses (e.g., [14–16, 22]) recovered molluscs as an early-derived offshoot of Lophotrochozoa (Spiralia), as had been proposed on morphological grounds [110]. Modern morphological studies suggest entoprocts as sister to molluscs [1], a view supported by mitochondrial genomics [105]. MicroRNA data [111] suggest Annelida is the sister to Mollusca, as recovered (but never robustly supported) by most of our analyses with a large outgroup taxon set (Suppl. Figure 1A). Our analyses did not resolve a consistent sister group to Mollusca. Yet permutations and pruning of our outgroup sampling did not affect ingroup topologies. We regard the molluscan sister group as an unanswered question, but not necessarily problematic to the question of internal molluscan phylogeny (if ingroup taxon sampling is sufficiently dense).

Our initial morphological character mapping (Figure 1(c)) suggests that the last common ancestor of living molluscs ("LAM") was a single-shelled conchiferan with a complex body, single (or few) paired shell retractors, single paired gills in a circumpedal mantle cavity, and an elaborated (cephalised) anterior body portion. There is little reason to assume that this hypothetical LAM resembled a chiton or monoplacophoran (e.g., [25]) or to suspect a segmented body organisation (e.g., [112]). Instead, the LAM may have resembled an untorted gastropod with a cap-like shell, perhaps similar to *Latouchella*, as assumed by morphologists before the discovery of the supposedly "living fossil" *Neopilina* and still advocated by some palaeontologists [70].

Our assessment of potential morphological apomorphies (Figure 1(c)) and the molecular clock results (Figure 3) would suggest that the Vendian (555 Ma) *Kimberella* [78] is a candidate stem-group mollusc appearing before the evolution of a dorsal shell field. The interpretation of *Kimberella* is controversial [113], but the true stem molluscs probably did have a large, bilaterally symmetrical body with subapical mouth on a snout with a likely bipartite radula [114], a broad ventral foot, many dorsoventral muscle bundles, and a dorsal mantle covered with a resistant dorsal cuticle with mineralised spicules, which are all molluscan features, but lacking a shell [115, 116]. During the latest Precambrian rise of predators and successive development of sediment bottoms [25], molluscan larvae or early juveniles may have calcified their plesiomorphic cap-shaped mantle cuticle for protective reasons. Answering Yochelson [117], the mollusc made a shell, but then the shell made the molluscs.

4. Conclusions

Only one (if any) of the dozens of proposed hypotheses on molluscan phylogeny reflects the true tree. Both the traditional palaeontological concept, with monoplacophorans giving rise to all other molluscan lineages, and the widely accepted morphocladistic Testaria hypothesis, with progressive evolution from vermiform molluscs to chitons and conchiferans [62, 118], are not supported by molecular evidence and are apparently incompatible with the chronological appearance of reliable fossils representing major molluscan lineages.

The Aculifera concept has been supported by phylogenomic results [2, 119], whose dichotomy is not inherently contradicted by the available fossil record if the last common molluscan ancestor was small and complex and had a shell (i.e., was conchiferan rather than chiton-like). Yet the branching patterns of living clades in available phylogenomic topologies appear to be incongruent with stratigraphic evidence. The debate on molluscan phylogeny can only be progressed using all available evidence, integrating morphological, fossil, and molecular data. To provide meaningful insights, molecular approaches must include all eight molluscan classes and cover the well-known diversity of living taxa.

Our results, despite using traditional markers that cover arguably less data than next-generation approaches, are based on a comprehensive taxon set with data quality checked exhaustively at all levels. Topologies recovered still may suffer from poor sampling especially of aplacophoran lineages and from heterogeneous evolution of ingroup clades such as cephalopods or patellogastropods. The data available, while extensive and of high quality, are small in comparison to the total genetic diversity of the phylum under study.

Nevertheless, our data sets, regimes, and analyses support and refine the Serialia hypothesis [9]. The topological results inferred herein cannot be refuted by recent research on shell building gene expression and mollusc palaeontology. In many well-studied molluscan taxa, shells are reduced or duplicated, bodies adapted to different environments and life styles such as benthic, interstitial, or pelagic realms, and features such as mantle cavities and radulae repeatedly were transformed, often drastically and rapidly. Heterochronic processes could already have occurred in the Palaeozoic, which would be consistent with the disparity known in living molluscs but which could also obscure deeper phylogenetic signal in morphological analyses. Ultimately, such complex diversification could have led to the fossil and extant molluscs that stand apart from other (noninsect) animals in terms of species diversity, body disparity, and variation of life traits. The true reconstruction of the early radiation of molluscs still is one of the major unresolved issues in evolutionary biology. Independent molecular evidence, such as microRNAs or phylogenomic data on a similarly comprehensive and dense taxon sampling as used herein, will be needed to further test these hypotheses.

Authors' Contributions

I. Stöger carried out the molecular genetic studies, performed the sequence alignments and the phylogenetic analyses, and drafted the paper. J. D. Sigwart participated in the design of the study and contributed to writing the paper; Y. Kano participated in original fieldwork and in the molecular genetic studies; T. Knebelsberger participated in the molecular genetic studies; B. A. Marshall carried out original fieldwork and helped draft the paper; E. Schwabe participated in original fieldwork and helped draft the paper; M. Schrödl conceived and designed the study, participated in original fieldwork, contributed to molecular genetic studies, and contributed to writing the paper.

Acknowledgments

Monoplacophoran material was collected during the expedition ANDEEP SYSTCO, from commercial fisheries bycatch, and by Taeko and Shoichi Kimura. Dirk Metzler (Munich) helped with molecular clocks, Patrick Kück (Bonn) performed some of the analyses with removed putative long branched molluscan classes by the help of a bioinformatic pipeline, Korbinian Trumpp provided computational support, and Gerhard Haszprunar (Munich) is thanked for critical discussions. Anonymous referees on a previous, short version of this paper greatly contributed to improve analyses and interpretations. The authors are grateful for additional comments from David Lindberg (Berkeley) and an anonymous reviewer. This study was funded by the Deep Metazoan Priority Programme of the German Research Foundation (DFG667/9-1 to MS). Further support came from the GeoBioCenter LMU and the Leibniz Computer Centre (both Munich).

References

[1] G. Haszprunar, C. Schander, and K. M. Halanych, "Relationships of higher taxa," in *Phylogeny and Evolution of the Mollusca*, W. F. Ponder and D. R. Lindberg, Eds., pp. 19–32, University of California Press, Berkeley, Calif, USA, 2008.

[2] M. J. Telford and G. E. Budd, "Invertebrate evolution: bringing order to the molluscan chaos," *Current Biology*, vol. 21, no. 23, pp. R964–R966, 2011.

[3] N. G. Wilson, D. Huang, M. C. Goldstein, H. Cha, G. Giribet, and G. W. Rouse, "Field collection of *Laevipilina hyalina* McLean, 1979 from southern California, the most accessible living monoplacophoran," *Journal of Molluscan Studies*, vol. 75, no. 2, pp. 195–197, 2009.

[4] Y. Kano, S. Kimura, Y. Kimura, and A. Warén, "Living Monoplacophora: morphological conservatism or recent diversification?" *Zoologica Scripta*, vol. 41, no. 5, pp. 471–488, 2012.

[5] K. M. Kocot, J. T. Cannon, C. Todt et al., "Phylogenomics reveals deep molluscan relationships," *Nature*, vol. 477, no. 7365, pp. 452–456, 2011.

[6] A. Meyer, A. Witek, and B. Lieb, "Selecting ribosomal protein genes for invertebrate phylogenetic inferences: how many genes to resolve the Mollusca?" *Methods in Ecology and Evolution*, vol. 2, no. 1, pp. 34–42, 2011.

[7] J. Vinther, E. A. Sperling, D. E. G. Briggs, and K. J. Peterson, "A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors," *Proceedings of the Royal Society B*, vol. 279, no. 1732, pp. 1259–1268, 2012.

[8] S. A. Smith, N. G. Wilson, F. E. Goetz et al., "Resolving the evolutionary relationships of molluscs with phylogenomic tools," *Nature*, vol. 480, no. 7377, pp. 364–367, 2011.

[9] G. Giribet, A. Okusu, A. R. Lindgren, S. W. Huff, M. Schrödl, and M. K. Nishiguchi, "Evidence for a clade composed of molluscs with serially repeated structures: monoplacophorans are related to chitons," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 20, pp. 7723–7728, 2006.

[10] N. G. Wilson, G. W. Rouse, and G. Giribet, "Assessing the molluscan hypothesis Serialia (Monoplacophora + Polyplacophora) using novel molecular data," *Molecular Phylogenetics and Evolution*, vol. 54, no. 1, pp. 187–193, 2010.

[11] C. Nielsen, G. Haszprunar, B. Ruthensteiner, and A. Wanninger, "Early development of the aplacophoran mollusc *Chaetoderma*," *Acta Zoologica*, vol. 88, no. 3, pp. 231–247, 2007.

[12] J. W. Wägele, H. Letsch, A. Klussmann-Kolb, C. Mayer, B. Misof, and H. Wägele, "Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny)," *Frontiers in Zoology*, vol. 6, no. 1, p. 12, 2009.

[13] A. Meyer, C. Todt, N. T. Mikkelsen, and B. Lieb, "Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 70, 2010.

[14] H. Philippe, H. Brinkmann, D. V. Lavrov et al., "Resolving difficult phylogenetic questions: why more sequences are not enough," *PLoS Biology*, vol. 9, no. 3, Article ID e1000602, 2011.

[15] H. Philippe, H. Brinkmann, R. R. Copley et al., "Acoelomorph flatworms are deuterostomes related to *Xenoturbella*," *Nature*, vol. 470, no. 7333, pp. 255–260, 2011.

[16] K. S. Pick, H. Philippe, F. Schreiber et al., "Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships," *Molecular Biology and Evolution*, vol. 27, no. 9, pp. 1983–1987, 2010.

[17] B. Misof and K. Misof, "A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion," *Systematic Biology*, vol. 58, no. 1, pp. 21–34, 2009.

[18] R. R. Stocsits, H. Letsch, J. Hertel, B. Misof, and P. F. Stadler, "Accurate and efficient reconstruction of deep phylogenies from structured RNAs," *Nucleic Acids Research*, vol. 37, no. 18, pp. 6184–6193, 2009.

[19] J. Mallatt, C. W. Craig, and M. J. Yoder, "Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction," *Molecular Phylogenetics and Evolution*, vol. 55, no. 1, pp. 1–17, 2010.

[20] J. Paps, J. Baguñà, and M. Riutort, "Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes," *Proceedings of the Royal Society B*, vol. 276, no. 1660, pp. 1245–1254, 2009.

[21] Y. J. Passamaneck, C. Schander, and K. M. Halanych, "Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences," *Molecular Phylogenetics and Evolution*, vol. 32, no. 1, pp. 25–38, 2004.

[22] C. W. Dunn, A. Hejnol, D. Q. Matus et al., "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature*, vol. 452, no. 7188, pp. 745–749, 2008.

[23] A. Hejnol, M. Obst, A. Stamatakis et al., "Assessing the root of bilaterian animals with scalable phylogenomic methods," *Proceedings of the Royal Society B*, vol. 276, no. 1677, pp. 4261–4270, 2009.

[24] M. R. Smith and J.-B. Caron, "Primitive soft-bodied cephalopods from the Cambrian," *Nature*, vol. 465, no. 7297, pp. 469–472, 2010.

[25] J.-B. Caron, A. H. Scheltema, C. Schander, and D. Rudkin, "A soft-bodied mollusc with radula from the Middle Cambrian Burgess Shale," *Nature*, vol. 442, no. 7099, pp. 159–163, 2006.

[26] J. Vinther and C. Nielsen, "The early Cambrian *Halkieria* is a mollusc," *Zoologica Scripta*, vol. 34, no. 1, pp. 81–89, 2005.

[27] S. C. Morris and J.-B. Caron, "Halwaxiids and the early evolution of the lophotrochozoans," *Science*, vol. 315, no. 5816, pp. 1255–1258, 2007.

[28] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006.

[29] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[30] K. Katoh, G. Asimenos, and H. Toh, "Multiple alignment of DNA sequences with MAFFT," *Methods in Molecular Biology*, vol. 537, pp. 39–64, 2009.

[31] P. Kück, K. Meusemann, J. Dambach et al., "Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees," *Frontiers in Zoology*, vol. 7, p. 10, 2010.

[32] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[33] P. Kück and K. Meusemann, *FASconCAT, Version 1. 0. Zool*, Forschungsmuseum A. Koenig, Bonn, Germany, 2010.

[34] A. J. Drummond and A. Rambaut, "BEAST: bayesian evolutionary analysis by sampling trees," *BMC Evolutionary Biology*, vol. 7, no. 1, pp. 754–755, 2007.

[35] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[36] G. Jobb, A. Von Haeseler, and K. Strimmer, "TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics," *BMC Evolutionary Biology*, vol. 4, p. 18, 2004.

[37] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, no. 4, pp. 371–373, 2001.

[38] D. Posada and K. A. Crandall, "MODELTEST: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[39] J. A. A. Nylander, *MrModeltest V2. 3. Program Distributed by the Author*, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden, 2008.

[40] D. L. Swofford, *PAUP. Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4*, Sinauer Associates, Sunderland, Mass, USA, 2003.

[41] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[42] A. Stamatakis, The RAxML 7.0.4 Manual. Department of Computer Science, Ludwig-Maximilian-Universität München, Germany.

[43] J. P. Huelsenbeck and F. Ronquist, "MRBAYES: bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[44] A. Rambaut and A. J. Drummond, "Tracer v1. 4," 2007, http://tree.bio.ed.ac.uk/software/tracer/.

[45] A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut, "Relaxed phylogenetics and dating with confidence," *PLoS Biology*, vol. 4, no. 5, p. e88, 2006.

[46] A. J. Drummond, S. Y. W. Ho, N. Rawlence, and A. Rambaut, *A Rough Guide to BEAST 1.4*, 2007.

[47] K. M. Jörger, I. Stöger, Y. Kano, H. Fukuda, T. Knebelsberger, and M. Schrödl, "On the origin of Acochlidia and other enigmatic euthyneuran gastropods, with implications for the systematics of Heterobranchia," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 323, 2010.

[48] C. A. Hipsley, L. Himmelmann, D. Metzler, and J. Müller, "Integration of bayesian molecular clock methods and fossil-based soft bounds reveals early cenozoic origin of African lacertid lizards," *BMC Evolutionary Biology*, vol. 9, no. 1, p. 151, 2009.

[49] J. W. Wägele and C. Mayer, "Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects," *BMC Evolutionary Biology*, vol. 7, p. 147, 2007.

[50] T. Furuhashi, C. Schwarzinger, I. Miksik, M. Smrz, and A. Beran, "Molluscan shell evolution with review of shell calcification hypothesis," *Comparative Biochemistry and Physiology B*, vol. 154, no. 3, pp. 351–371, 2009.

[51] T. R. Waller, "Origin of the molluscan class Bivalvia and a phylogeny of major groups," in *Bivalves: An Eon of Evolution*, P. A. Johnston and J. W. Haggard, Eds., pp. 1–45, University of Calgary Press, Calgary, Canada, 1998.

[52] G. Steiner and H. Dreyer, "Molecular phylogeny of Scaphopoda (Mollusca) inferred from 18s rDNA sequences: support for a Scaphopoda-Cephalopoda clade," *Zoologica Scripta*, vol. 32, no. 4, pp. 343–356, 2003.

[53] W. F. Ponder and D. R. Lindberg, *Phylogeny and Evolution of the Mollusca*, University of California Press, Berkeley, Calif, USA, 2008.

[54] P. C. E. Donoghue and M. J. Benton, "Rocks and clocks: calibrating the Tree of Life using fossils and molecules," *Trends in Ecology and Evolution*, vol. 22, no. 8, pp. 424–431, 2007.

[55] E. Mossel and M. Steel, "How much can evolved characters tell us about the tree that generated them?" in *Mathematics of Evolution and Phylogeny*, O. Gascuel, Ed., pp. 384–412, Oxford University Press, Oxford, UK, 2005.

[56] A. H. Scheltema, K. Kerth, and A. M. Kuzirian, "Original molluscan radula: comparisons among Aplacophora, Polyplacophora, Gastropoda, and the Cambrian fossil *Wiwaxia corrugata*," *Journal of Morphology*, vol. 257, no. 2, pp. 219–244, 2003.

[57] S. Shigeno, T. Sasaki, and G. Haszprunar, "Central nervous system of *Chaetoderma japonicum* (Caudofoveata, Aplacophora): implications for diversified ganglionic plans in early molluscan evolution," *Biological Bulletin*, vol. 213, no. 2, pp. 122–134, 2007.

[58] K. Lundin, C. Schander, and C. Todt, "Ultrastructure of epidermal cilia and ciliary rootlets in Scaphopoda," *Journal of Molluscan Studies*, vol. 75, no. 1, pp. 69–73, 2009.

[59] D. R. Lindberg and M. T. Ghiselin :, "Fact, theory and tradition in the study of molluscan origins," *Proceedings of the California Academy of Sciences*, vol. 54, pp. 663–686, 2003.

[60] L. Prendini, "Species or supraspecific taxa as terminals in Cladistic analysis? groundplans versus exemplars revisited," *Systematic Biology*, vol. 50, no. 2, pp. 290–300, 2001.

[61] G. Haszprunar, "Is the Aplacophora monophyletic? A cladistic point of view," *American Malacological Bulletin*, vol. 15, no. 2, pp. 115–130, 2000.

[62] G. Haszprunar and A. Wanninger, "Molluscs," *Current Biology*, vol. 22, no. 13, pp. R510–R514, 2012.

[63] L. Salvini-Plawen and G. Steiner, "Synapomorphies and plesiomorphies in higher classification of Mollusca," in *Origin and Evolutionary Radiation of the Mollusca*, J. Taylor, Ed., pp. 29–51, Oxford University Press, Oxford, UK, 1996.

[64] A. H. Scheltema, "Phylogenetic position of Sipuncula, Mollusca and the progenetic Aplacophora," in *Origin and Evolutionary Radiation of the Mollusca*, J. Taylor, Ed., pp. 53–58, Oxford University Press, Oxford, UK, 1996.

[65] M. Schrödl, K. M. Jörger, A. Klussmann-Kolb, and N. G. Wilson, "Bye bye "Opisthobranchia"! a review on the contribution of mesopsammic sea slugs to euthyneuran systematics," *Thalassas*, vol. 27, no. 2, pp. 101–112, 2011.

[66] M. Schrödl, K. M. Jörger, and N. G. Wilson, "A reply to Medina et al. (2011): crawling through time: transition of snails to slugs dating back to the Paleozoic based on mitochondrial phylogenomics," *Marine Genomics*, vol. 4, no. 4, pp. 301–303, 2011.

[67] A. Martynov, B. Brenzinger, Y. Hooker, and M. Schrödl, "3D-anatomy of a new tropical Peruvian nudibranch gastropod species, *Corambe mancorensis*, and novel hypotheses on dorid gill ontogeny and evolution," *Journal of Molluscan Studies*, vol. 77, no. 2, pp. 129–141, 2011.

[68] A. Martynov and M. Schrödl, "Phylogeny and evolution of corambid nudibranchs (Mollusca: Gastropoda)," *Zoological Journal of the Linnean Society*, vol. 163, no. 2, pp. 585–604, 2011.

[69] M. Schrödl and T. P. Neusser, "Towards a phylogeny and evolution of Acochlidia (Mollusca: Gastropoda: Opisthobranchia)," *Zoological Journal of the Linnean Society*, vol. 158, no. 1, pp. 124–154, 2010.

[70] P. Y. Parkhaev, "The early molluscan radiation," in *Phylogeny and Evolution of the Mollusca*, W. F. Ponder and D. R. Lindberg, Eds., pp. 33–69, University of California Press, Berkeley, Calif, USA, 2008.

[71] A. G. Smith, "Amphineura," in *Treatise on Invertebrate Palaeontology, Part 1: Mollusca 1*, R. C. Moore, Ed., pp. 141–176, Geological Society of America, New York, NY, USA, 1960.

[72] W. Yu, "Yangtze micromolluscan fauna in Yangtze region of China with notes on Precambrian-Cambrian boundary," in *Stratigraphy and Palaeontology Boundary in China Precambrian-Cambrian Boundary*, vol. 1, pp. 19–275, Nanjing Institute of Geology and Palaeontology Academia Sinica, Nanjing University, Jiangsu, China, 1987.

[73] D. R. Lindberg, "Monoplacophorans and the origin and relationships of mollusks," *Evolution*, vol. 2, no. 2, pp. 191–203, 2009.

[74] J. Pojeta Jr., "Cambrian Pelecypoda (Mollusca)," *American Malacological Bulletin*, vol. 15, no. 2, pp. 157–166, 2000.

[75] J. D. Sigwart and M. D. Sutton, "Deep molluscan phylogeny: synthesis of palaeontological and neontological data," *Proceedings of the Royal Society B*, vol. 274, no. 1624, pp. 2413–2419, 2007.

[76] M. D. Sutton and J. D. Sigwart, "A chiton without a foot," *Palaeontology*, vol. 55, no. 2, pp. 401–411, 2012.

[77] M. D. Sutton, D. E. G. Briggs, D. J. Siveter, D. J. Siveter, and J. D. Sigwart, "A Silurian armoured aplacophoran and implications for molluscan phylogeny," *Nature*, vol. 490, pp. 94–97, 2012.

[78] M. A. Fedonkin and B. M. Waggoner, "The late Precambrian fossil *Kimberella* is a mollusc-like bilaterian organism," *Nature*, vol. 388, no. 6645, pp. 868–871, 1997.

[79] G. D. Edgecombe, G. Giribet, C. W. Dunn et al., "Higher-level metazoan relationships: recent progress and remaining questions," *Organisms Diversity and Evolution*, vol. 11, no. 2, pp. 151–172, 2011.

[80] B. Runnegar and J. Pojeta Jr., "Molluscan phylogeny: the paleontological viewpoint," *Science*, vol. 186, no. 4161, pp. 311–317, 1974.

[81] E. L. Yochelson, "An alternative approach to the interpretation of the phylogeny of ancient molluscs," *Malacologia*, vol. 17, pp. 185–191, 1978.

[82] J. A. Harper and H. B. Rollins, "The bellerophont controversy revisited," *American Malacological Bulletin*, vol. 15, no. 2, pp. 147–156, 2000.

[83] J. Dzik, "Evolution of "small shelly fossils" assemblages of the early Paleozoic," *Acta Palaeontologica Polonica*, vol. 39, no. 3, pp. 247–313, 1994.

[84] W. Yu, "On merismoconchids," *Acta Oceanologica Sinica*, vol. 23, pp. 432–446, 1983.

[85] A. P. Gubanov and J. S. Peel, "Cambrian monoplacophoran molluscs (Class Helcionelloida)," *American Malacological Bulletin*, vol. 15, no. 2, pp. 139–145, 2000.

[86] E. L. Yochelson and C. H. Holland, "*Dentalium saturni* Goldfuss, 1841 (Eifelian: Mollusca): complex issues from a simple fossil," *Paläontologische Zeitschrift*, vol. 78, no. 1, pp. 97–102, 2004.

[87] J. S. Peel, "*Pinnocaris* and the origin of scaphopods," *Acta Palaeontologica Polonica*, vol. 49, no. 4, pp. 543–550, 2004.

[88] E. Yochelson, R. H. Flower, and G. F. Webers, "The bearing of new Late Cambrian monoplacophoran genus *Knightoconus*

upon the origin of Cephalopoda," *Lethaia*, vol. 6, no. 3, pp. 275–309, 1973.

[89] G. F. Webers and E. L. Yochelson, "Carboniferous Scaphopoda (Mollusca) and non-scaphopods from Scotland," *Scottish Journal of Geology*, vol. 47, no. 1, pp. 67–79, 2011.

[90] J. Dzik, "Brachiopod identity of the alleged monoplacophoran ancestors of cephalopods," *Malacologia*, vol. 52, no. 1, pp. 97–113, 2010.

[91] B. Kröger, J. Vinther, and D. Fuchs, "Cephalopod origin and evolution: a congruent picture emerging from fossils, development and molecules: extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators," *BioEssays*, vol. 33, no. 8, pp. 602–613, 2011.

[92] D. Mazurek and M. Zatoń, "Is *Nectocaris pteryx* a cephalopod?" *Lethaia*, vol. 44, no. 1, pp. 2–4, 2011.

[93] J.-Y. Chen, D.-Y. Huang, and D. J. Bottjer, "An Early Cambrian problematic fossil: *Vetustovermis* and its possible affinities," *Proceedings of the Royal Society B*, vol. 272, no. 1576, pp. 2003–2007, 2005.

[94] M. R. Smith and J.-B. Caron, "*Nectocaris* and early cephalopod evolution: reply to Mazurek & Zatoń," *Lethaia*, vol. 44, no. 4, pp. 369–372, 2011.

[95] L. Salvini-Plawen, *The Significance of the Placophora for Molluscan Phylogeny*, vol. 65, Venus, Elko, Nev, USA, 2006.

[96] K. G. Wingstrand, "On the anatomy and relationships of recent Monoplacophora," *Galathea Report*, vol. 16, pp. 7–94, 1985.

[97] G. Giribet, "Bivalvia," in *Phylogeny and Evolution of the Mollusca*, W. F. Ponder and D. R. Lindberg, Eds., pp. 105–141, University of California Press, Berkeley, Calif, USA, 2008.

[98] S. W. Aktipis and G. Giribet, "A phylogeny of Vetigastropoda and other "archaeogastropods": re-organizing old gastropod clades," *Invertebrate Biology*, vol. 129, no. 3, pp. 220–240, 2010.

[99] S. W. Aktipis and G. Giribet, "Testing relationships among the vetigastropod taxa: a molecular approach," *Journal of Molluscan Studies*, vol. 78, no. 1, pp. 12–27, 2012.

[100] G. Haszprunar, "Monoplacophora (Tryblidia)," in *Phylogeny and Evolution of the Mollusca*, W. F. Ponder and D. R. Lindberg, Eds., pp. 97–104, University of California Press, Berkeley, Calif, USA, 2008.

[101] L. R. L. Simone, "Comparative morphology among representatives of main taxa of Scaphopoda and basal protobranch Bivalvia (Mollusca)," *Papeis Avulsos de Zoologia*, vol. 49, no. 32, pp. 405–457, 2009.

[102] J. Pojeta and B. Runnegar, "The paleontology of rostroconch mollusks and the early history of the phylum Mollusca," *U.S. Geological Survey Professional Paper*, vol. 986, pp. 1–88, 1976.

[103] A. Wanninger and G. Haszprunar, "The expression of an engrailed protein during embryonic shell formation of the tusk-shell, *Antalis entalis* (Mollusca, Scaphopoda)," *Evolution and Development*, vol. 3, no. 5, pp. 312–321, 2001.

[104] S.-I. Yokobori, T. Iseto, S. Asakawa et al., "Complete nucleotide sequences of mitochondrial genomes of two solitary entoprocts, *Loxocorone allax* and *Loxosomella aloxiata*: implications for lophotrochozoan phylogeny," *Molecular Phylogenetics and Evolution*, vol. 47, no. 2, pp. 612–628, 2008.

[105] I. Stöger and M. Schrödl, "Mitogenomics does not resolve deep molluscan relationships (yet?)," *Molecular Phylogenetics and Evolution*, vol. 69, no. 2, pp. 376–392, 2013.

[106] S. Faller, B. H. Rothe, C. Todt, A. Schmidt-Rhaesa, and R. Loesel, "Comparative neuroanatomy of Caudofoveata, Solenogastres, Polyplacophora, and Scaphopoda (Mollusca) and its phylogenetic implications," *Zoomorphology*, pp. 1–22, 2012.

[107] M. P. Pelseneer, "Sur le quatrième orifice palléal des Pélécypodes," *Comptes Rendus Hebdonnaires des Séances de l'Académie des Sciences*, vol. 110, pp. 154–156, 1890.

[108] A. Wanninger and G. Haszprunar, "Chiton myogenesis: perspectives for the development and evolution of larval and adult muscle systems in molluscs," *Journal of Morphology*, vol. 251, no. 2, pp. 103–113, 2002.

[109] M. R. Smith, "Mouthparts of the Burgess Shale fossils *Odontogriphus* and *Wiwaxia*: implications for the ancestral molluscan radula," *Proceedings of the Royal Society B*, vol. 279, no. 1745, pp. 4287–4295, 2012.

[110] L. Salvini-Plawen, *Origin, Phylogeny and Classification of the Phylum Mollusca*, vol. 9, Iberus, Madrid, Spain, 1990.

[111] B. M. Wheeler, A. M. Heimberg, V. N. Moy et al., "The deep evolution of metazoan microRNAs," *Evolution and Development*, vol. 11, no. 1, pp. 50–68, 2009.

[112] R. D. Hoare, "Considerations on Paleozoic Polyplacophora including the description of *Plasiochiton curiosus* n. gen. and sp.," *American Malacological Bulletin*, vol. 15, no. 2, pp. 131–137, 2000.

[113] A. Y. Ivantsov, "Paleontological evidence for the supposed Precambrian occurrence of mollusks," *Paleontological Journal*, vol. 44, no. 12, pp. 1552–1559, 2010.

[114] A. Seilacher and J. W. Hagadorn, "Early molluscan evolution: evidence from the trace fossil record," *Palaios*, vol. 25, no. 9, pp. 565–575, 2010.

[115] M. A. Fedonkin, A. Simonetta, and A. Y. Ivantsov, "New data on *Kimberella*, the Vendian mollusc-like organism (White Sea region, Russia): palaeoecological and evolutionary implications," *Geological Society Special Publication*, no. 286, pp. 157–179, 2007.

[116] A. Y. Ivantsov, "New reconstruction of *Kimberella*, problematic Vendian metazoan," *Paleontological Journal*, vol. 43, no. 6, pp. 601–611, 2009.

[117] E. L. Yochelson, "Concerning the concept of extinct classes of Mollusca: or what may/may not be a class of mollusks," *American Malacological Bulletin*, vol. 15, no. 2, pp. 195–202, 2000.

[118] G. Haszprunar and B. Ruthensteiner, "Monoplacophora (Tryblidia)-some unanswered questions," *American Malacological Bulletin*, vol. 31, no. 1, pp. 189–194, 2013.

[119] K. Kocot, "Recent advances and unanswered questions in deep molluscan phylogenetics," *American Malacological Bulletin*, vol. 31, no. 1, pp. 195–208, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 353956, 14 pages
http://dx.doi.org/10.1155/2013/353956

Research Article

Speciation in *Thaparocleidus* (Monogenea: Dactylogyridae) Parasitizing Asian Pangasiid Catfishes

**Andrea Šimková,[1] Celine Serbielle,[2] Antoine Pariselle,[3]
Maarten P. M. Vanhove,[1,4] and Serge Morand[3]**

[1] *Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic*
[2] *Institut de Recherche sur la Biologie de l'Insecte, Faculté des Sciences et Techniques, UMR CNRS 6035, 37200 Tours, France*
[3] *Institut des Sciences de l'Evolution, IRD-CNRS-UM2, Université Montpellier 2, CC065, 34095 Montpellier Cedex 05, France*
[4] *Laboratory of Biodiversity and Evolutionary Genomics, University of Leuven, Charles Deberiotstraat 32, 3000 Leuven, Belgium*

Correspondence should be addressed to Andrea Šimková; simkova@sci.muni.cz

Received 1 March 2013; Revised 9 September 2013; Accepted 14 September 2013

Academic Editor: William Piel

The phylogeny of monogeneans of the genus *Thaparocleidus* that parasitize the gills of Pangasiidae in Borneo and Sumatra was inferred from molecular data to investigate parasite speciation. The phylogeny of the Pangasiidae was also reconstructed in order to investigate host-parasite coevolutionary history. The monophyly of *Thaparocleidus* parasitizing Pangasiidae was confirmed. Low intraspecies molecular variability was observed in three *Thaparocleidus* species collected from geographically distant localities. However, a high intraspecies molecular variability was observed in two *Thaparocleidus* species suggesting that these species represent a complex of species highly similar in morphology. Distance-based and tree-based methods revealed a significant global fit between parasite and host phylogenies. Parasite duplication (i.e., intrahost speciation) was recognized as the most common event in *Thaparocleidus*, while the numbers of cospeciation and host switches were lower and similar to each other. When collapsing nodes correspond to duplication cases, our results suggest host switches in the *Thaparocleidus*-Pangasiidae system precluding congruence between host and parasite trees. We found that the morphometric variability of the parasite attachment organ is not linked to phylogeny, suggesting that the attachment organ is under adaptive constraint. We showed that haptor morphometry is linked to host specificity, whereby nonspecific parasites display higher morphometric variability than specialists.

1. Introduction

The speciation of free-living organisms is thought to be caused by two main mechanisms: allopatric speciation, which results from reproductive isolation due to extrinsic factors such as geographical barriers [1], and nonallopatric speciation such as sympatric speciation, which requires intrinsic barriers for reproductive isolation [2, 3]. In parasites, speciation is usually linked to the evolutionary history of their host species, with host speciation inducing parasite speciation when each incipient host species has inherited parasite populations that subsequently diverge from a common ancestor [4]. Therefore, the allopatric speciation of parasites may occur when extrinsic barriers prevent parasite reproduction among isolated host populations. For example, this can occur when the host species are geographically isolated. On the other hand, sympatric speciation (geographic sympatry, within host sympatry or within microhabitat sympatry) can occur when the isolation of parasite populations is maintained by intrinsic barriers [5] and is therefore independent of host speciation events. Sympatric speciation could explain a large part of parasite diversity [4]. According to Kunz [6], sympatric speciation is more likely to occur in parasites than in free-living organisms, considering that the isolation of parasite populations seems to be accomplished more easily than in free-living organisms. The isolation processes and intrinsic barriers among parasite populations, such as host choice when the parasite shows a local host preference or mate choice when mating between two parasites is impossible, most likely will lead to parasite sympatric speciation [5]. Parasite sympatric speciation may occur within a single host species; that is, a parasite lineage has evolved within a single host species without any isolation of host populations. The key assumption here is that congeneric parasite species on the same host are sister species and that their occurrence is the result of one or more events of intrahost speciation [7].

Cophylogenetic studies comparing the evolutionary histories of parasites and their associated hosts may help us to further explore parasite speciation mechanisms [8]. Congruence of host and parasite phylogenies is considered evidence of cospeciation, that is, the concurrent speciation of both associated partners [8]. However, congruent trees are not always linked to cospeciation (see [9]). Incongruent phylogenies are often explained by host-switching events or parasite duplication. However, testing for cospeciation or codivergence (i.e., simultaneous speciation or divergence of host and parasite lineages, while cospeciation is a special kind of codivergence in which the end products of the divergence process are considered separate species) requires the combination of distance-based, tree-based, and data-based (these methods are used to determine the cause of topological incongruence between host and parasite trees) cophylogenetic methods [10]. Indeed, discriminating between trees that are concordant as a result of codivergence and trees that are concordant for reasons unrelated to codivergence necessitates the comparison of topological similarities between not only host and parasite trees but also timing of events [10].

Speciation in parasites has been mostly explained by their life-history traits, such as host specificity. Generally, a parasite living on/in one host species is considered a specialist, and a parasite living on/in at least two species is considered a generalist [11]. Brooks and McLennan [4] hypothesized that the chance of colonizing new host species, that is, host switching, and the subsequent speciation are inversely related to the degree of host specificity, which supposes that cospeciation and intrahost speciation are more frequent in parasites having a narrow host range. For example, in the highly host specific chewing lice parasitizing pocket gophers, cospeciation was found to be the main speciation event [12, 13].

Monogeneans, a group of mostly ectoparasitic flatworms predominantly found in fishes, seem to be an ideal model for investigating parasite diversification for at least three reasons. First, monogeneans are a highly diverse parasite group in terms of species richness [14]. Second, many monogenean species tend to be host specific, that is, infecting only one or a few host species [15], and also niche specific, that is, restricted to a particular habitat within the host species [7, 16]. Third, monogeneans are parasites with a direct life cycle (only one host species is involved in their life cycle), which may simplify the analyses of host-parasite associations compared to endoparasites with a complex life cycle (including intermediate and definitive host species throughout various stages of the life cycle). To date, several studies have investigated the speciation and diversification of different congeneric monogenean species. These studies do not show strong patterns of cospeciation, despite the high host specificity of monogenean parasites, but they suggest that monogeneans mostly diversify either through host switching [17–20] or by intrahost speciation [7, 21]. Host specificity, varying between the different monogenean models investigated, is considered to be an important parasite trait involved in monogenean speciation processes.

Monogeneans possess a posterior attachment organ, called a haptor, which is supposed to be linked to both specialization and adaptation [22]. Morand et al. [23] hypothesized that a link between morphological and phylogenetic distances may reflect a nonadaptive trend due to a high phylogenetic inertia, with sister species possessing similar haptors because they have inherited them from a common ancestor. Conversely, a link between the morphometrics of the monogenean haptor and host specificity may reveal a potential adaptation. Indeed, a higher variability of the attachment organ was shown in generalists compared to specialists in two groups of monogenean species [24, 25].

Thaparocleidus (Dactylogyridae, Ancylodiscoidinae) are gill monogeneans in siluriform fishes [26]. Nonmonophyly of *Thaparocleidus* was shown by Wu et al. [27]. To date, 43 *Thaparocleidus* species have been described from 18 species of Pangasiidae ([28] and references therein). These monogenean species are highly host specific: most of them are restricted to one host species. Pangasiid fishes are distributed in the main rivers and estuaries of South East Asia and occasionally in the sea [29]. The speciation processes involved in their diversification seem to be closely related to tectonic events that have formed the current river network in South East Asia as well as to sea level changes [30, 31]. Within this historical-biogeographical framework, we hypothesized that the speciation and diversification of specific *Thaparocleidus* should be closely related to the processes of Pangasiidae diversification. Moreover, we want to compare the diversification of *Thaparocleidus* species infecting the same host species to the processes observed in another well-studied freshwater dactylogyrid monogenean, *Dactylogyrus*. We hypothesize that, similar to *Dactylogyrus* parasitizing cyprinid fishes [7], intrahost speciation is an important speciation mechanism for *Thaparocleidus*.

The aim of this study was to use molecular phylogenetic reconstruction of *Thaparocleidus* parasitizing Pangasiidae in combination with cophylogenetic analyses to investigate speciation and diversification in this monogenean genus. We also investigated whether the morphometry of the attachment organ is phylogenetically constrained and linked to host specificity.

2. Materials and Methods

2.1. Parasite Sampling. Monogeneans identified as belonging to 16 different *Thaparocleidus* species were collected from freshwater fish species in the Indonesian islands Borneo and Sumatra (South East Asia). The fish were bought from the market or directly from fishermen and identified following Roberts and Vidthayanon [29]. After dissection, their left gill arches were preserved in 70% ethanol. Species identification was based on the morphology of sclerotized parts of the attachment and reproductive organs following Lim [32] and Pariselle et al. [28, 33–36]. All *Thaparocleidus* species collected for this study were found on a single host species except for *T. caecus*, which was collected from *Pangasius nasutus* and *Pangasianodon hypophthalmus* (this host species has been introduced over the entire region of Southeast Asia for aquaculture purposes). However, host specificity in this study (Table 1) was evaluated at global level; that is, the data on host range of each analyzed *Thaparocleidus* species were retrieved from published studies [28, 33–37]. Thus, parasite species were separated into two categories: specialist parasitizing a single host species and generalist parasitizing at least two different host species [38]. The geographical distribution of all fish species collected for this study was limited to Indonesia except for *Pangasius micronema*, which is distributed widely in South East Asia. The localities of the collected fish are given in Table 1.

2.2. DNA Extraction, Amplification, and Sequencing. DNA extraction was performed by the chelex method. Each monogenean specimen was disrupted in a 5% chelex solution with proteinase K (0.12 mg/mL). Partial 18S rDNA and entire ITS1 region were amplified in one round using the primers S1 (5′-ATTCCGATAACGAACGAGACT-3′) [39] and IR8 (5′-GCTAGCTGCGTTCTTCATCGA-3′) that anneal to the 18S and 5.8S rDNA genes, respectively [40]. Each amplification reaction was performed in a final volume of 25 μL containing 1.5 units of *Taq* polymerase, 1X buffer containing $MgCl_2$, 2.5 mM of each dNTP, 0.4 μM of each primer, and 5 μL of DNA. PCR was carried out using the following steps: 4 min at 95°C followed by 40 cycles of 1 min at 92°C, 1 min at 55°C, and 1 min 30 s at 72°C and 10 min of final elongation at 72°C. The PCR products were checked in a 1% agarose gel. The PCR products were purified by a ExoSAP-IT kit (USB) and were directly sequenced using the PCR primers. Sequencing was carried out using an ABI Prism BigDye Terminator Cycle Sequencing kit (Applied Biosystems) and electrophoresis was performed on an automated sequencer (MegaBace 500). New sequences were deposited in GenBank (see Table 1 for the accession numbers).

2.3. Phylogenetic Analyses. DNA sequences were aligned using ClustalW multiple alignment [41] in BioEdit Sequence Alignment Editor v.7.0.9 [42]. As the alignment for these closely related species was straightforward, gaps were included in the sequence alignment subjected to the phylogenetic analyses. Firstly, phylogenetic analyses were performed using partial 18S rDNA sequences, to allow inclusion of outgroups, in order to be able to subsequently root *Thaparocleidus* tree reconstructions. Four species, belonging to Ancylodiscoidinae and infecting nonpangasiid siluriforms, were used as outgroup to root the phylogeny of *Thaparocleidus* species parasitizing Pangasiidae: *Thaparocleidus siluri* (AJ490164), *T. vistulensis* (AJ490165) (both of them are monogenean parasites of the European silurid catfish species *Silurus glanis*), and two monogenean species from catfishes collected in West Africa, *Schilbetrema* sp. (HG491495) from the schilbeid catfish *Schilbe intermedius* and *Quadriancanthus* sp. (HG491496) from the airbreathing clariid *Heterobranchus bidorsalis*. As mentioned above, *Thaparocleidus* is not monophyletic; that is, *Thaparocleidus* parasitizing Siluridae and *Thaparocleidus* parasitizing Pangasiidae comprise different clades. Therefore, Wu et al. [27] proposed taxonomic revision of the species recently included in *Thaparocleidus*. Hence, *T. siluri* and *T. vistulensis* may be included as outgroup taxa. Subsequent phylogenetic analyses were performed using a concatenated dataset of partial 18S rDNA and ITS1 including only sequences of *Thaparocleidus* species parasitizing Pangasiidae. Intraspecific variability was explored using uncorrected p-distances (i.e., calculating the proportions of different nucleotide sites).

Phylogenetic analyses using minimum evolution (ME), maximum parsimony (MP), and maximum likelihood (ML) were performed in PAUP∗4b10 [43]. The Bayesian analyses were conducted using MrBayes 3.1 [44]. ModelTest [45] was used to select the optimal evolutionary model for each dataset, based on hierarchical likelihood ratio tests. The selected model was applied in the ME, ML, and BI tree reconstructions. ME analyses were performed using a heuristic search with a distance optimality criterion [46]. The search for the best ML tree was done via a heuristic search using the tree bisection reconnection branch-swapping algorithm (TBR). MP analyses were performed using a heuristic search algorithm with a stepwise random addition sequence running on unweighted informative characters and TBR branch swapping. The degree of "tree-likeness" in data under a parsimony model for character change was measured by consistency index (CI) and retention index (RI). Sequence alignment and trees were deposited to TreeBase, a database of phylogenetic knowledge, and are available at http://purl.org/phylo/treebase/phylows/study/TB2:S14752. Support values for internal nodes were estimated using a bootstrap resampling procedure with 1000 replicates [47]. BI analyses were performed using four Monte Carlo Markov Chains (MCMC) running on a given number of generations (starting at 10,000 and increasing until standard deviation was below 0.01), with trees being sampled every 100 generations. Log likelihoods of the saved trees were graphically inspected, and all trees before stationary were discarded as "burn-in." Two replicates were conducted for the Bayesian MCMC runs. The posterior probabilities for internal nodes were determined for all trees left in the plateau phase with the best likelihood scores. In accordance with Wahlberg et al. [48] and Yang et al. [49], clade support in phylogenetic trees indicated by bootstrap values (BP) or posterior probabilities (PP) was considered as follows: weak support 50–63%/0.5–0.69, moderate support 64–75%/0.7–0.84, good support 76–88%/0.85–0.94, and strong support 89–100%/0.95–1.00.

The sequences of cytochrome *b* obtained from Pouyaud et al. [30] were used for the phylogenetic reconstruction of Pangasiidae investigated in our study. Two Asian siluriforms, *Laides hexanema* and *Pseudeutropius brachypopterus*, belonging to Schilbeidae were used as outgroup [30]. Nucleotide sequences were aligned and then translated to amino acid sequences in MEGA v. 3.1 [50]. The best appropriate model of protein evolution was selected in ProtTest v. 2.4 [51] using the Akaike information criterion. The trees were reconstructed

Table 1: List of parasite species including their specificity (S: specialist, G: generalist), host species from which the parasite species was sequenced, and localities of collection and accession number. Host specificity was delimited using published records (see Section 2.)

Fish species	*Thaparocleidus* species	Location	Accession number
Pangasius nasutus	*T. caecus 1* (G)	Borneo	FJ493153
	T. alatus 2 (S)	Batang Hari River (Sumatra)	FJ493156
	T. citreum	Musi River (Sumatra)	FJ493145
	T. alatus 1 (S)	Musi River (Sumatra)	FJ493146
Pangasius micronema	*T. durandi 1* (S)	Borneo	FJ493151
	T. durandi 2 (S)	Batang Hari River (Sumatra)	FJ493162
	T. rukyanii 2 (S)	Batang Hari River (Sumatra)	FJ493163
	T. tacitus (S)	Batang Hari River (Sumatra)	FJ493161
	T. lebrunae (G)	Batang Hari River (Sumatra)	FJ493165
	T. summagracilis 2 (S)	Batang Hari River (Sumatra)	FJ493164
	T. sinepinae (G)	Musi River (Sumatra)	FJ493147
	T. rukyanii 1 (S)	Musi River (Sumatra)	FJ493148
	T. summagracilis 1 (S)	Musi River (Sumatra)	FJ493149
Pangasius djambal	*T. komarudini* (G)	Batang Hari River (Sumatra)	FJ493154
	T. combesi (G)	Batang Hari River (Sumatra)	FJ493155
Pangasius polyuranodon	*T. crassipenis* (S)	Batang Hari River (Sumatra)	FJ493157
	T. levangi (S)	Batang Hari River (Sumatra)	FJ493158
	T. legendrei (S)	Batang Hari River (Sumatra)	FJ493160
	T. turbinatio (G)	Batang Hari River (Sumatra)	FJ493159
Pangasianodon hypophthalmus	*T. caecus 2* (G)	Borneo	FJ493152
	T. siamensis (S)	Borneo	FJ493150

using both fast and slow strategies. Using the fast strategy, the tree with BIONJ topology and branch length optimized under the best-fit model was obtained. Using the slow strategy, the best ML tree with both branch length and topology optimized under the best-fit model was obtained. A nonparametric bootstrap was calculated in PhyML 3.0 using 1000 replicates [52].

2.4. Cophylogenetic Analyses. Two methods of coevolutionary analyses were applied: a distance-based method called ParaFit [53] implemented in CopyCat [54] and a tree-based method implemented in Jane 4.0 [55]. A tanglegram representing the host-parasite associations was reconstructed in TreeMap 1.0 [56].

Distance-based methods use host and parasite distance matrices and host-parasite associations to determine whether hosts and parasites are randomly associated. The global fit between host and parasite trees is computed and tested by randomizing individual host-parasite associations. ParaFit was also used to test whether a particular host-parasite association contributed to this global fit. The permutational tests of significance were calculated using 999 permutations. This method was shown to be useful for studying host-parasite cophylogeny using congeneric monogeneans parasitizing fish [17, 21].

Tree-based methods use tree topologies to assess the fit between host and parasite phylogenies. These methods are aimed at the reconstruction of shared evolutionary history between hosts and parasites with the smallest number of hypothesized historical events (this is expressed by "cost"). Each event has an attributed cost and the reconstruction with the lowest global cost is searched. Jane supports multihost parasites and multiparasite hosts. In addition, Jane 4 supports polytomies. In our study, we applied seven models with different event cost schemes. TreeMap and TreeFitter [57] models classically used four types of coevolutionary events, that is, cospeciation, duplication, host switching, and sorting event; however, Jane applies a fifth type of coevolutionary events called "failure to diverge" (host speciation is not followed by parasite speciation, and the same parasite species occur on the new host species). The cost for "failure to diverge" was added in TreeMap and TreeFitter models following Mendlová et al. [21]. The cophylogenetic analyses were performed using the following genetic parameters: 1000 generations and 100 as a population size. Statistical tests were computed using 999 randomizations using the method of random parasite tree. The *Thaparocleidus* tree inferred from the analysis of combined 18S rDNA and ITS1 data and the Pangasiidae tree inferred from cytochrome *b* were used in the cophylogenetic analyses.

2.5. Parasite Morphometry. A total of 9 specialist species and 6 generalist species sampled in this study and identified on the basis of morphology were measured (morphometric data for *T. citreum* were not available; see Table 1). Morphometric measurements were taken from 10 individuals from each parasite species. 20 variables describing the haptor were taken into account for statistical analyses. The variables used are represented in Figure 1. The Mahalanobis distances between species were calculated for haptor morphometrics.

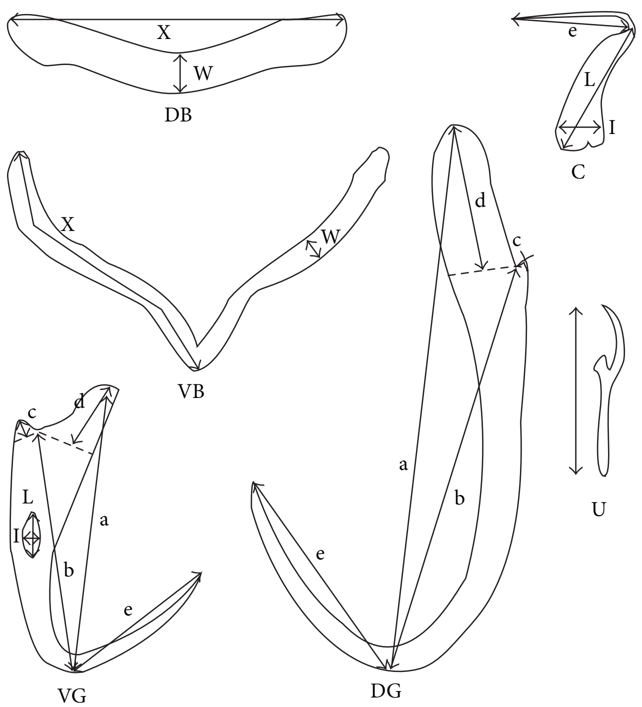


Figure 1: Measurements of the sclerotized part of haptor in *Thaparocleidus* (DB: dorsal transversal bar, VB: ventral transversal bar, VG: ventral gripus, DG: dorsal gripus, C: cuneus, and U: uncinulus, following [28]).

Phylogenetic distances expressed as patristic distances were calculated between species in PAUP∗4b10 [43]. The existence of a correlation between morphometric distances and phylogenetic distances was tested using a Mantel test. A significant correlation indicated phylogenetic inertia.

Next, we performed a principal component analysis (PCA) on the measurements of the haptor. We extracted the first two principal components as they represented 55.6% and 13.4% of the total variance. We tested the effect of host specificity on parasite haptor morphometry by performing a nonparametric Wald-Wolfowitz test on the values of the first two axes between the generalists and specialists.

3. Results

3.1. Intraspecies Variability. Within-species variability was compared between individuals collected from different host species (*T. caecus* 1 from *Pangasius nasutus* and *T. caecus* 2 from *Pangasianodon hypophthalmus*). Pairwise comparisons showed a 10.07% variability in combined partial 18S and ITS1 sequences. When comparing conspecific *Thaparocleidus* individuals parasitizing the same host species but collected from different islands (Borneo and Sumatra), pairwise comparisons revealed a 0.31% variability between *T. durandi* 1 (from *Pangasius micronema* collected in Borneo) and *T. durandi* 2 (from *Pangasius micronema* collected in Sumatra). Pairwise comparisons between individuals parasitizing the host species but collected from different rivers in Sumatra revealed a 0.62% variability between *T. rukyanii* 1 collected from Musi River and *T. rukyanii* 2 collected from Batang Hari River and a 0.30% variability between *T. summagracilis* 1 from Musi River and *T. summagracilis* 2 from Batang Hari River. A variability of 5.79% was recorded between *T. alatus* 1 collected from Musi River and *T. alatus* 2 collected from Batang Hari River.

3.2. Phylogenetic Analyses Using 18S rDNA. All of the presented phylogenetic analyses show that two individuals of *T. alatus* found on *P. nasutus* from different isolated localities occupy different positions in the phylogenetic trees. Similarly, two individuals morphologically identified as *T. caecus* but parasitizing different host species, *Pangasius nasutus* and *Pangasianodon hypophthalmus*, both collected in Borneo, did not represent sister species. Rather, these species were included in different well-supported clades.

The *Thaparocleidus* species parasitizing fish species of Pangasiidae and four other species of Ancylodiscoidinae were included in the first phylogenetic analyses (see Section 2). The sequence alignment was comprised of 393 unambiguously alignable positions (including 7 positions with gaps), of which 80 were variable and 45 were parsimony informative. The K80+G model was selected as the best appropriate model by ModelTest, and the heterogeneity and substitution rate were approximated by a gamma distribution with shape parameter $\alpha = 0.2165$ and substitution rate matrix: A-C = 1.0000, A-G = 6.2288, A-T = 2.5937, C-G = 2.5937, C-T = 6.2288, and G-T = 1.0000. The MP analysis provided 54 equally parsimonious trees with 120 steps (CI = 0.758, RI = 0.803). The BI analysis was conducted by running MCMC for 300,000 generations. All phylogenetic analyses yielded a similar tree topology (the ML tree is shown in Figure 2 including BP for ML, MP, ME, and PP for BI analyses). Several terminal clades including *Thaparocleidus* of Pangasiidae with moderate, good, or strong support based on ML, MP, and ME analyses and good or strong support based on BI analyses were recognized from the phylogenetic reconstruction using 18S rDNA data. However, many phylogenetic relationships displayed low resolution or were unresolved using ML, MP, and ME analyses, and some of them were resolved only under BI (PP from 0.62 to 0.93). However, the fact that the values of PP tend to be higher than of BP is well known [58]. *Thaparocleidus* of Pangasiidae form a monophyletic group.

3.3. Phylogenetic Analyses Using Combined Data 18S rDNA and ITS1. The congruence of 18S and ITS1 data sets was tested using the partition homogeneity test implemented in PAUP∗4b10 [43]. No significant difference was found between 18S rDNA and ITS1 ($P = 0.084$). Therefore, the next analyses were performed using the concatenated dataset, including 21 sequences of *Thaparocleidus* parasites of Pangasiidae. The phylogenetic trees were oriented using the results obtained from the 18S rDNA analyses.

The sequence alignment was comprised of 666 unambiguously alignable positions (including 36 positions with gaps), of which 232 were variable and 172 were parsimony informative. The TVM+I+G model was selected by ModelTest including equal frequencies of nucleotide bases,

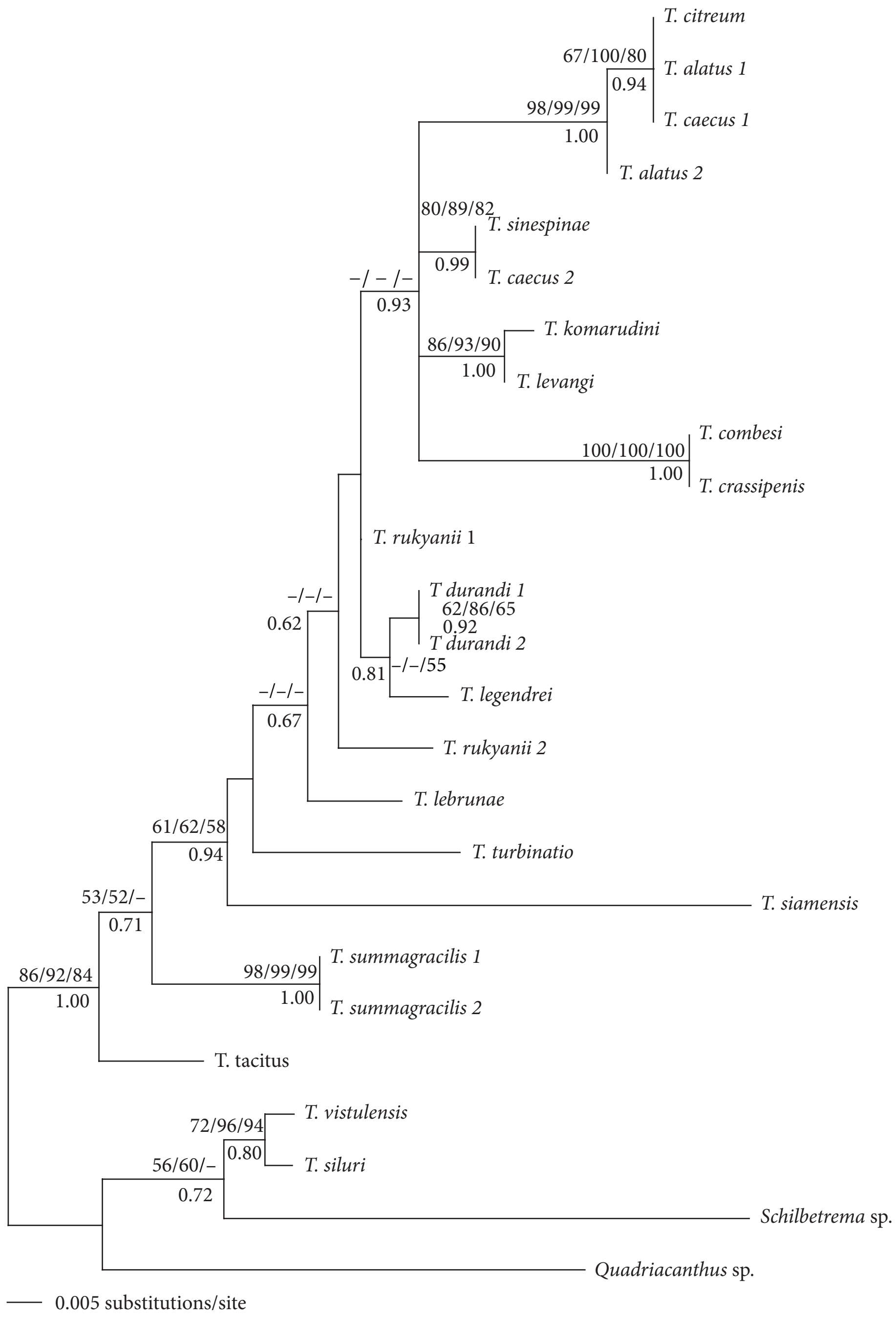


Figure 2: ML tree inferred from the analyses of partial 18S rDNA sequences of species belonging to Ancylodiscoidinae. Numbers above branches indicate bootstrap values resulting from ML/MP/ME analyses; numbers below branches indicate posterior probabilities resulting from BI analysis.

with a proportion of invariable sites pi = 0.3871, and the heterogeneity and substitution rate were approximated by a gamma distribution with shape parameter α = 0.6119, with the following substitution rate matrix: A-C = 0.8437, A-G = 5.6868, A-T = 3.0736, C-G = 0.8639, C-T = 5.6868, and G-T = 1.0000. The consensus tree obtained from the BI analysis is shown in Figure 3; the values of BP for ML, ME and MP analyses and PP for the BI analysis are included in this figure. The MP analysis provided 2 equally parsimonious trees with 493 steps (CI = 0.659, RI = 0.761). On the basis of the 18S rDNA analyses, *Thaparocleidus tacitus*, a parasite of *Pangasius micronema*, was used for rooting. All phylogenetic analyses yielded a similar tree topology; that is, five clades were recovered in all reconstructions (Figure 3). Three clades (group 1, group 2, and group 4) were strongly supported by BP or PP in all analyses, and two of them were weakly supported

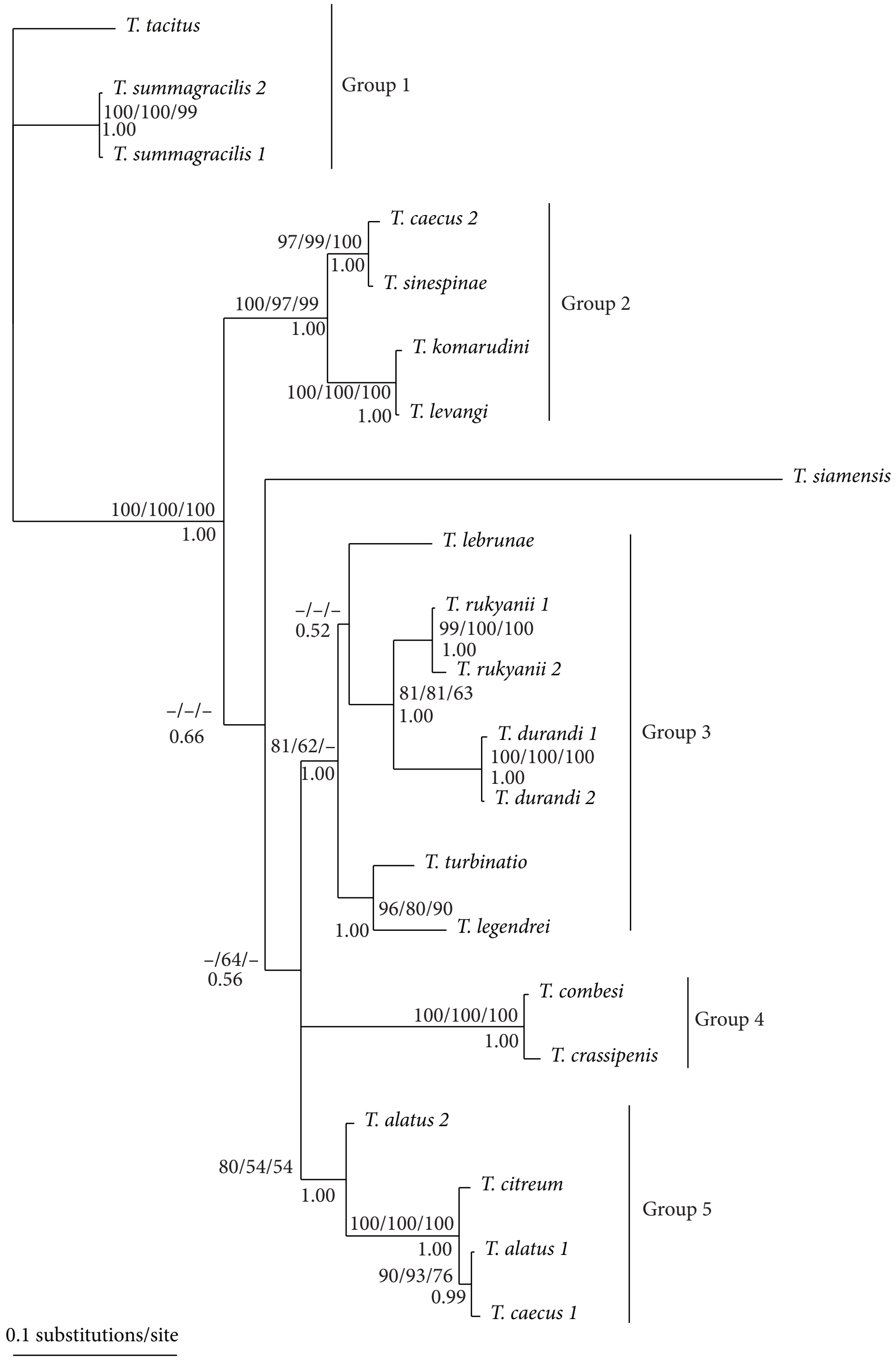


Figure 3: Bayesian topology for *Thaparocleidus* species of Pangasiidae based on combined data of partial 18S rDNA and ITS1 region. Numbers below branches indicate posterior probabilities resulting from BI analysis; numbers above branches indicate bootstrap values resulting from ML/MP/ME analyses.

by MP (group 3) or by MP and ME (group 5), well supported by ML, and strongly supported by BI analyses (group 3 and group 5). The position of *T. siamensis* was slightly variable using different methods, but this species was always included in a large strongly supported clade of *Thaparocleidus* species including group 2, group 3, group 4, and group 5 (Figure 3). The different position of *T. siamensis* in phylogenetic trees likely results from long branch attraction in the MP analysis and long branch repulsion in the ML analysis [59]. However, no effect of *T. siamensis* on the general topology of Asian *Thaparocleidus* was recognized; that is, an identical global topology was obtained when either excluding or including *T. siamensis*.

3.4. Fish Phylogeny. The sequence alignment of cytochrome *b* was comprised of 539 unambiguously alignable positions, of

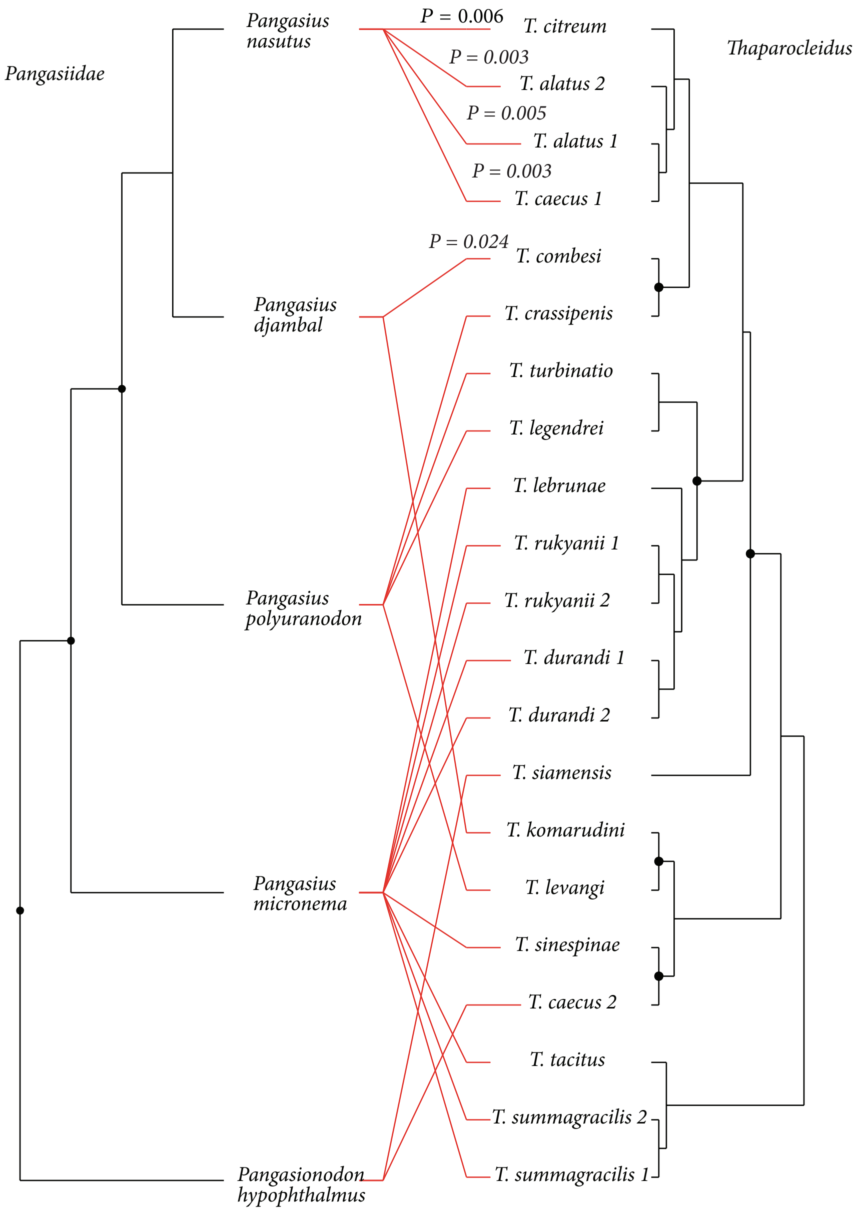


FIGURE 4: Tanglegram of *Thaparocleidus* and Pangasiidae species deduced from comparison of parasite tree inferred from combined data of 18S rDNA and ITS1 sequences and the fish tree obtained from cytochrome *b* analyses. P-values resulting from Parafit for significant host-parasite links are included.

which 197 were variable and 150 were parsimony informative. The amino acids alignment was comprised of 179 amino acids. MtMam+I was selected as the best appropriate model of protein evolution using both the slow and fast strategies in ProtTest with a proportion of invariable sites, pi = 0.822. The topology of the best ML tree is included in Figure 4. The phylogenetic relationships between the analyzed species of Pangasiidae were well resolved (all BP > 80) using both fast (BIONJ) and slow (ML) analyses and including *Laides hexanema* and *Pseudeutropius brachypopterus* as outgroup taxa. *Pangasionodon hypophthalmus* has a basal position to the group of the four *Pangasius* species that were analyzed.

3.5. Host-Parasite Associations. A tanglegram of the *Thaparocleidus* parasite species and their Pangasiidae fish hosts is shown in Figure 4. For *Thaparocleidus* species, the fully resolved ME topology was included (this tree is required to use TreeMap). Using ParaFit running in CopyCat, the overall cophylogenetic structure was highly significant ($P = 0.013$). The test computed for individual host-parasite links showed that five links out of 21 contributed significantly ($P < 0.05$) to this global fit (Figure 4).

We explored seven models with different event cost schemes (Table 2) previously applied for cophylogenetic studies using Jane, TreeMap, or TreeFitter programs. The

Table 2: Results of cophylogenetic analyses calculated in Jane 4 for *Thaparocleidus* parasites and Pangasiidae fish. The numbers of each event type necessary to reconcile host and parasite trees under different event cost schemes are shown. Event costs in the second column correspond to the following events: cospeciation, duplication, host switch, sorting event, and failure to diverge. The significant P values are shown in bold.

Model	Event costs	Total cost	Cospeciation	Duplication	Duplication & host switch	Sorting event	Failure to diverge	P value
Jane default model v. 4	1 1 1 1 1	26	4	11	5	1	0	**0.002**
Jane default model v. 3	0 1 1 2 1	23	4	11	5	1	0	**0.016**
TreeMap default model	0 1 1 1 1	22	4	11	5	1	0	**0.045**
TreeMap default model for building a jungle	0 2 1 1 1	37	5	10	5	2	0	0.270
TreeFitter default model	0 0 2 1 1	11	4	11	5	2	0	**0.005**
Host switch-adjusted TreeFitter model	0 0 1 1 1	6	4	11	5	1	0	**0.001**
Codivergence adjusted TreeFitter model	1 0 1 1 1	8	0	12	8	0	0	**0.003**

Note: the event cost schemes including cost for each evolutionary event are shown in the second column. Because it is assumed that host switch can only occur with duplication event, Jane 4 (unlike Jane 3, TreeMap, and TreeFitter) defined "duplication and host switch" instead of "host switch" with the default cost equal to 2 (i.e., cost of 1 for duplication and 1 for host switch is equivalent in Jane 4 to a cost of 1 for duplication and 2 for "duplication and host switch"). To avoid the misinterpretation of event cost schemes used in this study, in this table we retained the presentation using the classically applied event costs (i.e., cost for duplication and cost for host switch).

analyses revealed a significant global structure using all models except for one, assigning a higher cost to duplication than to other events (Table 2). The number of cospeciations and host switching events (always considered subsequent to duplication in Jane) was similar under all models except the codivergence-adjusted model, which assigned the same cost to cospeciation and host switch and no cost for duplication. Under all event cost settings, cophylogenetic reconciliation revealed that duplication (i.e., parasite speciation without corresponding host speciation) was the most frequent coevolutionary event (see Table 2). The tanglegram of *Thaparocleidus*-Pangasiidae associations demonstrated that intrahost duplication was mostly found in *P. micronema*, the fish species in which the highest diversity of *Thaparocleidus* species was recorded (see Table 1). However, as shown in this tanglegram, intrahost duplications were also reported in *P. nasutus* and *P. polyuranodon*. When the intrahost duplications were removed from reconstruction (i.e., after collapsing nodes in the parasite tree that correspond to duplications), no significant global structure ($P > 0.05$) was found using any of the models. The same number of cospeciation and host switch was inferred as in the analyses using all *Thaparocleidus* species.

3.6. Morphometric Variability of the Attachment Organ. No significant correlation between the Mahalanobis distances of the haptor and the patristic phylogenetic distances calculated between pairs of monogenean species was found (Mantel test, $F_{1,105} = 8.764$; $R^2 = 0.08$, $P = 0.14$). We performed a PCA on 20 morphometric variables of the haptor. The first two principal component axes were extracted (eigenvalue of the first axis = 11.12 with 55.6% of the total variance; the eigenvalue of the second axis = 2.68 with 13.40% of the total variance). We compared the values of the first two axes, PCA1 and PCA2, with host specificity (specific versus nonspecific) as a factor. Significant difference in haptor morphometry was found between the specific and nonspecific parasites using a Wald-Wolfowitz test ($Z = -6.064$, $P < 0.001$ for PC1 and $Z = -5.548$, $P < 0.001$ for PC2). Moreover, the nonspecific parasite species varied more in their morphometrics than the specific ones (F test, $P < 0.05$) (Figure 5).

4. Discussion

4.1. Thaparocleidus Species: Morphological versus Molecular Species Concept. In monogeneans, species identification is generally based on the morphology of two sclerotized organs, the attachment organ (the haptor) and the reproductive organ, including the copulatory piece and the vagina [60]. The morphology of the haptor is considered useful for parasite determination at the genus level, while the reproductive organ is more suitable for identification at the species level, probably because of its higher rate of change [27, 61, 62]. However, in some cases, identification is problematic at the generic as well as the specific level. Therefore, molecular identification is a helpful tool in resolving taxonomic problems in cases where the morphological "boundaries" among monogenean genera or species groups are ambiguous [27, 61, 62].

In our study, molecular phylogeny confirmed the monophyly of *Thaparocleidus* species parasitizing pangasiid hosts from Borneo and Sumatra. Wu et al. [27] showed that the species recently included in *Thaparocleidus* do not constitute a monophyletic group; that is, *Thaparocleidus* parasitizing Siluridae and *Thaparocleidus* parasitizing Pangasiidae form two divergent genetic lineages. Our results demonstrate that molecular data are not only helpful in recovering monophyly in monogeneans but also allow the discrimination of allegedly conspecific and morphologically identical species. In our study, intraspecies genetic variability was found between parasite individuals morphologically identified as belonging to the same species, which were collected from geographically distant locations, that is, for *T. summagracilis*, *T. durandi*, and *T. rukyanii*, each of them parasitizing a single host species collected from different Indonesian islands or two different rivers in Sumatra. However, intraspecies variability observed

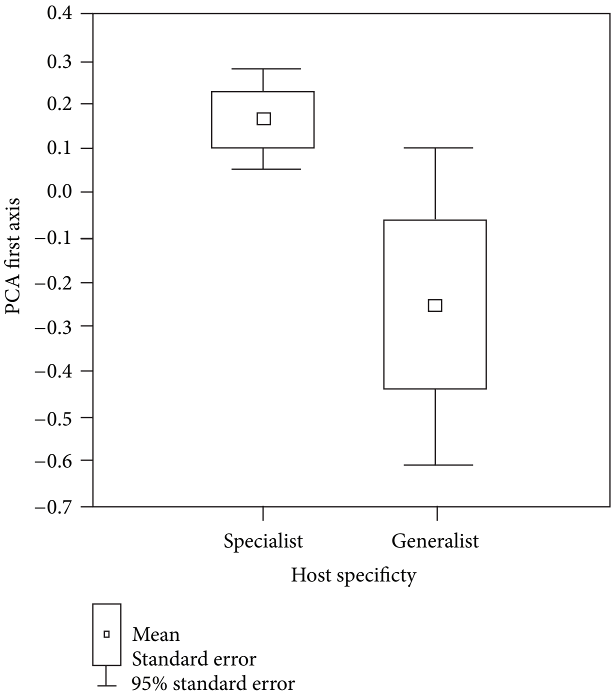


Figure 5: Morphometric variability of the parasite attachment organ estimated by the first axis of principal component analysis. The nonspecific parasites (i.e., generalists) differ significantly from specific parasites. The generalists show significant higher variance in haptor morphometry than specific parasites.

in *T. caecus* parasitizing two different host species and *T. alatus* parasitizing the same host species is very high, suggesting that each of these two species represents a potential complex of morphologically similar species. On the other hand, some morphologically well-identified *Thaparocleidus* species in our study showed a low level of molecular divergence. Similarly, consistently distinctive species separated by low genetic distances were also recognized within *Ligophorus* (also dactylogyrid gill monogeneans but parasitizing Mugilidae) ([63], Marchiori et al., unpublished). These are likely species complexes explained by recent rapid speciation and diversification.

The morphology of the male copulatory organ of *Thaparocleidus* was suggested as a key determinant for separating the clades recovered in molecular phylogenetic reconstruction [27]. A similar argument was used for monogeneans belonging to *Cichlidogyrus*, parasitizing Cichlidae, by Wu et al. [62]. However, Pouyaud et al. [61] and Vignon et al. [64] suggested that the morphology of the attachment organ is more suitable to infer phylogenetic relationships among major lineages in these cichlid monogeneans. The association between attachment organ morphology and (molecular based) phylogeny has also been documented for *Dactylogyrus*, a highly specific group of monogeneans of Cyprinidae [65]. However, species sharing the same morphological traits are not necessarily derived from a common ancestor. Moreover, the evolution of sclerotized organs is not neutral and seems to be under adaptive constraints in case of the attachment organ [23, 66]. In our study, the morphological variability of the haptor was not linked to phylogenetic distances, suggesting that the morphological variability of these sclerotized organs is not inherited from a common ancestor and may be under adaptive constraint.

4.2. Morphometric and Molecular Variability versus Host Specificity. Interspecific variability in ITS1 sequences and in haptor morphometry was previously shown in generalist monogeneans such as species belonging to *Dactylogyrus* and *Lamellodiscus* [24, 25]. Moreover, *Lamellodiscus* generalists have a higher intraspecies molecular variability and a higher variance of haptor morphometry than do specialists. In our study, we also found that the variance in haptor morphometry is higher in *Thaparocleidus* generalists than in specialists. Kaci-Chaouch et al. [25] proposed two alternative hypotheses to explain why the variance in haptor morphometry is higher in generalists than in specialists. Generalists exhibit a higher variance because (i) they use different host species representing a wide range of niches, which can exert different pressures on morphology, or (ii) they have a higher morphometric variability of the attachment organ which may allow parasites to colonize more host species. Therefore, morphometric variability can have a major impact on parasite speciation processes regardless of host speciation by restricting specialists within a particular host and habitat, thereby giving generalists the capability to have a larger host range and/or colonize several habitats.

4.3. Speciation and Diversification in Thaparocleidus. Several monophyletic groups, each of them including *Thaparocleidus* species parasitizing a single host species, were observed on the basis of molecular phylogenetic reconstruction. This strongly suggests the diversification of these monogeneans by within-host speciation. For example, the parasite species of *P. nasutus* are sister species resulting from several intrahost speciation events. In contrast with *P. nasutus*, which is restricted to the Borneo and Sumatra basins [29], *P. micronema* is largely distributed in Southeast Asia. In this host, *Thaparocleidus* species belong to three different phylogenetic lineages. Among these, two lineages (the first includes the species of group 1 and the next includes the species within group 3 in Figure 3) are formed by sister species which were the result of intrahost speciation. As the different *Thaparocleidus* lineages found in *P. micronema* are not closely related to each other, multiple colonization events are likely to have occurred within this host species. *Thaparocleidus* speciation has probably occurred independently in different host species, and different *Thaparocleidus* lineages could evolve in parallel within the same host.

The concept of sympatric speciation as an evolutionary diversification process remains controversial. According to Coyne [3] (see also [67]), there are four main requirements needed to prove sympatric speciation. The first is that the species must be largely or completely sympatric. In our case, if we consider the host species as a unit, all parasite species found within the same given host species are considered to be sympatric species. Secondly, these sympatric species must show reproductive isolation. It was demonstrated in other monogenean groups that congeneric species found in the same host species and occupying adjacent niches within the host (i.e., gill parts) differ in the morphology of their copulatory organs [68]. Thirdly, the sympatric species must be sister species, which is shown in our phylogenetic analyses where most *Thaparocleidus* species from a single host species form a monophyletic group. Fourthly and finally, the species did not seem to have undergone an allopatric diversification phase. However, a more intensive survey and analysis of the *Thaparocleidus* species will be needed to justify this assumption.

Sympatric speciation can usually be encountered when closely related species live in isolated island-like habitats. Host species are considered as islands for parasites, and in view of the parasite life cycle we can expect parasite speciation to occur at a higher rate than host speciation. This faster pace of evolution also favours intrahost speciation. Sympatric speciation in monogeneans has previously been observed in *Dactylogyrus* species parasitizing cyprinid fish in Central Europe [7]. In this system, the authors suggested that parasite diversification can be explained by sympatric speciation events (i.e., intrahost speciation). Intrahost or sympatric speciation is linked to reproductive isolation of sympatric parasite populations. Different mechanisms have been proposed to explain the reproductive isolation of parasites such as habitat selection (preferred niches are its consequence) or mate choice [22]. On the basis of tree-based cophylogenetic analysis using different event cost schemes, sympatric speciation (i.e., intrahost speciation) also appears as the dominant coevolutionary event involved in *Thaparocleidus* diversification. However, our study also evidenced some host switches in the *Thaparocleidus*-Pangasiidae system. Giraud et al. [69] showed that certain pathogen life traits (i.e., production of numerous propagules, gene exchange occurring within hosts, linkage of traits experiencing selection, and strong selection imposed by the hosts) likely render them prone to rapid ecological speciation by host shifts (i.e., speciation by specialization onto a novel host). As many such life traits have been recognized also for monogenean fish parasites, this may explain the evidence of host switches documented by cophylogenetic analyses in monogenean parasites.

The overall congruence between the *Thaparocleidus* and Pangasiidae phylogenies was statistically significant according to topology-based and distance-based methods. Using a tree-based method, a nonsignificant global fit between the phylogenies of *Thaparocleidus* parasites and pangasiid hosts was found only using the model with a higher cost for duplication than for host switch. When considering the fact that duplication is the most numerous coevolutionary event in congeneric monogeneans parasitizing freshwater fish hosts (e.g., [7, 21]), duplication is probably not so costly as host switch is (because many monogenean species are host specific; thus, they have a limited ability for dispersal to other host species). Therefore, the model with the cost of 2 for duplication and 1 for host switch seems to be less realistic.

While our results indicate the congruence between *Thaparocleidus* and Pangasiidae phylogenies, these results should be interpreted carefully. First, the number of the investigated host species could be small to detect potential cospeciation events even though we included all principal and commonly occurring pangasiid species from Indonesian islands. In addition to the tree-based and distance-based methods, the estimates of divergence times in host and parasite lineages are considered as critical components of cophylogenetic studies to detect cospeciation [10]. However, no timing information is available for our analysed model. Huyse and Volckaert [19] on the basis of tree-based methods found an overall fit between the phylogenies of *Gyrodactylus* parasites (viviparous monogeneans) and goby hosts, but an absolute timing of speciation events in hosts and parasites ruled out the possibility of synchronous speciation. Thus, they proposed that phylogenetically conserved host switching mimics the phylogenetic signature of cospeciation.

Bentz et al. [70] studied the evolution of African *Polystoma*, endoparasitic monogeneans of neobatrachian hosts, and proposed that distinctive larval behaviour of polystomes engenders isolation between parasite populations which precludes sympatric speciations, and thus cospeciation is another factor of diversification of *Polystoma* in the African continent. However, the majority of previous cophylogenetic studies on congeneric monogeneans parasitizing fish did not report cospeciation [7, 17, 18, 21]. De Vienne et al. [71] in their review study showed that convincing cases of cospeciation in host-parasite and host-mutualist associations are very rare and host switches may be the dominant mode of speciation over cospeciation. In addition, they suggested that cophylogenetic methods overestimate the occurrence of cospeciation. Different processes may generate apparent cospeciation [9, 71]. Our study indicates that such apparent cospeciation in *Thaparocleidus*-Pangasiidae may be generated by intrahost duplications and/or also caused by host-switching events. The sympatric occurrence of some pangasiid species may more likely support the evidence of host switches than cospeciation in *Thaparocleidus* diversification; for instance, *P. djambal* and *P. polyuranodon* live in the same basin, which could facilitate host switching.

5. Conclusion

Our study of closely related parasites within a relatively small geographical area emphasizes particularly that intrahost speciation is the dominant coevolutionary event in *Thaparocleidus* species diversification favored by high specificity. Our study may indicate that host switches rather than cospeciation play a more substantial role in *Thaparocleidus* diversification. However, Pangasiidae speciation is closely related to tectonic events and the variation of sea levels [31, 44]; we then expected a similar pattern in parasite evolution. Therefore, to infer a formal conclusion on the role of cospeciation and host switching for *Thaparocleidus* diversification, we need to study these monogenean species on a broader geographical scale also including additional host species. Our study indicates that the morphological variability of attachment organ in *Thaparocleidus* parasites is not inherited from a common ancestor and could be potentially under adaptive constraint.

Acknowledgment

This study was supported by the Czech Science Foundation, Project no. P505/12/G112 (ECIP).

References

[1] J. A. Coyne and H. A. Orr, "The evolutionary genetics of speciation," *Philosophical Transactions of the Royal Society B*, vol. 353, no. 1366, pp. 287–305, 1998.

[2] S. Via, "Sympatric speciation in animals: the ugly duckling grows up," *Trends in Ecology and Evolution*, vol. 16, no. 7, pp. 381–390, 2001.

[3] J. A. Coyne, "Sympatric speciation," *Current Biology*, vol. 17, no. 18, pp. R787–R788, 2007.

[4] D. R. Brooks and D. A. McLennan, *Phylogeny, Ecology and Behavior: A Research Program in Comparative Biology*, The University of Chicago Press, London, UK, 1991.

[5] K. D. McCoy, "Sympatric speciation in parasites—what is sympatry?" *Trends in Parasitology*, vol. 19, no. 9, pp. 400–404, 2003.

[6] W. Kunz, "When is a parasite species a species?" *Trends in Parasitology*, vol. 18, no. 3, pp. 121–124, 2002.

[7] A. Šimková, S. Morand, E. Jobet, M. Gelnar, and O. Verneau, "Molecular phylogeny of congeneric monogenean parasites (*Dactylogyrus*): a case of intrahost speciation," *Evolution*, vol. 58, no. 5, pp. 1001–1018, 2004.

[8] R. D. M. Page, *Tangled Trees: Phylogeny, Cospeciation and Coevolution*, University of Chicago Press, Chicago, Ill, USA, 2003.

[9] D. M. de Vienne, T. Giraud, and J. A. Shykoff, "When can host shifts produce congruent host and parasite phylogenies? A simulation approach," *Journal of Evolutionary Biology*, vol. 20, no. 4, pp. 1428–1438, 2007.

[10] J. E. Light and M. S. Hafner, "Codivergence in heteromyid rodents (Rodentia: Heteromyidae) and their sucking lice of the genus *Fahrenholzia* (Phthiraptera: Anoplura)," *Systematic Biology*, vol. 57, no. 3, pp. 449–465, 2008.

[11] I. Humphery-Smith, "The evolution of phylogenetic specificity among parasitic organisms," *Parasitology Today*, vol. 5, no. 12, pp. 385–387, 1989.

[12] M. S. Hafner and S. A. Nadler, "Phylogenetic trees support the coevolution of parasites and their hosts," *Nature*, vol. 332, no. 6161, pp. 258–259, 1988.

[13] M. S. Hafner, P. D. Sudman, F. X. Villablanca, T. A. Spradling, J. W. Demastes, and S. A. Nadler, "Disparate rates of molecular evolution in cospeciating hosts and parasites," *Science*, vol. 265, no. 5175, pp. 1087–1090, 1994.

[14] L. S. Roberts and J. J. Javony, *Foundations of Parasitology*, William C. Brown, Dubuque, Iowa, USA, 1996.

[15] R. Poulin, "Determinants of host-specificity in parasites of freshwater fishes," *International Journal for Parasitology*, vol. 22, no. 6, pp. 753–758, 1992.

[16] A. Šimková, Y. Desdevises, M. Gelnar, and S. Morand, "Co-existence of nine gill ectoparasites (*Dactylogyrus*: Monogenea) parasitising the roach (*Rutilus rutilus* L.): history and present ecology," *International Journal for Parasitology*, vol. 30, no. 10, pp. 1077–1088, 2000.

[17] Y. Desdevises, S. Morand, O. Jousson, and P. Legendre, "Coevolution between *Lamellodiscus* (Monogenea: Diplectanidae) and Sparidae (Teleostei): the study of a complex host-parasite system," *Evolution*, vol. 56, no. 12, pp. 2459–2471, 2002.

[18] M. S. Zietara and J. Lumme, "Speciation by host switch and adaptive radiation in a fish parasite genus *Gyrodactylus* (Monogenea, Gyrodactylidae)," *Evolution*, vol. 56, no. 12, pp. 2445–2458, 2002.

[19] T. Huyse and F. A. M. Volckaert, "Comparing host and parasite phylogenies: *Gyrodactylus* flatworms jumping from goby to goby," *Systematic Biology*, vol. 54, no. 5, pp. 710–718, 2005.

[20] L. Plaisance, D. T. J. Littlewood, P. D. Olson, and S. Morand, "Molecular phylogeny of gill monogeneans (Platyhelminthes, Monogenea, Dactylogyridae) and colonization of Indo-West Pacific butterflyfish hosts (Perciformes, Chaetodontidae)," *Zoologica Scripta*, vol. 34, no. 4, pp. 425–436, 2005.

[21] M. Mendlová, Y. Desdevises, K. Civáňová, A. Pariselle, and A. Šimková, "Monogeneans of West African cichlid fish: evolution and cophylogenetic interactions," *PloS ONE*, vol. 7, no. 5, Article ID e37268, 2012.

[22] K. Rohde, "Critical evaluation of intrinsic and extrinsic factors responsible for niche restriction in parasites," *The American Naturalist*, vol. 114, no. 5, pp. 648–671, 1979.

[23] S. Morand, A. Šimková, I. Matějusová, L. Plaisance, O. Verneau, and Y. Desdevises, "Investigating patterns may reveal processes: evolutionary ecology of ectoparasitic monogeneans," *International Journal for Parasitology*, vol. 32, no. 2, pp. 111–119, 2002.

[24] A. Šimková, M. Pečínková, E. Řehulková, M. Vyskočilová, and M. Ondračková, "*Dactylogyrus* species parasitizing European *Barbus* species: morphometric and molecular variability," *Parasitology*, vol. 134, no. 12, pp. 1751–1765, 2007.

[25] T. Kaci-Chaouch, O. Verneau, and Y. Desdevises, "Host specificity is linked to intraspecific variability in the genus *Lamellodiscus* (Monogenea)," *Parasitology*, vol. 135, no. 5, pp. 607–616, 2008.

[26] L. H. S. Lim, T. A. Timofeeva, and D. I. Gibson, "Dactylogyridean monogeneans of the siluriform fishes of the Old World," *Systematic Parasitology*, vol. 50, no. 3, pp. 159–197, 2001.

[27] X.-Y. Wu, X.-Q. Zhu, M.-Q. Xie, J. Q. Wang, and A.-X. Li, "The radiation of *Thaparocleidus* (Monogenoidea: Dactylogyridae: Ancylodiscoidinae): phylogenetic analyses and taxonomic implications inferred from ribosomal DNA sequences," *Parasitology Research*, vol. 102, no. 2, pp. 283–288, 2008.

[28] A. Pariselle, L. H. S. Lim, and A. Lambert, "Monogeneans from Pangasiidae (Siluriformes) in Southeast Asia: X. Six new species of *Thaparocleidus* Jain, 1952 (Ancylodiscoididae) from *Pangasius micronema*," *Parasite*, vol. 13, no. 4, pp. 283–290, 2006.

[29] T. R. Roberts and C. Vidthayanon, "Systematic revision of the Asian catfish family Pangasiidae, with biological observations and descriptions of three new species," *Proceedings of the Academy of Natural Sciences of Philadelphia*, vol. 143, pp. 97–144, 1991.

[30] L. Pouyaud, G. G. Teugels, R. Gustiano, and M. Legendre, "Contribution to the phylogeny of pangasiid catfishes based on allozymes and mitochondrial DNA," *Journal of Fish Biology*, vol. 56, no. 6, pp. 1509–1538, 2000.

[31] H. K. Voris, "Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations," *Journal of Biogeography*, vol. 27, no. 5, pp. 1153–1167, 2000.

[32] L. H. S. Lim, "*Silurodiscoides* Gussev, 1961 (Monogenea: Ancyrocephalidae) from *Pangasius sutchi* Fowler, 1931 (Pangasiidae) cultured in Penninsular Malaysia," *Raffles Bulletin of Zoology*, vol. 38, no. 1, pp. 55–53, 1990.

[33] A. Pariselle, L. H. S. Lim, and A. Lambert, "Monogeneans from Pangasiidae (Siluriformes) in Southeast Asia: III. Five new species of *Thaparocleidus* Jain, 1952 (Ancylodiscoididae) from *Pangasius bocourti*, *P. djambal* and *P. hypophthalmus*," *Parasite*, vol. 9, no. 3, pp. 207–217, 2002.

[34] A. Pariselle, L. H. S. Lim, and A. Lambert, "Monogeneans from Pangasiidae (Siluriformes) in Southeast Asia: V. Five new species of *Thaparocleidus* Jain, 1952 (Ancylodiscoididae) from *Pangasius nasutus*," *Parasite*, vol. 10, no. 4, pp. 317–323, 2003.

[35] A. Pariselle, L. H. S. Lim, and A. Lambert, "Monogeneans from pangasiidae (Siluriformes) in Southeast Asia: VII. Six new host-specific species of *Thaparocleidus* Jain, 1952 (Ancylodiscoididae) from *Pangasius polyuranodon*," *Parasite*, vol. 11, no. 4, pp. 365–372, 2004.

[36] A. Pariselle, L. H. S. Lim, and A. Lambert, "Monogeneans from Pangasiidae (Siluriformes) in Southeast Asia: VIII. Four new non-specific species of *Thaparocleidus* Jain, 1952 (Ancylodiscoididae) from *Pangasius polyuranodon* and *P. elongatus*," *Parasite*, vol. 12, no. 1, pp. 23–29, 2005.

[37] A. Pariselle, L. H. S. Lim, and A. Lambert, "Monogeneans from Pangasiidae (Siluriformes) in Southeast Asia: IX. Two new species of *Thaparocleidus* Jain, 1952 (Ancylodiscoididae) from *Pangasius mahakamensis*," *Parasite*, vol. 12, no. 4, pp. 325–329, 2005.

[38] L. Euzet and C. Combes, "Les problèmes de l'espèce chez les animaux parasites," in *Les Problèmes de l'espèce dans le Règne Animal*, C. Boquet, J. Genermont, and M. Lamotte, Eds., vol. 40 of *Memoires de la Societé zoologique de France*, pp. 239–285, 1980.

[39] N. D. Sinnappah, L. H. S. Lim, K. Rohde, R. Tinsley, C. Combes, and O. Verneau, "A paedomorphic parasite associated with a neotenic amphibian host: phylogenetic evidence suggests a revised systematic position for Sphyranuridae within anuran and turtle Polystomatoineans," *Molecular Phylogenetics and Evolution*, vol. 18, no. 2, pp. 189–201, 2001.

[40] A. Šimková, L. Plaisance, I. Matějusová, S. Morand, and O. Verneau, "Phylogenetic relationships of the Dactylogyridae Bychowsky, 1933 (Monogenea: Dactylogyridea): the need for the systematic revision of the Ancyrocephalinae Bychowsky, 1937," *Systematic Parasitology*, vol. 54, no. 1, pp. 1–11, 2003.

[41] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[42] T. A. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Research*, vol. 41, no. 41, pp. 95–98, 1999.

[43] D. L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*, Version 4.0b10, Sinauer Associates, Sunderland, Mass, USA, 2002.

[44] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[45] D. Posada and K. A. Crandall, "Modeltest: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[46] A. Rzhetsky and M. Nei, "A simple method for estimating and testing minimum-evolution trees," *Molecular Biology and Evolution*, vol. 9, no. 5, pp. 945–967, 1992.

[47] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[48] N. Wahlberg, E. Weingartner, and S. Nylin, "Towards a better understanding of the higher systematics of Nymphalidae (Lepidoptera: Papilionoidea)," *Molecular Phylogenetics and Evolution*, vol. 28, no. 3, pp. 473–484, 2003.

[49] J. Yang, S. He, J. Freyhof, K. Witte, and H. Liu, "The phylogenetic relationships of the Gobioninae (Teleostei: Cyprinidae) inferred from mitochondrial cytochrome b gene sequences," *Hydrobiologia*, vol. 553, no. 1, pp. 255–266, 2006.

[50] S. Kumar, K. Tamura, and M. Nei, "MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment," *Briefings in Bioinformatics*, vol. 5, no. 2, pp. 150–163, 2004.

[51] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[52] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[53] P. Legendre, Y. Desdevises, and E. Bazin, "A statistical test for host-parasite coevolution," *Systematic Biology*, vol. 51, no. 2, pp. 217–234, 2002.

[54] J. P. Meier-Kolthoff, A. F. Auch, D. H. Huson, and M. Göker, "CopyCat: cophylogenetic analysis tool," *Bioinformatics*, vol. 23, no. 7, pp. 898–900, 2007.

[55] C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas, "Jane: a new tool for the cophylogeny reconstruction problem," *Algorithms for Molecular Biology*, vol. 5, article 16, 2010.

[56] R. D. M. Page, "Parallel phylogenies: reconstructing the history of host-parasite assemblages," *Cladistics*, vol. 10, no. 2, pp. 155–173, 1994.

[57] F. Ronquist, "TreeFitter, program and documentation," 2001, http://www.softpedia.com/progMoreBy/Publisher-Fredrik-Ronquist-Maxim-Teslenko83341.html.

[58] C. J. Douady, F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery, "Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability," *Molecular Biology and Evolution*, vol. 20, no. 2, pp. 248–254, 2003.

[59] D. Pol and M. E. Siddall, "Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case," *Cladistics*, vol. 17, no. 3, pp. 266–281, 2001.

[60] A. V. Gusev, *Identification of Freshwater Fish Parasites*, Nauka, Leningrad, Russia, 1985.

[61] L. Pouyaud, E. Desmarais, M. Deveney, and A. Pariselle, "Phylogenetic relationships among monogenean gill parasites (Dactylogyridea, Ancyrocephalidae) infesting tilapiine hosts (Cichlidae): systematic and evolutionary implications," *Molecular Phylogenetics and Evolution*, vol. 38, no. 1, pp. 241–249, 2006.

[62] X.-Y. Wu, X.-Q. Zhu, M.-Q. Xie, and A.-X. Li, "The evaluation for generic-level monophyly of Ancyrocephalinae (Monogenea, Dactylogyridae) using ribosomal DNA sequence data," *Molecular Phylogenetics and Evolution*, vol. 44, no. 2, pp. 530–544, 2007.

[63] I. Blasco-Costa, R. Míguez-Lozano, V. Sarabeev, and J. A. Balbuena, "Molecular phylogeny of species of *Ligophorus* (Monogenea: Dactylogyridae) and their affinities within the Dactylogyridae," *Parasitology International*, vol. 61, pp. 619–627, 2012.

[64] M. Vignon, A. Pariselle, and M. P. M. Vanhove, "Modularity in attachment organs of African *Cichlidogyrus* (Platyhelminthes: Monogenea: Ancyrocephalidae) reflects phylogeny rather than host specificity or geographic distribution," *Biological Journal of the Linnean Society*, vol. 102, no. 3, pp. 694–706, 2011.

[65] A. Šimková, O. Verneau, M. Gelnar, and S. Morand, "Specificity and specialization of congeneric monogeneans parasitizing cyprinid fish," *Evolution*, vol. 60, no. 5, pp. 1023–1037, 2006.

[66] J. Jarkovský, S. Morand, A. Šimková, and M. Gelnar, "Reproductive barriers between congeneric monogenean parasites (*Dactylogyrus*: Monogenea): attachment apparatus morphology or copulatory organ incompatibility?" *Parasitology Research*, vol. 92, no. 2, pp. 95–105, 2004.

[67] S. Morand, A. Šimková, and S. Gourbière, "Beyond the paradigms of cospeciation and host-switch: is sympatric speciation an important mode of speciation for parasites?" *Life and Environment*, vol. 58, no. 2, pp. 125–132, 2008.

[68] A. Šimková, M. Ondračková, M. Gelnar, and S. Morand, "Morphology and coexistence of congeneric ectoparasite species: reinforcement of reproductive isolation?" *Biological Journal of the Linnean Society*, vol. 76, no. 1, pp. 125–135, 2002.

[69] T. Giraud, P. Gladieux, and S. Gavrilets, "Linking the emergence of fungal plant diseases with ecological speciation," *Trends in Ecology and Evolution*, vol. 25, no. 7, pp. 387–395, 2010.

[70] S. Bentz, S. Leroy, L. Du Preez, J. Mariaux, C. Vaucher, and O. Verneau, "Origin and evolution of African *Polystoma* (Monogenea: Polystomatidae) assessed by molecular methods," *International Journal for Parasitology*, vol. 31, no. 7, pp. 697–705, 2001.

[71] D. M. de Vienne, G. Refrégier, M. López-Villavicencio, A. Tellier, M. E. Hood, and T. Giraud, "Cospeciation versus host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution," *New Phytologist*, vol. 198, no. 2, pp. 347–385, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 856521, 12 pages
http://dx.doi.org/10.1155/2013/856521

Research Article

The Evolutionary Pattern and the Regulation of Stearoyl-CoA Desaturase Genes

Xiaoyun Wu,[1,2] Xiaoju Zou,[3] Qing Chang,[2] Yuru Zhang,[2] Yunhai Li,[2] Linqiang Zhang,[2] Jingfei Huang,[1] and Bin Liang[2]

[1] *State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*

[2] *Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences and Yunnan Province, Kunming Institute of Zoology, Kunming, Yunnan 650223, China*

[3] *Department of Life Science and Biotechnology, Kunming University, Kunming 650214, China*

Correspondence should be addressed to Bin Liang; liangb@mail.kiz.ac.cn

Received 23 January 2013; Accepted 14 September 2013

Academic Editor: William Piel

Stearoyl-CoA desaturase (SCD) is a key enzyme that converts saturated fatty acids (SFAs) to monounsaturated fatty acids (MUFAs) in the biosynthesis of fat. To date, two isoforms of *scd* gene (*scd1* and *scd5*) have been found widely existent in most of the vertebrate animals. However, the evolutionary patterns of both isofoms and the function of *scd5* are poorly understandable. Herein, we aim to characterize the evolutionary pattern of *scd* genes and further predict the function differentiation of *scd* genes. The sequences of *scd* genes were highly conserved among eukaryote. Phylogenetic analysis identified two duplications of *scd* gene early in vertebrate evolution. The relative rate ratio test, branch-specific *dN/dS* ratio tests, and branch-site *dN/dS* ratio tests all suggested that the *scd* genes were evolved at a similar rate. The evolution of *scd* genes among eukaryote was under strictly purifying selection though several sites in *scd1* and *scd5* were undergone a relaxed selection pressure. The variable binding sites by transcriptional factors at the 5′-UTR and by miRNAs at 3′-UTR of *scd* genes suggested that the regulators of *scd5* may be different from that of *scd1*. This study promotes our understanding of the evolutionary patterns and function of SCD genes in eukaryote.

1. Introduction

Stearoyl-CoA desaturase (SCD) is an intrinsic membrane protein that binds to the endoplasmic reticulum (ER), composed of four transmembrane domains [1–3]. SCD is the rate-limiting enzyme that introduces the first cis-double bond at the delta-9 position of saturated fatty acids (SFAs) to thereby generate monounsaturated fatty acids (MUFAs) [4], which are major substrates for biosynthesis of polyunsaturated fatty acids (PUFAs) and complex lipids such as triglycerides, phospholipids, cholesterol esters, and wax esters being as energy storage, components of biological membrane, and signaling molecules. The ratio of unsaturated fatty acids to saturated fatty acids plays a vital role in cell signaling and membrane fluidity, in which imbalance of this ratio is often associated with diseases like diabetes, cardiovascular diseases, fatty liver, cancers and stresses resistance, and so forth [5].

The *scd* genes are universally present in living organisms. The number of *scd* genes varies from one to five, which are generally called *scd1, scd2, scd3, scd4,* and *scd5* in different organisms [4, 6], but with other distinct names in invertebrates such as *fat-5*, *fat-6*, and *fat-7* in *Caenorhabditis elegans* [7–9] and *ole1* in *Saccharomyces cerevisiae* [10]. The yeast genome contains only *ole-1* gene encoded SCD, and *ole-1* mutant requires unsaturated fatty acids for growth [10]. The desaturase of *C. elegans* FAT-5, FAT-6, and FAT-7 displays substrate preferences, in which both FAT-6 and FAT-7 mainly desaturate stearic acid (18 : 0) and have less activity on palmitic acid (16 : 0). On the contrary, FAT-5 desaturates palmitic acid (16 : 0) but has nearly undetectable activity on stearic acid (18 : 0) [7]. The evolutionary history revealed that the *scd* genes in vertebrates could be distinctly classified into *scd5* type [3, 6, 11] and *scd1* type including its homologs *scd2, scd3, and scd4* [6, 12]. The divergence of *scd1* and *scd5* genes occurred early in vertebrate evolution due to the whole genome duplication (2R) [6]. However, the *scd* genes may have distinct fates after gene duplication event. It is unknown whether one *scd* evolved faster and acquired new function more rapidly than the other, and whether the selective patterns on both *scd* genes were similarly changed following the duplication.

Interestingly, though the enzymes of *scd* genes display similar delta-9 desaturation activity [4], the expression pattern of *scd1* and *scd5* is variable that *scd1* is ubiquitous, but *scd5* is mainly in the brain and pancreas even in different species [3, 6, 11], implying that the regulation of *scd1* and *scd5* expression and biological function may be distinct. The promoter region of *scd1* contains many consensus binding sites for numerous transcription factors, for example, SREBP1, LXR, PPARα, C/EBP-α, NF-1, NF-Y, and Sp1 [13]. However, it is unclear whether *scd5* contains similar or completely different consensus binding sites with *scd1*. Meanwhile, it is completely unknown that the 3′-UTR of *scd1* and *scd5* that may also contain similar or different target sites of microRNAs regulating their expression.

Therefore, to address the above questions, we compared the sequence characteristics of *scd* paralogs and then reconstructed the phylogenetic trees of *scd* genes in eukaryote species to determine the evolutionary history of *scd* genes. We used the relative rate ratio test, branch-specific dN/dS ratio tests, and branch-site dN/dS ratio tests to analyze the evolutionary forces after gene duplication. Furthermore, we characterized the binding sites by transcript factors in the 5′-UTR and the target sites by microRNAs in the 3′-UTR of *scd1* and *scd5* genes to investigate the regulation mechanisms of both *scd* genes.

2. Material and Methods

2.1. SCD Homologs BLAST, Sequence Alignment, and Phylogenetic Analysis. SCD homologs were retrieved by key word "Stearoyl-CoA desaturase" from NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and Ensemble genome database (http://asia.ensembl.org/index.html). In addition, the sequences of human SCD proteins were used to blast available genomes from NCBI GenBank and Ensemble database. Eventually, 73 *scd* nucleotide sequences from 39 representative eukaryote species were retrieved (see Table S1 in the Supplementart Material available online at http://dx.doi.org/10.1155/2013/856521). Sequence alignment of 73 *scd* nucleotides was performed with MegAlign implemented in DNAStar 6.0 software package (DNASTAR, Madison, USA) and then was confirmed visually by BioEdit 7.0.9 [14]. The ambiguous regions of alignment were discarded and eventually 720 nucleotide bases were obtained.

Phylogenetic tree was reconstructed based on the full alignment of 73 sequences by using Maximum Likelihood (ML) analysis in PHYML [15] and approximately Maximum Likelihood (ML) analysis in FastTree 2.1.3 [16]. The yeast *scd* ortholog, *ole1*, was used as the outgroup to root the tree. For ML analysis, supports for nodes among branches were evaluated using nonparametric bootstrapping [17] with 1000 bootstrap replications. For FastTree 2 analysis, a heuristics search strategy was employed with an estimated rate of evolution for each site (the "CAT" approximation), minimum-evolution subtree-pruning regrafting (SPRs), and maximum-likelihood nearest-neighbor interchanges (NNIs). The local support values were provided based on the Shimodaira-Hasegawa (SH) test [18, 19].

To evaluate the evolutionary conservation of the SCD1 and SCD5, the amino acid sequences of SCD1 and SCD5 of 11 model organisms including human, rhesus monkey, mouse, rat, tree shrew, zebrafish, *Drosophila melanogaster*, and *C. elegans* were retrieved and then aligned using Muscle (http://www.ebi.ac.uk/Tools/msa/muscle/), followed by manual adjustment with BioEdit 7.0.9 [14]. Additionally, a Neighbouring-Joining (NJ) tree was reconstructed with the amino acid sequences of SCDs from human, rhesus monkey, mouse, rat, tree shrew, and *C. elegans* by MEGA 4.0 [20] using amino acid p-distance model. Support for nodes among branches was evaluated using nonparametric bootstrapping [17] with 1000 bootstrap replications.

2.2. Regulation Prediction in 3′-UTR and 5′-UTR of scd Genes. Searching for the transcription factor-binding sites (TFBS) in the 5′-UTR of *scd* genes was carried out based on the positive effectors of transcription in the promoter region of *scd1* from human, mouse, and chicken [13]. The length of 5′-UTR for this analysis was about 2500 bp upstream of the translational start sites of *scd5* gene. The TFBSs were estimated by Match 1.0 with the TRANSFAC database v. 6.0 and Promo with TRANSFAC database v. 8.3 [21, 22]. The cut-off parameters were set as 0.75 for the core similarity and 0.85 for matrix similarity in Match 1.0 analysis. In Promo analysis, the species of factor and site were only constrained to animals. MultiSearchSite was used to search for binding sites sharing 15% maximum matrix dissimilarity rate in the promoter sequences of human, rhesus monkey, tree shrew, and chicken.

The microRNA targets sites in the 3′-UTR region of *scd* genes were predicted by using TargetScan release 6.2 (http://www.targetscan.org/). The lengths of the 3′-UTR region of *scd1* and *scd5* genes were about 4000 bp and 1790 bp, respectively. Only the broadly conserved sites for miRNA families among vertebrates were considered in this study. The predicted miRNAs were then introduced to the miR2Disease Base (http://www.mir2disease.org/) to establish the relationship between miRNAs and human diseases.

2.3. Relative Rate Test. The substitution rates of the *scd* genes were compared among different paralogs inferred from the phylogenetic tree using the relative rate test implemented in RRTree [23]. Three phylogroups were defined as vertebrate *scd1*, vertebrate *scd5*, and invertebrate *scd* gene. The yeast *ole1* gene was used as outgroup.

2.4. Selective Pattern Analysis. The ratio of synonymous substitution to nonsynonymous substitution ($\omega = dN/dS$) is a good indicator to estimate the evolutionary selective pressure of protein-coding regions. The ratio of $\omega = 1$, <1, and >1 indicates a neutral selection, a purifying selection, and positive selection, respectively. The ω ratios between pairwise sequences were estimated following the method of Yang and Nielsen [24].

The codon-substitution models were implemented using CODEML in PAML package [25]. All models fixed the transition/transversion rate and codon usage biases (F3×4). To determine the evolutionary selective patterns of two *scd* genes, the branch-specific model was applied to the data, which assumed that the foreground clade had different ratios from the background clade [26]. In model B, *scd1* and *scd5* genes were set as the foreground clade. In model C, *scd1*, *scd5,* and the invertebrate SCD homologs were set as three clades. In addition, we also determined the sites evolving under positive selection in a specific clade with the branch-site model that allows variation in ω across individual codons on a specific lineage [27]. We applied the modified branch-site model A (test 1 and test 2) [27], which permits variation of the ω ratio both among sites and lineages. The likelihood ratio tests (LRTs) were constructed to compare the fit to the data of two nested models. The significant difference between two models was evaluated by calculating twice the log-likelihood difference, and followed an χ^2 distribution with the number degree of freedom equal to the difference in the number of free parameters.

3. Results

3.1. The Sequence Characteristics of SCD Orthologs. In human, the size of *scd1* gene is about 17 kb and 170 kb for *scd5* gene. Though the remarkably different sizes of two *scd* genes, the full lengths of both *scd* encoded proteins are very close that SCD1 has 359 aa and SCD5 330 aa (Figure 1). To determine the conservation of SCD orthologs, we first investigated the sequence characteristics of SCD proteins. Comparison of the SCD amino acid sequences from several animal organisms revealed that the three histidine motifs HRLWSH, HRAHH, and HNYHH that exist in human SCD are also highly conserved in all alignments (Figure 1). But, the three histidine motifs also display minor changes in some organisms. For example, HRLWAH exists in *C. elegans* FAT-5 and *Drospholia* SCD genes; HNFHH in *C. elegans* FAT-6 and FAT-7 (Figure 1). The four transmembrane hydrophobic domains marked underline are also conserved in all alignments (Figure 1). Then, we investigated the sizes and order of exons of *scd* genes in several representative eukaryote organisms (Figure 2). Most of the *scd1* genes (e.g., chicken, human, etc.) are consisted of 6 exons. However, some vertebrate *scd1* genes only have 5 exons, like platypus and zebrafish. All of the *scd5* genes are consisted of 5 exons. Very interestingly, except the exon 1, the sizes and order of other exons (exon 2 (131), exon 3 (206), exon 4 (233), and exon 5 (191)) of *scd5* genes were not only separately equal but also very similar to the sizes and order of the third to sixth exons of *scd1* genes (exon 3 (131), exon 4 (206), exon 5 (233) and exon 6 (200)) in eukaryote organisms (Figures 2(a) and 2(b)).

3.2. Phylogenetic Inference of scd Gene Lineages. The phylogenetic tree of *scd* genes based on the 73 nucleotide sequences from 39 species is shown in Figure 3(a) (TreeBASE Accession URL http://purl.org/phylo/treebase/phylows/study/TB2:S14739). The *scd* orthologs of invertebrate species are placed at the base of the tree using *scd* ortholog yeast *ole1* as outgroup. The *C. elegans fat-5*, *fat-6*, and *fat-7* are placed at the most basal position of the tree. In addition, the *scd1a, scd1b,*and *scd1c* from *Ciona savignyi* and amphioxus *Branchiostoma floridae* are just located out of the vertebrate lineages. Intriguingly, the *scd* genes in vertebrates are split into two lineages with strong support (support value = 99%) according to the *scd* gene classification, suggesting that independent duplication events occurred in vertebrates after separation from invertebrates during evolution. In teleost fish, two *scd1* paralogs were also diverged into two independent clades with high support, but the *scd5* gene was lost. This evolutionary pattern might suggest that the teleost fish *scd* experienced an ancient gene duplication event [12] or the genome duplication [6].

3.3. Evolutionary Rates and Selective Pattern in scd Genes. To determine whether the paralogs of *scd* evolve at the similar rates, the relative rate analysis was performed among *scd* gene and in which the invertebrate *scd* genes, vertebrate *scd1* and *scd5* genes were separated into 3 groups using the yeast *ole1* as outgroup. The analysis revealed that the *scd* genes were evolved at the similar evolutionary rate ($P < 0.05$).

To address the selective constraint pattern within *scd* genes, the ratios of nonsynonymous (dn) to synonymous (ds) were estimated between two sequences. The analysis suggested that nearly almost pairwise comparisons of *scd* genes had a $\omega < 1$, indicating a strong purifying selection. Intriguingly, the pairwise comparisons among *scd1* genes of human, gorilla, and chimpanzee had a $\omega = \infty$, which might result from that the nonsynonymous substitution occurred while the synonymous substitution did not in *scd1* sequences probably because of the very close relationships among these three species.

The selective pattern of *scd* genes was further performed using the condon-based maximum likelihood analysis (Table 1). In this analysis, the yeast *ole1* was excluded. The estimated one ratio of ω_0 (0.08684) over all sites and branches from the *scd* genes was substantially smaller than 1, suggesting a strong purifying selection (Table 1). In the branch-specific models, Model B assumes *scd1* gene and *scd5* gene as the foreground clades, respectively. In this model the estimated ω value was 0.09207 for *scd1* gene and 0.07951 for the background clades. The estimated ω value was 0.06146 for *scd5* gene and 0.09735 for the background clades. The LRT test suggested that the two-ratio model was not fit for the data better than the one-ratio model for *scd1* gene ($P > 0.05$) but fit better for *scd5* gene ($P < 0.001$). Under Model C, ω estimates for *scd1*, *scd5,* and invertebrate *scd* gene were 0.06140, 0.09198, and 0.11788, respectively. The LRT test indicated that Model C

```
C. elegans FAT5              M-------------T---------QIKVDAIISKQFLAAD-----------LNEIRQMQEE----SKKQV
C. elegans FAT6              M-------------TVKTRSNIAKKIEKDGGPETQYLAVD-----------PNEIIQLQEE----SKKIP
C. elegans FAT7              M-------------TVKTRASIAKKIEKD-GLDSQYLFMD-----------PNEVLQVQEE----SKKIP
Fruit fly SCD                MP----PNAQAGAQSISDSLIAAASAAADAGQSPTKLQEDSTGVLFECDV-ETTDGGLVKDITVMKKAEK
Zebrafish SCD1a              M----------PDSDVKAPVLQPQLEAM-----------------------EDEFDPLYKE----KPGPK
Zebrafish SCD1b              M------------ADVTTTT-------------------------------EDVFDDSYVE----KPGPS
Xenopus tropicalis SCD1      M-------TFRGSQTVSTVMESPSIIQDEIGADRVMT--------------DDIFDTTYIK----KVDFK
Mouse SCD1                   MPAHMLQE-ISSSYTTTTTITAPPSG---NEREKVKTVPLHLEEDIRPEMKEDIHDPTYQD----EEGPP
Tree shrew SCD1              MPAHMLQDEISSSYTTTTTITAPPSRVLQNGGGKLEKTSLCFDEDIRPEISDDIHDPSYQD----KEGPA
Rhesus monkey SCD1           MPAHLLQEDISSSYTTTTTITAPPSRVLQNGRDKLETTPLYLEEDVRPDIKDDIYDPTYKD----KEGPS
Human SCD1                   MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKD----KEGPS
Tree shrew SCD5              M--------PGLAADAGKVPFCDAKEEIRAG---L----------------EGSEGGGGPE----RPDAR
Rhesus monkey SCD5           M--------PGPATDAGKIPFCDAKEEIRAG---L----------------ESSEGGGGPE----RPGAR
Human SCD5                   M--------PGPATDAGKIPFCDAKEEIRAG---L----------------ESSEGGGGPE----RPGAR

C. elegans FAT5              IKME-IVWKNVALFVALHIGALVGLYQLVFQAKWATVGWVFLLHTLGSMGVTGGAHRLWAHRAYKATLSW
C. elegans FAT6              YKME-IVWRNVALFAALHFAAAIGLYQLIFEAKWQTVIFTFLLYVFGGFGITAGAHRLWSHKSYKATTPM
C. elegans FAT7              YKME-IVWRNVALFAALHVAAAIGLYELVFHAKWQTAVFSFALYVFSGFGITAGAHRLWSHKSYKATTPM
Fruit fly SCD                RRLK-LVWRNIIAFGYLHLAALYGAYLMVTSAKWQTCILAYFLYVISGLGITAGAHRLWAHRSYKAKWPL
Zebrafish SCD1a              PPMK-IVWRNVILMSLLHIAAVYGLF-LIPSAHPLTLLWAFACFVYGGLGITAGVHRLWSHRSYKATLPL
Zebrafish SCD1b              PPVQ-IVWRNVILMTLLHLGALYGMT-ILPFVSSLTLIWTGVCFMVSALGITAGAHRLWSHRSYRASLPL
Xenopus tropicalis SCD1      PPIK-LVWRNVILMALLHFGAFYGLF-MIPAAKPITLAWAILCFMLSALGVTAGAHRLWSHRSYKAKLPL
Mouse SCD1                   PKLE-YVWRNIILMVLLHLGGLYGII-LVPSCKLYTCLFGIFYYMTSALGITAGAHRLWSHRTYKARLPL
Tree shrew SCD1              PKLE-YVWRNIILMSLLHLGALYGIV-LFPTSKFYTWLWVLFYYLVSALGITAGAHRLWSHRTYKARLPL
Rhesus monkey SCD1           PKVE-YVWRNIILMSLLHLGALYGIT-LIPTCKLYTCLWGLFYYVVSALGITAGAHRLWSHRSYKARLPL
Human SCD1                   PKVE-YVWRNIILMSLLHLGALYGIT-LIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPL
Tree shrew SCD5              GRRQNIVWRNVVLMSLLHLGAVYSLV-LIPKAKPLTLLWAYFCFLLTALGVTAGAHRLWSHRSYKAKLPL
Rhesus monkey SCD5           GQRQNIVWRNVVLMSLLHLGAVYSLV-LIPKAKPLTLLWAYFCFLLAALGVTAGAHRLWSHRSYKAKLPL
Human SCD5                   GQRQNIVWRNVVLMSLLHLGAVYSLV-LIPKAKPLTLLWAYFCFLLAALGVTAGAHRLWSHRSYRAKLPL

C. elegans FAT5              RVFLMLINSIAFQNDIIDWARDHRCHHKWTDTDADPHSTNRGMFFAHMGWLLVKKHDQLKIQGGKLDLSD
C. elegans FAT6              RIFLMILNNIALQNDVIEWARDHRCHHKWTDTDADPHNTTRGFFFAHMGWLLVRKHPQVKEQGAKLDMSD
C. elegans FAT7              RIFLMLLNNIALQNDIIEWARDHRCHHKWTDTDADPHNTTRGFFFTHMGWLLVRKHPQVKEHGGKLDLSD
Fruit fly SCD                RVILVIFNTIAFQDAAYHWARDHRVHHKYSETDADPHNATRGFFFSHVGWLLCKKHPEVKAKGKGVDLSD
Zebrafish SCD1a              RIFLAIGNSMAFQNDIYEWSRDHRVHHKYSETDADPHNSNRGFFFSHVGWLLVRKHPEVIERGRKLELTD
Zebrafish SCD1b              RIFLAVANSMAFQNDIYEWARDHRVHHKFSETDADPHNARRGFFFAHIGWLLVRKHPEVIDKGRKLTFED
Xenopus tropicalis SCD1      RIFLAVVNSMAFQNDIYEWARDHRVHHKYSETDADPHNAVRGFFFSHIGWLLMRKHPDVIEKGKKLDLSD
Mouse SCD1                   RIFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGGKLDMSD
Tree shrew SCD1              RLFLIIANTMAFQNDVYEWARDHRVHHKFSETHADPHNARRGFFFSHVGWLLVRKHPAVKEKGALLDLSD
Rhesus monkey SCD1           RLFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGATLDLSD
Human SCD1                   RLFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSD
Tree shrew SCD5              RIFLAAANSMAFQNDIFEWCRDHRVHHKYSETDADPHNARRGFFFSHIGWLFVRKHRDVIEKGRKLDFTD
Rhesus monkey SCD5           RIFLAVANSMAFQNDIFEWSRDHRAHHKYSETDADPHNARRGFFFSHIGWLFVRKHPDVIEKGRKLDVTD
Human SCD5                   RIFLAVANSMAFQNDIFEWSRDHRAHHKYSETDADPHNARRGFFFSHIGWLFVRKHRDVIEKGRKLDVTD

C. elegans FAT5              LYEDPVLMFQRKNYLPLVGIFCFALPTFIPVVLWGESAFIAFYTAALFRYCFTLHATWCINSVSHWVGWQ
C. elegans FAT6              LLSDPVLVFQRKHYFPLVILCCFILPTIIPVYFWKETAFIAFYTAGTFRYCFTLHATWCINSAAHYFGWK
C. elegans FAT7              LFSDPVLVFQRKHYFPLVILFCFILPTIIPVYFWKETAFIAFYVAGTFRYCFTLHATWCINSAAHYFGWK
Fruit fly SCD                LRADPILMFQKKYYMILMPIACFIIPTVVPMYAWGESFMNAWFVATMFRWCFILNVTWLVNSAAHKFGGR
Zebrafish SCD1a              LKADKVVMFQRRFYKLSVVLMCFVVPTVVPCYMWGESLWIAYFIPTLLRYALGLNSTWLVNSAAHMWGNR
Zebrafish SCD1b              LKADSVVMFQRRHYKLSVVVMCFLIPTLVPWFFWEESLWTAYLVPCLLRYAVVLNATWLVNSAAHMWGMR
Xenopus tropicalis SCD1      LKADKVVMFQRRNYKLSILVMCFILPTVIPWYFWDESFSVAFYVPCLLRYALVLNATWLVNSAAHMYGNR
Mouse SCD1                   LKAEKLVMFQRRYYKPGLLLMCFILPTLVPWYCWGETFVNSLFVSTFLRYTLVLNATWLVNSAAHLYGYR
Tree shrew SCD1              LRAEKLVMFQRRYYVPGVLLMCFILPTLVPWCLWGETFMHSLYVATLLRYATVLNATWLVNSAAHLYGYR
Rhesus monkey SCD1           LEAEKLVMFQRRYYKPGLLLMCFILPTLVPWCFWGETFQHSVFVATFLRYAIVLNVTWLVNSAAHLFGYR
Human SCD1                   LEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYR
Tree shrew SCD5              LLADPVVRFQRKYYKITVVLMCFAVPTLVPWYIWGESLWNSYFLASILRYTISLNVTWLVNSAAHMYGNR
Rhesus monkey SCD5           LLADPVVRIQRKYYKISVVLMCFAVPTLVPWYIWGESLWNSYFLASILRYTISLNVAWLVNSAAHMYGNR
Human SCD5                   LLADPVVRIQRKYYKISVVLMCFVVPTLVPWYIWGESLWNSYFLASILRYTISLNISWLVNSAAHMYGNR

C. elegans FAT5              PYDHQASSVDNLWTSIAAVGEGGHNYHHTFPQDYRTSEHA-EFLNWTRVLIDFGASIGMVYDRKTTPEEV
C. elegans FAT6              PYDSSITPVENVFTTIAAVGEGGHNFHHTFPQDYRTSEYS-LKYNWTRVLIDTAAALGLVYDRKTACDEI
C. elegans FAT7              PYDTSVSAVENVFTTVVAVGEGGHNFHHTFPQDYRASEYS-LIYNWTRVLIDTAAVLGLVYDRKTIADEF
Fruit fly SCD                PYDKFINPSENISVAILAFGEGWHNYHHVFPWDYKTAEFGKYSLNFTTAFIDFFAKIGWAYDLKTVSTDI
Zebrafish SCD1a              PYDGNIGPRENRFVTFSAIGEGYHNYHHTFPYDYSTSEYG-WKLNLTTIFVDTMCFLGLASNRKRVSKEL
Zebrafish SCD1b              PYDHNINPRENKFVAFSAIGEGFHNYHHTFPHDYATSEFG-SRLNVTKAFIDLMCFLGLANDCRRVTHET
Xenopus tropicalis SCD1      PYDQTINPRENPLVAIGAIGEGFHNYHHTFPFDYSTSEFG-LKFNITTGFIDLMCLLGLANDCKRVSKET
Mouse SCD1                   PYDKNIQSRENILVSLGAVGEGFHNYHHTFPFDYSASEYR-WHINFTTFFIDCMAALGLAYDRKKVSKAT
Tree shrew SCD1              PYDKTINPRENILVSLGAVGEGFHNYHHTFPYDYSASEYR-WHINLTTFFIDCMAALGLAYDRKKVSKAA
Rhesus monkey SCD1           PYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYR-WHINFTTFFIDCMA---------------
Human SCD1                   PYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYR-WHINFTTFFIDCMAALGLAYDRKKVSKAA
Tree shrew SCD5              PYNKHISPRQNPLVTLGAIGEGFHNYHHTFPFDYSASEFG-LNFNPTTWFIDFMCWLGLATDRKRAPKPM
Rhesus monkey SCD5           PYDKHISPRQNPLVALGAIGEGFHNYHHTFPFDYSASEFG-LNFNPTTWFIDLMCWLGLATDRKRATKPM
Human SCD5                   PYDKHISPRQNPLVALGAIGEGFHNYHHTFPFDYSASEFG-LNFNPTTWFIDFMCWLGLATDRKRATKPM
```

(a)

Figure 1: Continued.

```
C. elegans FAT5              IQRQCKKFGCETEREKMLHKLG-----------------
C. elegans FAT6              IGRQVSNHGCDIQRGKSIM--------------------
C. elegans FAT7              ISRQVANHGSEESRKKSIM--------------------
Fruit fly SCD                IKKRVKRTGDGTHATWGWGDVDQPKEEIEDAVITHKKSE
Zebrafish SCD1a              ILARVKRTGDGSYRSG-----------------------
Zebrafish SCD1b              ILARVQRTGDGSHKSG-----------------------
Xenopus tropicalis SCD1      IMARKKRTGDGSHRSG-----------------------
Mouse SCD1                   VLARIKRTGDGSHKSS-----------------------
Tree shrew SCD1              VLARIKRTGDGTYKSG-----------------------
Rhesus monkey SCD1           ---------------------------------------
Human SCD1                   ILARIKRTGDGNYKSG-----------------------
Tree shrew SCD5              IEAQKARTGDGSA--------------------------
Rhesus monkey SCD5           IEARKARTGDSSA--------------------------
Human SCD5                   IEARKARTGDSSA--------------------------
```

(b)

Figure 1: Alignment of inferred SCD protein sequences from 8 model animals. The three histidine motifs are in bold, and the four transmembrane hydrophobic domains were marked underline.

was significantly better fit for the data than did the one ratio model (M0) ($P < 0.001$).

In addition, we determined the amino acid sites under positive selection at SCD1 and SCD5 clades on the phylogeny using the branch-site model. In this model, the SCD1 and SCD5 clades were assumed as the foreground clades, respectively. As seen in Table 1, the results of test 1 analysis designated several amino acid sites under the relaxed selection ($P < 0.001$) in both the *scd1* and *scd5* genes. However, none of the LRT test for *scd* genes was significant in test 2 analysis, indicating that the null hypothesis of the test 2 could not be rejected in both of the *scd* genes, and none of the two *scd* genes was underrelaxed selective constraint or under positive selection. Thus, we did not find any evidence for positive selection in both of the *scd* genes under these analyses.

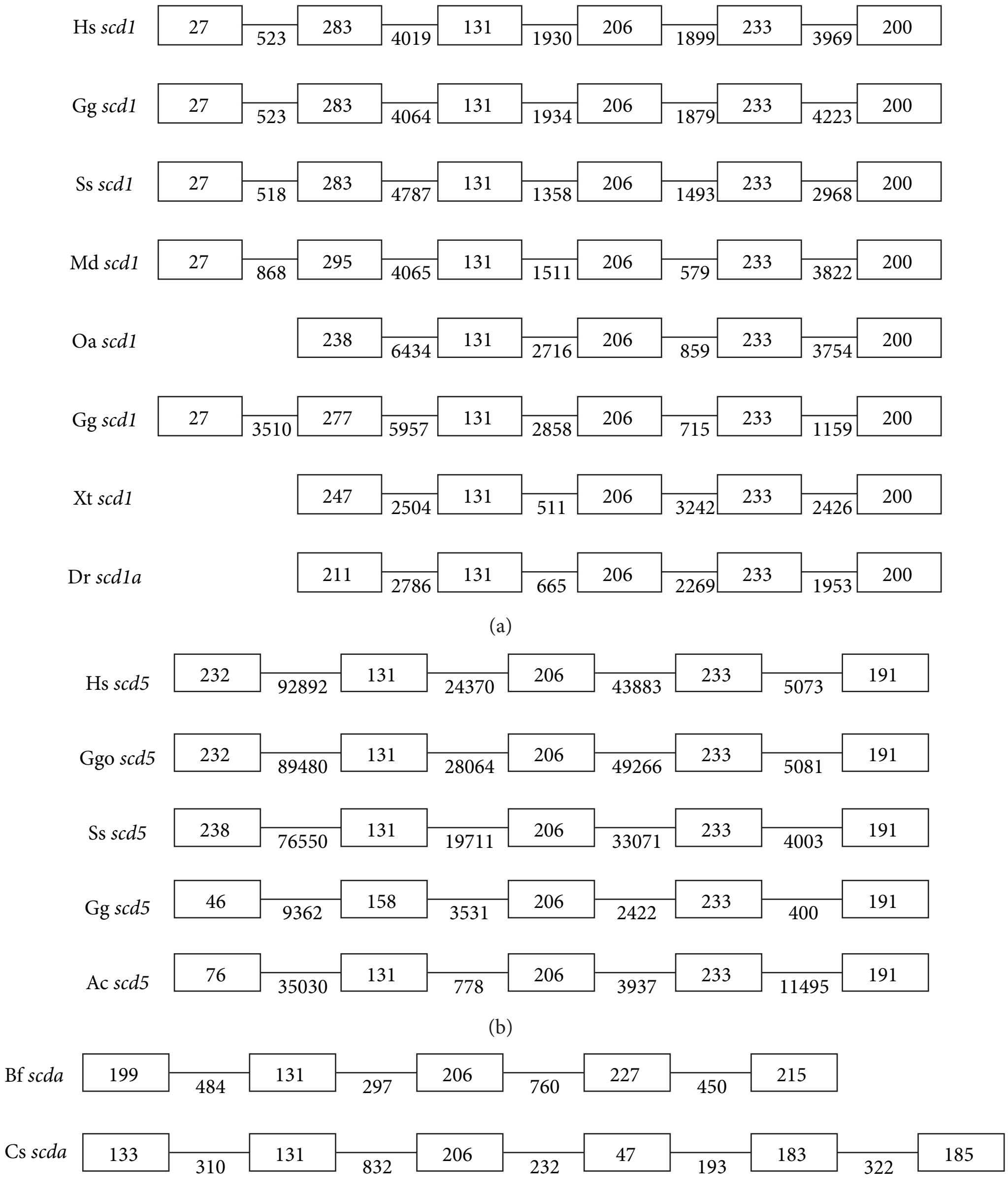
3.4. The Regulation Analysis of scd Genes.

Numerous transcription factors, for example, SREBP1, LXR, PPARα, C/EBP-α, NF-1, NF-Y, and Sp1, have been revealed to bind to the *scd1* promotor region [13]. The consensus binding sites for the SREBP1, PPAR-α, C/EBP-α, NF-1, and NF-Y were known to mediate the insulin response, whereas the binding sites for Sp1 and AP1 were known to be the leptin response element. To determine whether these transcription factors also bind to *scd5* promotor region, the transcription binding site prediction was performed by using TRANSFAC and Promo. C/EBP-α, AP1, SP1, NF-1, NF-Y, and SREBP1 were detected at the promoter region of *scd5* gene of four species (Table 2). But SREBP1 was not detected in the promoter region of *scd5* gene in other mammals (results not shown). Because SREBPs are weak transcriptional activators on their own, they interact with their target promoters in cooperation with additional regulators, most commonly including one or both of the transcription factors NFY and SP1 [28–30], and their binding sites were possessed a high degree of overlap [31]. We also detected the binding sites of NFY and SP1 near the binding site of SREBP1 in human. In this analysis, we detected the binding site of PPARα by Promo, but not by TRANSFAC. However, the binding site of PPARα detected in *scd5* gene was different from that of in *scd1* gene (Table 2). Though most of the transcription factor binding sites in *scd1* gene could be detected in *scd5* gene, the regulation of these transcription factors on *scd5* gene still needs further experimental verification.

In order to compare the microRNAs regulation on *scd* genes, we predicted the microRNA target sites at the 3′-UTR region of *scd1* and *scd5* genes using TargetScan. The lengths of 3′-UTR region of *scd1* and *scd5* gene were about 4000 bp and 1790 bp, respectively (Figure 4). Within the 3′-UTR region of *scd1* gene, 8 conserved sites of microRNA families were predicted among vertebrates and 5 conserved sites were predicted among mammals (Figure 4(a)). Among these 13 microRNA families, almost all of them were closely associated with the cancers, for example, the miR-128, Let 7, miR-206, and miR-124a linked to breast cancer [32–35], hepatocellular carcinoma [36–38], and pancreatic cancer [39, 40]. In addition, plenty of evidence has described that the *scd1* acted as a potential target to prevent or treat metabolic syndrome. Among these microRNA families, several microRNAs were associated with the nonalcoholic fatty liver disease (NAFLD), type 2 diabetes, and diabetic nephropathy; for example, miR-429 and Let 7cde were closely related to NAFLD [41, 42]; miR-181a related to diabetes [43]; miR-216a related to diabetic nephropathy [44]. At the 3′-UTR region of *scd5* gene, 5 conserved sites of microRNA families were predicted among vertebrates and 2 conserved sites were predicted among mammals (Figure 4(b)). All of these microRNA families were closely associated with cancers. *scd5* gene was mainly expressed in brain and pancreas. Several microRNAs were associated with the neurological disorder and pancreatic cancers. miR-106a was associated with autism [45], miR-17 with glioma [46], miR-20b with schizophrenia [47]. miR-205, miR-221, miR-222, miR-17-5p, and miR-20a were associated with pancreatic cancers [39, 48–50]. Only 2 microRNAs, miR-200ab and miR-17, were linked to NAFLD [41, 42].

4. Discussion

The phylogenetic trees show that homologs of *scd* gene from invertebrates were all placed at the basal position of the tree,


Hs *scd1* 27 523 283 4019 131 1930 206 1899 233 3969 200
Gg *scd1* 27 523 283 4064 131 1934 206 1879 233 4223 200
Ss *scd1* 27 518 283 4787 131 1358 206 1493 233 2968 200
Md *scd1* 27 868 295 4065 131 1511 206 579 233 3822 200
Oa *scd1* 238 6434 131 2716 206 859 233 3754 200
Gg *scd1* 27 3510 277 5957 131 2858 206 715 233 1159 200
Xt *scd1* 247 2504 131 511 206 3242 233 2426 200
Dr *scd1a* 211 2786 131 665 206 2269 233 1953 200
Hs *scd5* 232 92892 131 24370 206 43883 233 5073 191
Ggo *scd5* 232 89480 131 28064 206 49266 233 5081 191
Ss *scd5* 238 76550 131 19711 206 33071 233 4003 191
Gg *scd5* 46 9362 158 3531 206 2422 233 400 191
Ac *scd5* 76 35030 131 778 206 3937 233 11495 191
Bf *scda* 199 484 131 297 206 760 227 450 215
Cs *scda* 133 310 131 832 206 232 47 193 183 322 185

(a)

(b)

(c)

Figure 2: The exons size changes of *scd* genes. (a) Exon size changes of *scd1* gene in vertebrates. (b) Exon size changes of *scd5* gene in vertebrates. (c) Exon size changes of *scd* genes in invertebrates. Numbers in box represent the sizes of exons and numbers under bars represent the sizes of introns. Hs, *Homo sapiens*; Ggo, *Gorilla gorilla*; Ss, *Sus scrofa*; Md, *Monodelphis domestica*; Oa, *Ornithorhynchus anatinus*; Gg, *Gallus gallus*; Xt, *Xenopus tropicalis*; Dr, *Danio rerio*; Ac, *Anolis carolinensis*; Bf, *Branchiostoma floridae*; Cs, *Ciona savignyi*.

whereas the *scd* genes in vertebrates were diverged into two independently duplicated genes early in vertebrate evolution with strong support, in which all *scd1* genes form a distinct clade and all *scd5* genes clustered into another clade (Figures 3(a) and 3(b)). Our phylogenetic analysis was consistent with the previous studies by Castro et al. [6] and Lengi and Corl [11]. This pattern of duplication might be resulted from part of the two rounds of genome duplication in vertebrate ancestry [6].

When a gene duplication event occurs, the duplicated genes have redundant functions. The fate of the duplicated genes might be loss of function, gaining a new function, or subfunctionalization [51]. Subfunctionalization occurred when both duplicates can be stably maintained in the genome [52]. The division of gene expression after gene duplication appears to be a general form of subfunctionalization [53, 54]. In this model, after gene duplication, complementary degenerate mutations are fixed randomly underrelaxed functional constraints [55]. Previous studies suggested that both *scd1* and *scd5* encode the same functional delta-9 desaturase and are localized on endoplasmic reticulum (ER) membrane [3, 56]. However, the *scd1* gene expressed ubiquitous, and *scd5*

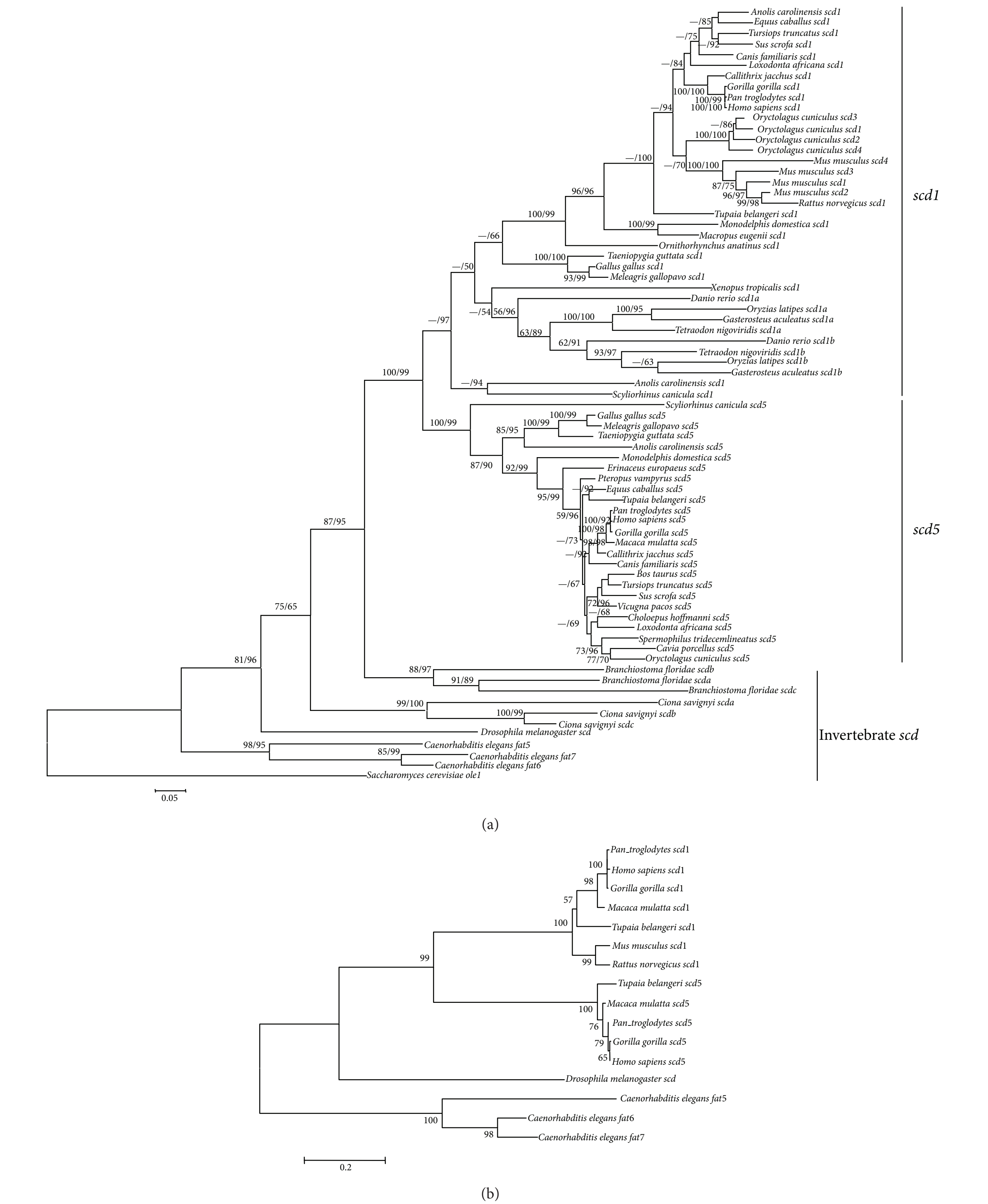


Figure 3: Phylogenetic trees of eukaryote *scd* isoforms. (a) Phylogenetic trees based on the nucleotide sequence data. The numbers on nodes indicated the support values, the former number was calculated using PHYML, and the latter number was calculated by FastTree 2.1.3. If bootstrap values were less than 50%, they were defaulted. Trees were rooted by yeast *ole1* gene. (b) Phylogenetic trees based on the amino acid sequences of 9 model animals with MEGA 4.0. The numbers on nodes indicated the support values. If bootstrap values were less than 50%, they were defaulted. Trees were rooted by *C. elegans* SCD paralogs.

Table 1: Selective patterns of scd genes estimated in CODEML.

Model	ln L	Parameters estimates	$2\Delta L$	Positively selected sites
		Branch-specific models		
M0	−20340.727636	$\omega = 0.08684$		
Model B				
Scd1 two ratio	−20339.045042	$\omega_0 = 0.09207, \omega_1 = 0.07951$	3.365188	
Scd5 two ratio	−20318.455800	$\omega_0 = 0.06146, \omega_1 = 0.09735$	$44.542672^{###}$	
Model C Three ratio	−20315.506084	$\omega_{scd1} = 0.06140$ $\omega_{scd5} = 0.09198$ $\omega_{\text{invertebrate}\ scd} = 0.11788$	$50.443104^{###}$	
		Branch-site models		
Scd1				
Model A1	−20079.883939	$\omega_0 = 0.06891, \omega_1 = 1, \omega_2 = 1$ $P_0 = 0.85305, P_1 = 0.04177$		
M1a	−20178.290653	$\omega_0 = 0.07950, \omega_1 = 1$ $P_0 = 0.91928, P_1 = 0.08072$	$196.813428^{###}$	
Model A	−20079.883939	$\omega_0 = 0.06891, \omega_1 = 1, \omega_2 = 1$ $P_0 = 0.85305, P_1 = 0.04177$	0	108L^{**}, 109F^{**}, 201A^{**}, 206S, 212K^{**}, 247Y^{**}, 254A, 255I^{*}, 276K^{**}, 289V^{*}, 315P, 330Y, 339A
Scd5				
Model A1	−20129.099054	$\omega_0 = 0.07500, \omega_1 = 1, \omega_2 = 1$ $P_0 = 0.88996, P_1 = 0.05968$		
M1a	−20168.513281	$\omega_0 = 0.07951, \omega_1 = 1$ $P_0 = 0.91900, P_1 = 0.08100$	$78.828454^{###}$	
Model A	−20129.099054	$\omega_0 = 0.07500, \omega_1 = 1, \omega_2 = 1$ $P_0 = 0.88996, P_1 = 0.05968$	0	157A^{**}, 194P^{**}, 215M^{**}, 223P, 230I, 338A^{**}

$^{###}P < 0.001$; $^{##}0.001 < P < 0.01$;
$^{**}P > 0.99$; $^{*}P > 0.95$.

Table 2: Transcription factor binding sites predicted at the 5′UTR of *hscd5*.

Transcription factor	Binding sites	Position (*hscd1*)[$]	Position (*hscd5*)
C/EBPα	GMAAA	−219	−1061, −1648
AP1	TGACC	−204, −271	−580, −643
SP1	GGCGG	−304, −314, −551	−286, −946
NF-Y	CCAAT	−458, −501	−397, −976
NF-1	TTGGC	−459, −502	−395
SREBP1	TCACC	−517	−892[*]
PPARα	AAAG/GGTCA	−1186	−579[#]
T3R	GGTCA	−2228	−1223, −2245

T3R: tri-iodothyronine receptor; AP1: activator protein 1; NF-1/Y: nuclear factor 1/Y; SREBP1: sterol regulatory element binding protein; PPARα: peroxisome proliferator-activated receptor; C/EBPα: CAAT/enhancer binding protein.

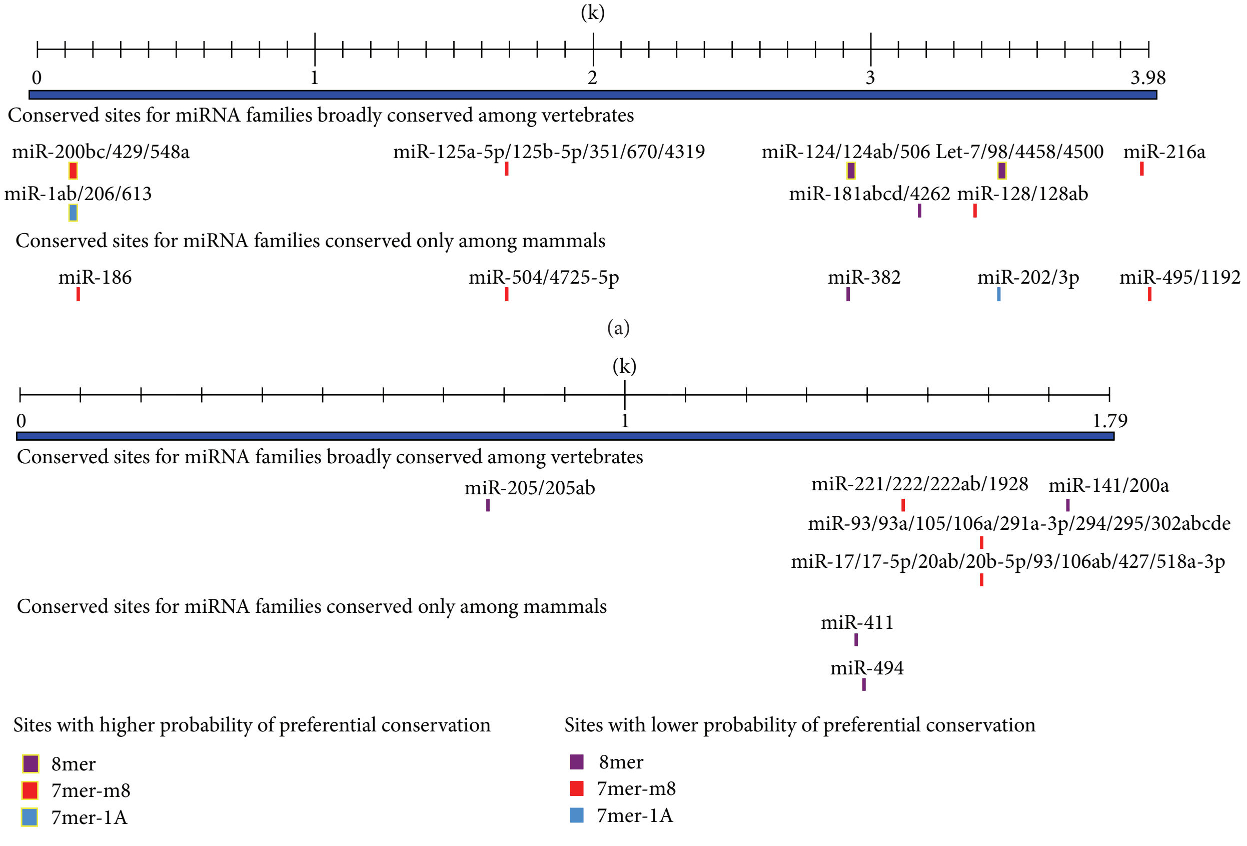
[$] These transcription factor binding sites were from [64].

[#] Predicted by PROMO.

[*] Only found in human using TRANSFAC.

gene expressed mainly in brain in different species [3, 6, 11]. We inferred that the evolution of *scd* genes might be a division of gene expression subsequent to gene duplication. This pattern was supported by the evolutionary forces behind the expression division of duplicate genes. The relative rate test suggested that the two duplicated *scd* genes evolved at the similar rate. The selective constraints analysis suggested that the *scd1* and *scd5* were both under strict purifying selection (Table 1), which was consistent with the conserved delta-9 desaturase of both *scd* genes. Intriguingly, in the branch-site analysis, we detected that some sites within *scd1* and *scd5* were underrelaxed selective pressure. These sites might be resulted from the random fixation of the complementary degenerate mutations that were underrelaxed functional constraints.

Though both of *scd1* and *scd5* encoded delta-9 desaturase, producing a palmitoleic acid (16:1n7) and oleic acid (18:1n9)


(k)
0
1
2
3
3.98
Conserved sites for miRNA families broadly conserved among vertebrates
miR-200bc/429/548a
miR-125a-5p/125b-5p/351/670/4319
miR-124/124ab/506
Let-7/98/4458/4500
miR-216a
miR-1ab/206/613
miR-181abcd/4262
miR-128/128ab
Conserved sites for miRNA families conserved only among mammals
miR-186
miR-504/4725-5p
miR-382
miR-202/3p
miR-495/1192
(a)
(k)
0
1
1.79
Conserved sites for miRNA families broadly conserved among vertebrates
miR-205/205ab
miR-221/222/222ab/1928
miR-141/200a
miR-93/93a/105/106a/291a-3p/294/295/302abcde
miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p
Conserved sites for miRNA families conserved only among mammals
miR-411
miR-494
Sites with higher probability of preferential conservation
8mer
7mer-m8
7mer-1A
Sites with lower probability of preferential conservation
8mer
7mer-m8
7mer-1A

(b)

Figure 4: The target sites for miRNA families conserved among mammals and vertebrates at the 3′-UTR region of human *scd1* gene (a) and *scd5* gene (b). The sites with different probability of preferential conservation were marked in different colors. The target sites sharing among miRNAs separated by slash were marked with same color.

[3, 56], they expressed diversely in the physiological process. Previous studies had proposed that *scd1* were associated with a variety of diseases including cancers, type 2 diabetes, and cardiovascular disorders [13], whereas *scd5* might act a potential role for maintaining the optimum levels of oleic acid in brain development and physiological activities [3, 57]. Castro et al. proposed that the major distinction between *scd5* and *scd1* would be at the regulatory level, in which *scd1* gene expression was mainly modulated at the transcriptional level by a wide variety of hormones and nutrients, whereas *scd5* was not responsive to external inputs like food sources [6]. In this study, we predicted the transcription factor binding sites at the 5′-UTR region and the miRNA target sites at the 3′-UTR region of human *scd5* gene. The transcription factor binding sites detected in *scd1* gene [13] could also be detected in *scd5* gene. However, the SREBP1 binding site only presents in human *scd5* gene, but not in other mammals, for example, rhesus monkey, pig and others. This might be that the prediction of transcription factor (TF) binding sites was based on known TF binding sites so that some new TF binding sites can not be detected. Recent studies have suggested that SREBP1 regulates the expression of *scd5* in primary cultures of human skeletal muscle cells [58], or directly binding to the promoter region of *scd5* in bovine [59]. In contrast, a study on human hepatocyte cell line suggested that SREBP1 only binds to the *scd1* gene, but not to *scd5* gene [31]. This discrepancy might be the distinct expression of *scd5* gene in different species or tissues. From our prediction, we conclude that the TF binding sites predicted in *scd5* gene were very similar to these of *scd1* gene, suggesting that the regulators may also be similar between two *scd* genes. Certainly, these TF predictions need further experimental verification.

miRNAs regulation is another gene regulatory mechanism in posttranscriptional regulation. Gu et al. estimated the time of vertebrate miRNA duplication events and suggested that gene/genome duplications in the early stage of vertebrates may expand the protein-encoding genes and miRNAs simultaneously [60]. Gene duplication events, followed by subfunctionalization and neofunctionalization processes, are considered to be a major source for emergence of novel miRNA genes [61]. In this study, the lengths of 3′-UTR of *scd1* and *scd5* gene are about 4000 bp and 1790 bp, respectively (Figure 4). A previous study suggested that genes with longer 3′-UTRs are regulated by more distinct types of miRNAs [62]. In our analysis, 13 miRNAs targeting sites are detected in the 3′-UTR of *scd1* gene, while 7 miRNAs targeting sites are detected in the 3′-UTR of *scd5* gene. Additionally, the length changes of 3′-UTRs in these two *scd* genes might suggest a differentiation of the regulatory mechanisms. miRNAs predicted to target the 3′-UTR region of *scd1* gene are associated with breast cancers, hepatocellular carcinoma, and metabolic syndromes such as diabetes, NAFLD. However, most of the miRNAs predicted to target the 3′-UTR region of *scd5* gene are related to the neurogenic disease and pancreatic cancer; and only 2 microRNAs are associated with the NAFLD. This regulatory pattern might be due to the high expression of *scd5* gene in brain and pancreas [3]. Additionally, a recent study has reported that the *scd5* gene plays a key role in the regulation of the neuronal cell proliferation and differentiation [56]. These results might indicate that the expression of *scd5* is implicated in brain development and physiological activity.

In addition, we also investigated the size and order of exons of *scd* genes. We found that the *scd1* gene has an extra exon (exon1) compared to *scd5* gene (Figure 2). The first 45 amino acids of SCD1 were highly different from those of SCD5 (Figure 1). Though there is no histidine domain and transmembrane domain exists in this part of SCD1, about 30 residues constitute a motif responsible for the rapid degradation of SCD [63]. This result indicated that the degradation of two SCD might be very different. However, due to no information on the degradation of SCD5, the evolutionary changes of regulation on both *scd* genes and SCD proteins still need further investigation.

5. Conclusion

In summary, this study of evolutionary pattern of *scd* genes showed that *scd1* and *scd5* genes emerged due to the duplication event as well as that they may play different roles. We also detected that the *scd* genes were evolved at the similar rate and were under strictly purifying selection, consistent with the conserved function of delta-9 desaturase of both SCD. Furthermore, our study revealed several potentially adaptive amino acid changes, which might be resulted from the random fixation of the complementary degenerate mutations underrelaxed functional constraints. The prediction of transcriptional factor binding sites at the 5′-UTR and miRNAs at 3′-UTR of *scd* genes suggested that the regulators of *scd5* may be different from *scd1* gene, supportting the differentiation at the regulatory levels between *scd5* and *scd1*. These findings increase the current knowledge of evolutionary patterns and function of *scd* genes in eukaryote. Yet, further experimental investigations need to elucidate the regulation and function of *scd* genes, especially the *scd5* gene.

Conflict of Interests

All authors declared no conflict of interests.

Authors' Contribution

Xiaoyun Wu and Xiaoju Zou contributed equally to this paper.

Acknowledgments

The authors thank the anonymous referees for their constructive and helpful comments and suggestions. This work was supported by the Innovative Program of the Chinese Academy of Sciences (KSCX2-EW-R-12 and KSCX2-EW-J-23) to Bin Liang, the National 863 Project of China (2012AA021801 and 2012AA022402), the National Natural Science Foundation of China (Y301261041, U1202223, 31160216, and 31171134), and the West Light Foundation to Xiaoyun Wu from the Chinese Academy of Sciences, and State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (GREKF13-03).

References

[1] F. S. Heinemann and J. Ozols, "Stearoyl-CoA desaturase, a short-lived protein of endoplasmic reticulum with multiple control mechanisms," *Prostaglandins Leukotrienes and Essential Fatty Acids*, vol. 68, no. 2, pp. 123–133, 2003.

[2] P. Strittmatter, L. Spatz, and D. Corcoran, "Purification and properties of rat liver microsomal stearyl coenzyme A desaturase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 71, no. 11, pp. 4565–4569, 1974.

[3] J. Wang, L. Yu, R. E. Schmidt et al., "Characterization of HSCD5, a novel human stearoyl-CoA desaturase unique to primates," *Biochemical and Biophysical Research Communications*, vol. 332, no. 3, pp. 735–742, 2005.

[4] C. M. Paton and J. M. Ntambi, "Biochemical and physiological function of stearoyl-CoA desaturase," *American Journal of Physiology—Endocrinology and Metabolism*, vol. 297, no. 1, pp. E28–E37, 2009.

[5] M. T. Flowers and J. M. Ntambi, "Role of stearoyl-coenzyme A desaturase in regulating lipid metabolism," *Current Opinion in Lipidology*, vol. 19, no. 3, pp. 248–256, 2008.

[6] L. F. C. Castro, J. M. Wilson, O. Gonçalves, S. Galante-Oliveira, E. Rocha, and I. Cunha, "The evolutionary history of the stearoyl-CoA desaturase gene family in vertebrates," *BMC Evolutionary Biology*, vol. 11, no. 1, article 132, 2011.

[7] J. L. Watts and J. Browse, "A palmitoyl-CoA-specific Δ9 fatty acid desaturase from Caenorhabditis elegans," *Biochemical and Biophysical Research Communications*, vol. 272, no. 1, pp. 263–269, 2000.

[8] T. J. Brock, J. Browse, and J. L. Watts, "Genetic regulation of unsaturated fatty acid composition in C. elegans," *PLoS Genetics*, vol. 2, no. 7, article e108, 2006.

[9] T. J. Brock, J. Browse, and J. L. Watts, "Fatty acid desaturation and the regulation of adiposity in Caenorhabditis elegans," *Genetics*, vol. 176, no. 2, pp. 865–875, 2007.

[10] J. E. Stukey, V. M. McDonough, and C. E. Martin, "Isolation and characterization of OLE1, a gene affecting fatty acid desaturation from Saccharomyces cerevisiae," *The Journal of Biological Chemistry*, vol. 264, no. 28, pp. 16537–16544, 1989.

[11] A. J. Lengi and B. A. Corl, "Comparison of pig, sheep and chicken SCD5 homologs: evidence for an early gene duplication event," *Comparative Biochemistry and Physiology B*, vol. 150, no. 4, pp. 440–446, 2008.

[12] H. Evans, T. De Tomaso, M. Quail et al., "Ancient and modern duplication events and the evolution of stearoyl-CoA desaturases in teleost fishes," *Physiological Genomics*, vol. 35, no. 1, pp. 18–29, 2008.

[13] D. Mauvoisin and C. Mounier, "Hormonal and nutritional regulation of SCD1 gene expression," *Biochimie*, vol. 93, no. 1, pp. 78–86, 2011.

[14] T. A. Hall, "Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/Nt," *Nucleic Acids Symposium Series*, vol. 41, pp. 95–98, 1999.

[15] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[16] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2—approximately maximum-likelihood trees for large alignments," *PLoS ONE*, vol. 5, no. 3, Article ID e9490, 2010.

[17] J. Felsenstein, "Confidence limits on phylogenies: an approach using bootstrap," *Evolution*, vol. 39, pp. 783–791, 1985.

[18] S. Guindon, F. Delsuc, J.-F. Dufayard, and O. Gascuel, "Estimating maximum likelihood phylogenies with PhyML," *Methods in Molecular Biology*, vol. 537, pp. 113–137, 2009.

[19] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0," *Systematic Biology*, vol. 59, no. 3, pp. 307–321, 2010.

[20] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[21] X. Messeguer, R. Escudero, D. Farré, O. Núñez, J. Martínez, and M. M. Albà, "PROMO: detection of known transcription regulatory elements using species-tailored searches," *Bioinformatics*, vol. 18, no. 2, pp. 333–334, 2002.

[22] D. Farré, R. Roset, M. Huerta et al., "Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3651–3653, 2003.

[23] M. Robinson-Rechavi and D. Huchon, "RRTree: relative-rate tests between groups of sequences on a phylogenetic tree," *Bioinformatics*, vol. 16, no. 3, pp. 296–297, 2000.

[24] Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 32–43, 2000.

[25] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[26] Z. Yang, "Inference of selection from multiple species alignments," *Current Opinion in Genetics and Development*, vol. 12, no. 6, pp. 688–694, 2002.

[27] J. Zhang, R. Nielsen, and Z. Yang, "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level," *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2472–2479, 2005.

[28] H. B. Sanchez, L. Yieh, and T. F. Osborne, "Cooperation by sterol regulatory element-binding protein and Sp1 in sterol regulation of low density lipoprotein receptor gene," *The Journal of Biological Chemistry*, vol. 270, no. 3, pp. 1161–1169, 1995.

[29] S. M. Jackson, J. Ericsson, R. Mantovani, and P. A. Edwards, "Synergistic activation of transcription by nuclear factor Y and sterol regulatory element binding protein," *Journal of Lipid Research*, vol. 39, no. 4, pp. 767–776, 1998.

[30] K. A. Dooley, S. Millinder, and T. F. Osborne, "Sterol regulation of 3-hydroxy-3-methylglutaryl-coenzyme A synthase gene through a direct interaction between sterol regulatory element binding protein and the trimeric CCAAT-binding factor/nuclear factor Y," *The Journal of Biological Chemistry*, vol. 273, no. 3, pp. 1349–1356, 1998.

[31] B. D. Reed, A. E. Charos, A. M. Szekely, S. M. Weissman, and M. Snyder, "Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000133, 2008.

[32] J. A. Foekens, A. M. Sieuwerts, M. Smid et al., "Four miRNAs associated with aggressiveness of lymph node-negative, estrogen receptor-positive human breast cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 13021–13026, 2008.

[33] M. V. Iorio, M. Ferracin, C.-G. Liu et al., "MicroRNA gene expression deregulation in human breast cancer," *Cancer Research*, vol. 65, no. 16, pp. 7065–7070, 2005.

[34] N. Kondo, T. Toyama, H. Sugiura, Y. Fujii, and H. Yamashita, "MiR-206 expression is down-regulated in estrogen receptor α-positive human breast cancer," *Cancer Research*, vol. 68, no. 13, pp. 5004–5008, 2008.

[35] A. Lujambio, S. Ropero, E. Ballestar et al., "Genetic unmasking of an epigenetically silenced microRNA in human cancer cells," *Cancer Research*, vol. 67, no. 4, pp. 1424–1429, 2007.

[36] X.-H. Huang, Q. Wang, J.-S. Chen et al., "Bead-based microarray analysis of microRNA expression in hepatocellular carcinoma: MiR-338 is downregulated," *Hepatology Research*, vol. 39, no. 8, pp. 786–794, 2009.

[37] L. Gramantieri, M. Ferracin, F. Fornari et al., "Cyclin G1 is a target of miR-122a, a MicroRNA frequently down-regulated in human hepatocellular carcinoma," *Cancer Research*, vol. 67, no. 13, pp. 6092–6099, 2007.

[38] Y. Zhao, H.-L. Jia, H.-J. Zhou et al., "Identification of metastasis-related microRNAs of hepatocellular carcinoma in hepatocellular carcinoma cell lines by quantitative real time PCR," *Chinese Journal of Hepatology*, vol. 17, no. 7, pp. 526–530, 2009.

[39] S. Volinia, G. A. Calin, C. G. Liu et al., "A microRNA expression signature of human solid tumors defines cancer gene targets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 7, pp. 2257–2261, 2006.

[40] J. L. Eun, Y. Gusev, J. Jiang et al., "Expression profiling identifies microRNA signature in pancreatic cancer," *International Journal of Cancer*, vol. 120, no. 5, pp. 1046–1054, 2007.

[41] L. Zheng, G.-C. Lv, J. Sheng, and Y.-D. Yang, "Effect of miRNA-10b in regulating cellular steatosis level by targeting PPAR-α expression, a novel mechanism for the pathogenesis of NAFLD," *Journal of Gastroenterology and Hepatology*, vol. 25, no. 1, pp. 156–163, 2010.

[42] A. Alisi, L. Da Sacco, G. Bruscalupi et al., "Mirnome analysis reveals novel molecular determinants in the pathogenesis of diet-induced nonalcoholic fatty liver disease," *Laboratory Investigation*, vol. 91, no. 2, pp. 283–293, 2011.

[43] N. Klöting, S. Berthold, P. Kovacs et al., "MicroRNA expression in human omental and subcutaneous adipose tissue," *PLoS ONE*, vol. 4, no. 3, Article ID e4699, 2009.

[44] M. Kato, S. Putta, M. Wang et al., "TGF-β activates Akt kinase through a microRNA-dependent amplifying circuit targeting PTEN," *Nature Cell Biology*, vol. 11, no. 7, pp. 881–889, 2009.

[45] K. Abu-Elneel, T. Liu, F. S. Gazzaniga et al., "Heterogeneous dysregulation of microRNAs across the autism spectrum," *Neurogenetics*, vol. 9, no. 3, pp. 153–161, 2008.

[46] B. Malzkorn, M. Wolter, F. Liesenberg et al., "Identification and functional characterization of microRNAs involved in the malignant progression of gliomas," *Brain Pathology*, vol. 20, no. 3, pp. 539–550, 2010.

[47] D. O. Perkins, C. D. Jeffries, L. F. Jarskog et al., "MicroRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder," *Genome Biology*, vol. 8, no. 2, article R27, 2007.

[48] M. Bloomston, W. L. Frankel, F. Petrocca et al., "MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from normal pancreas and chronic pancreatitis," *The Journal of the American Medical Association*, vol. 297, no. 17, pp. 1901–1908, 2007.

[49] J.-K. Park, E. J. Lee, C. Esau, and T. D. Schmittgen, "Antisense inhibition of microRNA-21 or -221 arrests cell cycle, induces apoptosis, and sensitizes the effects of gemcitabine in pancreatic adenocarcinoma," *Pancreas*, vol. 38, no. 7, pp. e190–e199, 2009.

[50] T. Greither, L. F. Grochola, A. Udelnow, C. Lautenschläger, P. Würl, and H. Taubert, "Elevated expression of microRNAs 155, 203, 210 and 222 in pancreatic tumors is associated with poorer survival," *International Journal of Cancer*, vol. 126, no. 1, pp. 73–80, 2010.

[51] J. Z. Zhang, "Evolution by gene duplication: an update," *Trends in Ecology and Evolution*, vol. 18, no. 6, pp. 292–298, 2003.

[52] M. A. Nowak, M. C. Boerlijst, J. Cooke, and J. M. Smith, "Evolution of genetic redundancy," *Nature*, vol. 388, no. 6638, pp. 167–171, 1997.

[53] A. Wagner, "Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 12, pp. 6579–6584, 2000.

[54] Z. Gu, D. Nicolae, H. H.-S. Lu, and W.-H. Li, "Rapid divergence in expression between duplicate genes inferred from microarray data," *Trends in Genetics*, vol. 18, no. 12, pp. 609–613, 2002.

[55] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y.-L. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[56] D. I. Sinner, G. J. Kim, G. C. Henderson, and R. A. Igal, "StearoylCoA desaturase-5: a novel regulator of neuronal cell proliferation and differentiation," *PLoS One*, vol. 7, no. 6, Article ID e39787, 2012.

[57] A. J. Lengi and B. A. Corl, "Identification and characterization of a novel bovine stearoyl-CoA desaturase isoform with homology to human SCD5," *Lipids*, vol. 42, no. 6, pp. 499–508, 2007.

[58] S. Rome, V. Lecomte, E. Meugnier et al., "Microarray analyses of SREBP-1a and SREBP-1c target genes identify new regulatory pathways in muscle," *Physiological Genomics*, vol. 34, no. 3, pp. 327–337, 2008.

[59] A. J. Lengi and B. A. Corl, "Regulation of the bovine SCD5 promoter by EGR2 and SREBP1," *Biochemical and Biophysical Research Communications*, vol. 421, no. 2, pp. 375–379, 2012.

[60] X. Gu, Z. Su, and Y. Huang, "Simultaneous expansions of micrornas and protein-coding genes by gene/genome duplications in early vertebrates," *Journal of Experimental Zoology B*, vol. 312, no. 3, pp. 164–170, 2009.

[61] E. Berezikov, "Evolution of microRNA diversity and regulation in animals," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 846–860, 2011.

[62] C. Cheng, N. Bhardwaj, and M. Gerstein, "The relationship between the evolution of microRNA targets and the length of their UTRs," *BMC Genomics*, vol. 10, article 431, 2009.

[63] H. Mziaut, G. Korza, and J. Ozols, "The N terminus of microsomal Δ9 stearoyl-CoA desaturase contains the sequence determinant for its rapid degradation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 16, pp. 8883–8888, 2000.

[64] L. Zhang, L. Ge, T. Tran, K. Stenn, and S. M. Prouty, "Isolation and characterization of the human stearoyl-CoA desaturase gene promoter: requirement of a conserved CCAAT cis-element," *Biochemical Journal*, vol. 357, part 1, pp. 183–193, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 347242, 6 pages
http://dx.doi.org/10.1155/2013/347242

Research Article

Phylogenetic Relationships of *Pseudorasbora*, *Pseudopungtungia*, and *Pungtungia* (Teleostei; Cypriniformes; Gobioninae) Inferred from Multiple Nuclear Gene Sequences

Keun-Yong Kim,[1] Myeong-Hun Ko,[2] Huanzhang Liu,[3] Qiongying Tang,[3] Xianglin Chen,[4] Jun-Ichi Miyazaki,[5] and In-Chul Bang[2]

[1] *Department of Research and Development, NLP Co., Ltd., Busan 619-912, Republic of Korea*
[2] *Department of Life Sciences & Biotechnology, Soonchunhyang University, Asan 336-745, Republic of Korea*
[3] *Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*
[4] *School of Life Science, South China Normal University, Guangzhou 510631, China*
[5] *Faculty of Education and Human Sciences, University of Yamanashi, Yamanashi 400-8510, Japan*

Correspondence should be addressed to In-Chul Bang; incbang@sch.ac.kr

Received 1 March 2013; Accepted 7 August 2013

Academic Editor: William Piel

Gobionine species belonging to the genera *Pseudorasbora*, *Pseudopungtungia*, and *Pungtungia* (Teleostei; Cypriniformes; Cyprinidae) have been heavily studied because of problems on taxonomy, threats of extinction, invasion, and human health. Nucleotide sequences of three nuclear genes, that is, recombination activating protein gene 1 (*rag1*), recombination activating gene 2 (*rag2*), and early growth response 1 gene (*egr1*), from *Pseudorasbora*, *Pseudopungtungia*, and *Pungtungia* species residing in China, Japan, and Korea, were analyzed to elucidate their intergeneric and interspecific phylogenetic relationships. In the phylogenetic tree inferred from their multiple gene sequences, *Pseudorasbora*, *Pseudopungtungia* and *Pungtungia* species ramified into three phylogenetically distinct clades; the "*tenuicorpa*" clade composed of *Pseudopungtungia tenuicorpa*, the "*parva*" clade composed of all *Pseudorasbora* species/subspecies, and the "*herzi*" clade composed of *Pseudopungtungia nigra*, and *Pungtungia herzi*. The genus *Pseudorasbora* was recovered as monophyletic, while the genus *Pseudopungtungia* was recovered as polyphyletic. Our phylogenetic result implies the unstable taxonomic status of the genus *Pseudopungtungia*.

1. Introduction

Species of the subfamily Gobioninae (or gudgeons) (Teleostei; Cypriniformes; Cyprinidae) are mostly distributed in East Asia [1–4] except several species belonging to the genera *Gobio* and *Romanogobio* in Europe [5]. Among them, the slender topmouth gudgeon *Pseudorasbora elongata* endemic to China faces a high risk of extinction due to habitat degradation and loss and fishing [6]. *Pseudorasbora interrupta* was recently erected as novel species [7]. The topmouth gudgeon (or the stone morocco) *Pseudorasbora parva*, which was first reported from Nagasaki, Japan, is widely distributed in East Asia [1] and has rapidly extended its habitats either naturally or artificially to all of Europe and parts of North Africa during the last 50 years [8, 9]. Moreover, this freshwater fish is notorious as the second intermediate host of the liver fluke *Clonorchis sinensis* [10] and is a carrier of the rosette agent (*Sphaerothecum destruens*) which inhibits spawning and causes increased mortality in native European fish species [11]. *Pseudorasbora pumila pumila* and *Pseudorasbora pumila* subsp. originally inhabited limited areas in northern and middle parts on Honshu of Japan, respectively, but their distributions have been further restricted to patchily discrete locations due to loss of their habitats and invasion of *Pseudorasbora parva* into their habitats. Thus, the Japanese Ministry of the Environment designated them as critically endangered subspecies. The striped shiner *Pungtungia herzi* Herzenstein, which was first reported from Chungju, Korea [1, 4, 12], resides in China, Japan, and Korea [1]. The black shiner *Pseudopungtungia nigra* endemic to Korea was reported as a novel genus and species by Mori [13]. This species was reported to inhabit the Geum River, the Mangyeong River, and the Ungcheon Stream in Korea [14, 15] but believed to be regionally extinct in the latter due to water impoundment and pollution [4, 16]. Because of the threat of extinction, it was designated as an endangered species in 1997 by the Ministry of Environment of Korea and protected by national legislation. The slender shiner *Pseudopungtungia tenuicorpa* endemic to Korea, inhabits the upper reaches of the Han and Imjin Rivers [4, 16]. This fish species was also designated as an endangered species since 2005 due to deterioration of its natural habitats.

Despite great concerns on conservation the taxonomic positions of *Pseudorasbora elongata*, *Pseudopungtungia nigra,* and *Pseudopungtungia tenuicorpa* are still unsettled. For example, it was interesting that *Pseudorasbora elongata* showed closer phylogenetic affiliation to *Pungtungia herzi* rather than congeneric species based on the mitochondrially encoded cytochrome *b* gene (*mt-cyb*) sequences [17]. Meanwhile, Kang [18] suggested transferring two *Pseudopungtungia* species to the genus *Pungtungia* based on synapomorphic osteological characters such as jaws supporting the form of mouth, suspensorial elements, and hyoid arch, but they still remain in the former genus [4].

Despite problems with taxonomy, threats of extinction, invasion, and human health, there are not enough molecular data for *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* species to provide compelling answers to questions about their intergeneric and interspecific phylogenetic relationships (e.g., Yang et al. [17]), genetic variation for conservation (e.g., Konishi and Takata [19]), phylogeography (e.g., Watanabe et al. [20]), divergence time estimation (e.g., Liu et al. [21]), and developing monitoring markers for tracing their dispersal route. The genetic data available for those species up to date are mostly composed of nucleotide sequences from a single mitochondrially encoded gene: the *mt-cyb*, which is maternally inherited and thus provides insufficient evidence for resolving their phylogenetic relationships.

Recently, phylogenetic markers of nuclear genes were deciphered and successfully applied for reconstructing phylogenetic trees across diverse Cypriniform species [22–25]. In this study, we analyzed multiple nuclear gene sequences of eight species and subspecies of *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* residing in China, Japan, and Korea to elucidate their molecular phylogenetic relationships.

2. Materials and Methods

2.1. Specimen and Genomic DNA Extraction. Fish specimens used in this study were captured with a spoon net (mesh size: 4 × 4 mm) from river drainages of China, Japan, and Korea. The specimens we used were transported to the laboratory alive and killed rapidly with formaldehyde after anaesthetizing them by submerging into a solution containing a fish anaesthetic agent, Tricaine Methane Sulphonate (MS222) (Aqualife TMS, Syndel Laboratories, Ltd., Canada). The specimens were deposited in the fish collection of Soonchunhyang University (SUC; Asan, Republic of Korea), Chinese Academy of Sciences (Wuhan, China), and South China Normal University (Guangzhou, China). Their detailed sampling information was provided in Table 1.

A small piece of a pectoral or anal fin tissue was excised from each specimen to extract genomic DNA (gDNA). It was incubated in 500 μL of TNES-Urea buffer (10 mM Tris-HCl, pH 8.0; 125 mM NaCl; 10 mM EDTA, pH 8.0; 1% SDS; 6 M urea; [26]) containing 100 μg of proteinase K (Sigma-Aldrich, St. Louis, MO, USA) at 37°C for a week, followed by separation with phenol : chloroform : isoamyl alcohol (25 : 24 : 1) solution and ethanol precipitation. The extracted gDNA was finally resuspended in 50 μL of TE buffer (10 mM Tris-HCl, pH 8.0; 1 mM EDTA, pH 8.0). Its quantity and quality were checked using a spectrophotometer, NanoDrop 1000 (Thermo Fisher Scientific, Wilmington, DE, USA) and by electrophoresis in a 0.7% agarose gel after staining with GelRed Nucleic Acid Gel Stain (Biotium, Hayward, CA, USA).

2.2. PCR Amplification and Sequencing. For phylogenetic analysis, three nuclear genes, that is, recombination activating gene 1 (*rag1*), recombination activating gene 2 (*rag2*), and early growth response 1 gene (*egr1*), were selected based on previous studies [23, 24]. Information for the primers used in this study is shown in Table 2. PCR reactions were carried out in a 20 μL reaction volume using *AccuPower* PCR Premix (Bioneer, Daejeon, Republic of Korea), including 50 ng of gDNA and 0.2 μM of forward and reverse primers.

PCR was run with the following thermal cycling profile in a DNA Engine DYAD Peltier Thermal Cycler (MJ Research Inc., Waltham, MA, USA): an initial denaturation at 94°C for 3 min, 25–35 cycles of denaturation at 94°C for 30 s, annealing at 50–52°C for 30 s, and elongation at 72°C for 1 min. The reaction was completed with a final elongation at 72°C for 7 min. The PCR product was purified with the *AccuPrep* PCR Purification Kit (Bioneer). After cycle sequencing with the ABI PRISM BigDye Terminator v3.1 Cycle Sequencing Ready Reaction Kit (Applied Biosystems Inc., Foster City, CA, USA), the purified product was directly sequenced on an ABI 3730xl DNA Analyzer (Applied Biosystems Inc.) with PCR primers by a commercial company, Macrogen Inc. (Seoul, Republic of Korea). Electropherograms were assembled in BioEdit 7.0.5 [28] and corrected manually. The sequences analyzed in this study were deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers KF468594-KF468626 (Table 1).

2.3. Phylogenetic Analyses. Nucleotide sequences of the *rag1*, *rag2,* and *egr1* genes of eight *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* species analyzed in this study (Table 1) were aligned with ClustalW in BioEdit [28]. Two *Sarcocheilichthys* species were used as outgroups, based on previous molecular phylogenetic [17, 25, 29] and morphological [18] studies. The three nuclear genes were concatenated according to genes. There were no indels in the nucleotide matrix that consisted of 1,488, 1,120, and 1,087 bp for each

Table 1: Sampling information of gobionine species used in the phylogenetic analyses.

Species	Voucher no.	Sampling site	Drainage	GenBank acc. no.		
				rag1	*rag2*	*egr1*
Pseudopungtungia nigra	SUC-0777	Geumsan, Korea	Geum River	KF468619	KF468608	KF468597
Pseudopungtungia tenuicorpa	SUC-0683	Yangpyeong, Korea	Han River	KF468618	KF468607	KF468596
*Pseudorasbora elongata**	—	China	—	KF468621	KF468610	KF468599
*Pseudorasbora interrupta**	—	China	—	KF468623	KF468612	KF468601
Pseudorasbora parva	SUC-6658	Kasumigaura, Ibaraki Pref., Japan	A small tributary of Lake Kasumigaura	KF468626	KF468615	KF468604
Pseudorasbora pumila pumila	SUC-6656	Nagano, Nagano Pref., Japan	An irrigative pond	KF468624	KF468613	KF468602
Pseudorasbora pumila subsp.	SUC-6657	Bred in Lake Biwa Museum, Shiga Pref., Japan	—	KF468625	KF468614	KF468603
Pungtungia herzi	SUC-0706	Yangpyeong, Korea	Han River	KF468620	KF468609	KF468598
Sarcocheilichthys nigripinnis morii	SUC-1833	Seocheon, Korea	Gilsan Stream	KF468617	KF468606	KF468595
Sarcocheilichthys variegatus wakiyae	SUC-0774	Geumsan, Korea	Geum River	KF468616	KF468605	KF468594

*Detailed information is not provided by the authors for protecting their natural habitats.

Table 2: Information of PCR primers used in this study.

Gene	Primer	Sequence (5′ → 3′)	Reference
Recombination activating gene 1 (*rag1*)	RAG1-1495f3	CAGTAYCAYAAGATGTACCG	Kim and Bang [27]
	RAG1-3067r	TTGTGAGCYTCCATRAACTT	Kim and Bang [27]
Recombination activating gene 2 (*rag2*)	RAG2-108f	CCVARACGCTCATGTCCAAC	This study
	RAG2-1324r	TGGARCAGWAGATCATKGC	This study
Early growth response 1 gene (*egr1*)	EGR1-291f	CACAGGMCGTTTCACCCTYG	Modified from Chen et al. [24]
	EGR1-1456r	GACAGGRGARCTGTAGATGTT	Modified from Chen et al. [24]

gene, respectively. The nucleotide matrix is available upon request.

Maximum likelihood (ML) analysis was performed with RAxML 7.0.4 [30, 31]. The concatenated nucleotide matrix was partitioned according to genes. The RAxML search was executed for the best-scoring ML tree in one single program run (the "-f a" option) instead of the default maximum parsimony starting tree. The best-scoring ML tree of a thorough ML analysis was determined under the GTRMIX model in 200 inferences. Statistical support was evaluated with 1,000 nonparametric bootstrap inferences.

Bayesian inference (BI) analysis was carried out in MrBayes 3.1.2 [32] after partitioning the nucleotide matrix according to genes. MrModeltest 2.3 [33] in PAUP* 4.0b10 [34] was used to determine the best-fit evolutionary model by Akaike Information Criterion (AIC) for each gene and selected the SYM+Γ, K80+I, and HKY+I models, for the *rag1*, *rag2,* and *egr1* genes, respectively. All model parameters were unlinked across partitions, and all partitions were allowed to have different rates. Two independent Metropolis-coupled Markov chain Monte Carlo (MCMCMC) runs were performed with four simultaneous chains (three heated and one cold) and random starting trees for 5,000,000 generations, sampling parameters, and topologies every 100 generations. Burn-in was determined by checking the convergence of likelihood values across MCMCMC. A total of 500 out of 50,001 resulting trees were discarded as "burn-in." The last trees after convergence were used to construct a 50% majority-rule consensus tree and to summarize posterior probability support for each node.

3. Results

ML and BI trees inferred from the multiple nuclear gene sequences generated identical tree topologies. In the phylogenetic tree, species belonging to the genera *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* formed a monophyletic group with the highest level of confidence with respect to the *Sarcocheilichthys* outgroups (Figure 1).

In the phylogenetic tree, *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* species were ramified into three distinct clades; the "*tenuicorpa*" clade composed of a single species, *Pseudopungtungia tenuicorpa*, the "*herzi*" clade of *Pseudorasbora nigra* and *Pungtungia herzi*, and the "*parva*" clade of *Pseudorasbora elongata*, *Pseudorasbora interrupta*,

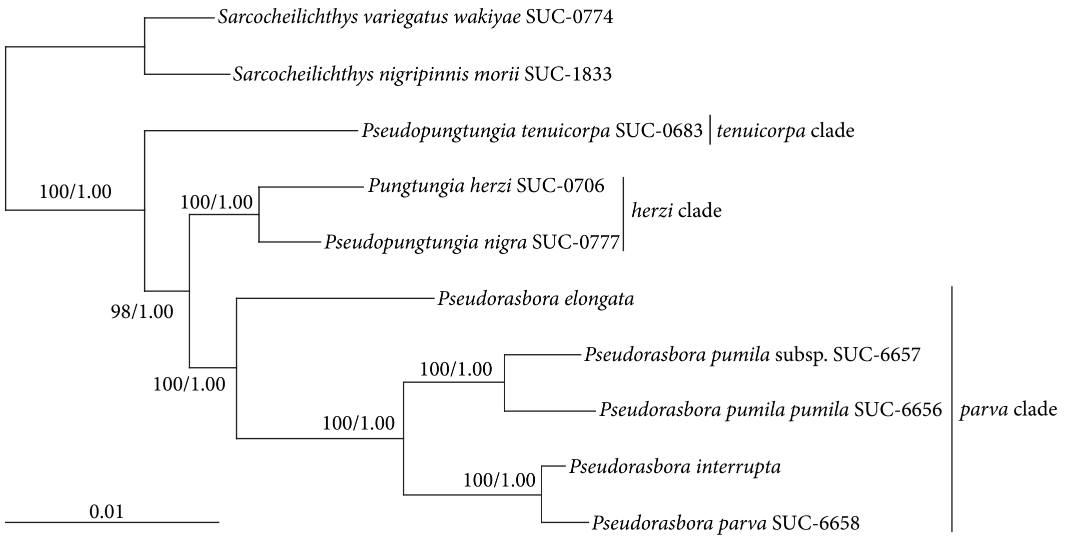

Sarcocheilichthys variegatus wakiyae SUC-0774
Sarcocheilichthys nigripinnis morii SUC-1833
Pseudopungtungia tenuicorpa SUC-0683
tenuicorpa clade
100/1.00
100/1.00
Pungtungia herzi SUC-0706
herzi clade
Pseudopungtungia nigra SUC-0777
98/1.00
Pseudorasbora elongata
100/1.00
Pseudorasbora pumila subsp. SUC-6657
100/1.00
Pseudorasbora pumila pumila SUC-6656
parva clade
100/1.00
Pseudorasbora interrupta
100/1.00
0.01
Pseudorasbora parva SUC-6658

Figure 1: Maximum likelihood (ML) trees of gobionine species belonging to the genera *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* inferred from multiple nuclear genes, that is, the recombination activating gene 1 (*rag1*), recombination activating gene 2 (*rag2*), and early growth response 1 gene (*egr1*). ML and Bayesian inference (BI) trees were reconstructed after partitioning the concatenated nucleotide matrix according to genes. Bootstrap values above 50% of ML analysis and posterior probabilities above 0.90 of BI analysis were shown at each branch node. The scale bar indicates substitutions/site.

Pseudorasbora parva and *Pseudorasbora pumila pumila*, *Pseudorasbora pumila* subsp. (Figure 1). Among those clades, "*tenuicorpa*" clade placed at the basal position, giving rise to two ramifying "*herzi*" and "*parva*" clades, supported by 98% bootstrap value in ML tree and 1.00 posterior probability in BI tree. The "*herzi*" and "*parva*" clades were supported with the highest statistical supports. Within the "*parva*" clade, *Pseudorasbora elongata* formed the sister-group relationship to the lineage composed of *Pseudorasbora pumila pumila*, *Pseudorasbora pumila* subsp., *Pseudorasbora interrupta,* and *Pseudorasbora parva*. The former two consistently separated from the latter two.

4. Discussion

In the phylogenetic tree, the genus *Pseudorasbora* was recovered as monophyletic, but the genus *Pseudopungtungia* was recovered as polyphyletic; the monotypic *Pungtungia herzi* was closely affiliated to *Pseudopungtungia nigra* with highest statistical supports, and *Pseudopungtungia tenuicorpa* was placed at the basal position among *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* species.

Previous molecular phylogenetic studies inferred from nuclear or mitochondrial gene sequences [17, 25, 28] clearly revealed the monophyly of the genera *Pseudorasbora* and *Pungtungia*. However, those studies did not include a closely related genus (i.e., *Pseudopungtungia*) and all *Pseudorasbora* species (i.e., *Pseudorasbora interrupta* and both *Pseudorasbora pumila* subspecies). Overall tree topologies generated in this study after including all those species revealed the clear monophyletic nature of the three genera *Pseudorasbora*, *Pungtungia,* and *Pseudopungtungia* from China, Japan, and Korea with respect to *Sarcocheilichthys* outgroups. This is completely or partially congruent with previous phylogenetic assumptions based on osteology [18, 35] and anatomy (vertebral formula; [36]). Meanwhile, our phylogenetic trees recovered the genus *Pseudorasbora* as monophyletic and the genus *Pseudopungtungia* as polyphyletic. *Pseudopungtungia nigra* showed the closest phylogenetic affiliation to *Pungtungia herzi*, and *Pseudopungtungia tenuicorpa* was clearly separated not only from those two species but also from the five *Pseudorasbora* species and subspecies.

In accordance with the polyphyletic nature of the genus *Pseudopungtungia*, Kim [37] mentioned significant morphological differences between *Pseudopungtungia nigra* and *Pseudopungtungia tenuicorpa* in the body shape and crossbars in fins except the mouth shape and the unstable taxonomic status of the genus *Pseudopungtungia*. Mori [13] morphologically differentiated *Pseudopungtungia nigra* from *Pungtungia herzi* by the mouth shape and fin coloration and erected the former as a novel genus and species. However, Banarescu [38] described their similarities in the mouth shape, lips, and jaws that are congruent with our molecular phylogenetic result. The close relationship can also be explained by the occurrence of a natural hybrid between them [39]. Independently, Kim et al. [40] carried out the polyacrylamide gel electrophoresis of muscle proteins extracted from Korean gobionine species to investigate their systematic relationships. Their result revealed many similarities among species of *Coreoleuciscus*, *Pseudorasbora*, *Pseudopungtungia,* and *Pungtungia* and the close relationship between *Pseudopungtungia nigra* and *Pungtungia herzi* among them, which is also congruent with our results.

The monophyly of the genus *Pseudorasbora* reflected the current taxonomic classification, which includes *Pseudorasbora elongata*, *Pseudorasbora interrupta*, *Pseudorasbora* *parva*, *Pseudorasbora pumila pumila,* and *Pseudorasbora pumila* subsp. Banarescu and Nalbant [41] mentioned that *Pseudorasbora elongata* has a notably distinct taxonomic position from other *Pseudorasbora* species, because of its elongated body and snout and longitudinal blackish stripes as *Pungtungia*. This is congruent with our phylogenetic tree, because *Pseudorasbora elongata* was placed at the basal position separated from other *Pseudorasbora* species and subspecies. This is also congruent with the result of Yang et al. [17] based on the *mt-cyb* gene. However, Yang et al. [17] and Liu et al. [21] showed that *Pseudorasbora elongata* consistently clustered with *Pungtungia herzi* and clearly separated from congeneric *Pseudorasbora parva* and *Pseudorasbora pumila* in their *mt-cyb* trees. Xiao et al. [7] mentioned the close relationship of *Pseudorasbora interrupta* to *Pseudorasbora parva* and *Pseudorasbora pumila*, which is congruent with our phylogenetic tree. In this study, *Pseudorasbora interrupta* has a closer phylogenetic relationship to *Pseudorasbora parva* than the two *Pseudorasbora pumila* subspecies.

Our result shows the unstable taxonomic status of the genus *Pseudopungtungia* and suggests a novel genus should be erected to accommodate *Pseudopungtungia tenuicorpa* in a future taxonomic study. Besides resolving phylogenetic relationships, the nucleotide sequence information presented in this study will provide useful baseline data for developing recovery plans of endangered species and subspecies investigated in this study (i.e., *Pseudopungtungia nigra* and *Pseudopungtungia tenuicorpa*, *Pseudorasbora elongata* and *Pseudorasbora pumila* subsp.) because clarification of their phylogenetic positions is the prerequisite for such efforts.

Acknowledgments

The authors express their sincere thanks to Drs. Koji Tojo and Keisuke Takata (Shinshu University) and Dr. Masanari Matsuda (Lake Biwa Museum) for their assistance in collecting *Pseudorasbora pumila* and providing useful information on their biology. This work was supported by the Soonchunhyang University Research Fund.

References

[1] P. Banarescu and T. T. Nalbant, *Pisces, Teleostei, Cyprinidae (Gobioninae)*, vol. 93 of *Das Tierreich Lieferung*, 1973.

[2] Y. Chen, X. Chu, Y. Luo et al., *Fauna Sinica: Osteichthyes Cypriniformes II*, Science Press, Beijing, China, 1998, (Chinese).

[3] T. Nakabo, *Fishes of Japan with Pictorial Keys to the Species*, Tokai University Press, Tokyo, Japan, 2002.

[4] I.-S. Kim, Y. Choi, C.-L. Lee, Y.-J. Lee, B.-J. Kim, and J.-H. Kim, *Illustrated Book of Korean Fishes*, Kyohak Publishing, Seoul, Republic of Korea, 2005, (Korean).

[5] M. Nowak, J. Koščo, and W. Popek, "Review of the current status of systematics of gudgeons (Gobioninae, Cyprinidae) in Europe," *AACL Bioflux*, vol. 1, no. 1, pp. 27–38, 2008.

[6] K. DePing, C. GuiHua, and Y. JunXing, "Threatened fishes of the world: *Pseudorasbora elongata* Wu, 1939 (Cyprinidae)," *Environmental Biology of Fishes*, vol. 76, no. 1, pp. 69–70, 2006.

[7] Z. Xiao, Z.-H. Lan, and X.-L. Chen, "A new species of the genus *Pseudorasbora* from Guangdong Province, China (Cypriniformes, Cyprinidae)," *Acta Zootaxonomica Sinica*, vol. 32, no. 4, pp. 977–980, 2007.

[8] A. Witkowski, "NOBANIS—Invasive Alien Species Fact Sheet—*Pseudorasbora parva*—From: Online Database of the North European and Baltic Network on Invasive Alien Species—NOBANIS," 2006, http://www.nobanis.org.

[9] R. E. Gozlan, D. Andreou, T. Asaeda et al., "Pan-continental invasion of *Pseudorasbora parva*: towards a better understanding of freshwater fish invasions," *Fish and Fisheries*, vol. 11, no. 4, pp. 315–340, 2010.

[10] H.-J. Rim, "Clonorchiasis in Korea," *Korean Journal of Parasitology*, vol. 28, supplement, pp. 63–78, 1990.

[11] R. E. Gozlan, S. St.-Hilaire, S. W. Feist, P. Martin, and M. L. Kent, "Biodiversity: disease threat to European fish," *Nature*, vol. 435, no. 7045, p. 1046, 2005.

[12] S. M. Herzenstein, "Ichthyologische Bemerkungen aus dem Zoologischen Museum der Kaiserlichen Akademie Wissenschaften. III," *Mélanges Biologiques, Tirés du Bulletin Physico-Mathématique de l'Académie Impériale des Sciences de St. Pétersbourg*, vol. 13, part 2, pp. 219–235, 1892.

[13] T. Mori, "Descriptions of two new genera and seven new species of Cyprinidae from Chosen," *Annotations of Zoologicae Japonenses*, vol. 15, no. 2, pp. 161–181, 1935.

[14] K. C. Choi, "On the geographical distribution of freshwater fishes south of DMZ in Korea," *Korean Journal of Limnology*, vol. 6, no. 3, pp. 29–36, 1973 (Korean).

[15] S. R. Jeon, "Ecological studies on the *Pseudopungtungia nigra* from Korea," *Korean Journal of Limnology*, vol. 10, no. 1, pp. 33–46, 1977 (Korean).

[16] I.-S. Kim and J.-Y. Park, *Freshwater Fishes of Korea*, Kyohak Publishing, Seoul, Republic of Korea, 2002, (Korean).

[17] J. Yang, S. He, J. Freyhof, K. Witte, and H. Liu, "The phylogenetic relationships of the Gobioninae (Teleostei: Cyprinidae) inferred from mitochondrial cytochrome *b* gene sequences," *Hydrobiologia*, vol. 553, no. 1, pp. 255–266, 2006.

[18] E.-J. Kang, *Phylogenetic study on the subfamily gobioninae (Pisces: Cyprinidae) from Korea as evidenced by their comparative osteology and myology [Ph.D. thesis]*, Chonbuk National University, Jeonju, Republic of Korea, 1991, (Korean).

[19] M. Konishi and K. Takata, "Isolation and characterization of polymorphic microsatellite DNA markers in topmouth gudgeon, *Pseudorasbora* (Teleostei: Cyprinidae)," *Molecular Ecology Notes*, vol. 4, no. 1, pp. 64–66, 2004.

[20] K. Watanabe, K. Iguchi, K. Hosoya, and M. Nishida, "Phylogenetic relationships of the Japanese minnows, *Pseudorasbora* (Cyprinidae), as inferred from mitochondrial 16S rRNA gene sequences," *Ichthyological Research*, vol. 47, no. 1, pp. 43–50, 2000.

[21] H. Z. Liu, J. Q. Yang, and Q. Y. Tang, "Estimated evolutionary tempo of East Asian gobionid fishes (Teleostei: Cyprinidae) from mitochondrial DNA sequence data," *Chinese Science Bulletin*, vol. 55, no. 15, pp. 1501–1510, 2010.

[22] V. Šlechtová, J. Bohlen, and A. Perdices, "Molecular phylogeny of the freshwater fish family Cobitidae (Cypriniformes: Teleostei): delimitation of genera, mitochondrial introgression and evolution of sexual dimorphism," *Molecular Phylogenetics and Evolution*, vol. 47, no. 2, pp. 812–831, 2008.

[23] X. Wang, J. Li, and S. He, "Molecular evidence for the monophyly of East Asian groups of Cyprinidae (Teleostei: Cypriniformes) derived from the nuclear recombination activating

gene 2 sequences," *Molecular Phylogenetics and Evolution*, vol. 42, no. 1, pp. 157–170, 2007.

[24] W.-J. Chen, M. Miya, K. Saitoh, and R. L. Mayden, "Phylogenetic utility of two existing and four novel nuclear gene loci in reconstructing tree of life of ray-finned fishes: the order Cypriniformes (Ostariophysi) as a case study," *Gene*, vol. 423, no. 2, pp. 125–134, 2008.

[25] R. L. Mayden, K. L. Tang, R. M. Wood et al., "Inferring the tree of life of the order Cypriniformes, the earth's most diverse clade of freshwater fishes: implications of varied taxon and character sampling," *Journal of Systematics and Evolution*, vol. 46, no. 3, pp. 424–438, 2008.

[26] T. Asahida, T. Kobayashi, K. Saitoh, and I. Nakayama, "Tissue preservation and total DNA extraction from fish stored at ambient temperature using buffers containing high concentration of urea," *Fisheries Science*, vol. 62, no. 5, pp. 727–730, 1996.

[27] K.-Y. Kim and I.-C. Bang, "Molecular phylogenetic position of *Abbottina springeri* (Cypriniformes, Cyprinidae) based on nucleotide sequences of RAG1 gene," *Korean Journal of Ichthyology*, vol. 22, no. 4, pp. 273–278, 2010.

[28] T. A. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Symposium Series*, no. 41, pp. 95–98, 1999.

[29] K. Saitoh, T. Sado, R. L. Mayden et al., "Mitogenomic evolution and interrelationships of the Cypriniformes (Actinopterygii: Ostariophysi): the first evidence toward resolution of higher-level relationships of the world's largest freshwater fish clade based on 59 whole mitogenome sequences," *Journal of Molecular Evolution*, vol. 63, no. 6, pp. 826–841, 2006.

[30] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006.

[31] A. Stamatakis, P. Hoover, and J. Rougemont, "A rapid bootstrap algorithm for the RAxML web servers," *Systematic Biology*, vol. 57, no. 5, pp. 758–771, 2008.

[32] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[33] J. A. A. Nylander, *MrModeltest v2.2*, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden, 2004.

[34] D. L. Swofford, *"PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)," ver.4, Sinauer Associates, Sunderland, UK, 2002*.

[35] I.-S. Kim and E.-J. Kang, "Comparative study on the urohyal of the subfamily Gobioninae of Korea," *Korean Journal of Ichthyology*, vol. 1, no. 1, pp. 24–34, 1989 (Korean).

[36] A. M. Naseka, "Comparative study on the vertebral column in the Gobioninae (Cyprinidae, Pisces) with special reference to its systematics," *Publicaciones Especiales Instituto Español de Oceanografía*, vol. 21, pp. 149–167, 1996.

[37] I.-S. Kim, "The taxonomic study of gudgeons of the subfamily Gobioninae (Cyprinidae) in Korea," *Bulletin of Korean Fisheries Society*, vol. 17, no. 5, pp. 436–448, 1984.

[38] P. M. Banarescu, "A critical updated checklist of Gobioninae (Pisces, Cyprinidae)," *Travaux du Muséum d'Histoire Naturelle "Grigore Antipa"*, vol. 32, pp. 303–330, 1992.

[39] I.-S. Kim, Y. Choi, and J.-H. Shim, "An occurrence of intergeneric hybrid cross, *Pungtungia herzi* × *Pseudopungtungia nigra* from the Ungcheon River, Korea," *Korean Journal of Ichthyology*, vol. 3, no. 1, pp. 42–47, 1991.

[40] J.-S. Kim, I.-S. Kim, and J. W. Shim, "Electrophoretic study on the muscle proteins and systematic relationships of the gudgeons of the subfamily Gobioninae (Cyprinidae) in Korea," *Korean Journal of Limnology*, vol. 17, no. 3, pp. 55–61, 1984.

[41] P. Banarescu and T. T. Nalbant, "Studies on the systematic of Gobioninae (Pisces, Cyprinidae)," *Revue Roumaine de Biologie*, vol. 10, no. 4, pp. 219–229, 1965.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 147064, 11 pages
http://dx.doi.org/10.1155/2013/147064

Research Article

Evolutionary Relations of Hexanchiformes Deep-Sea Sharks Elucidated by Whole Mitochondrial Genome Sequences

Keiko Tanaka,[1] Takashi Shiina,[1] Taketeru Tomita,[2] Shingo Suzuki,[1] Kazuyoshi Hosomichi,[3] Kazumi Sano,[4] Hiroyuki Doi,[5] Azumi Kono,[1] Tomoyoshi Komiyama,[6] Hidetoshi Inoko,[1] Jerzy K. Kulski,[1,7] and Sho Tanaka[8]

[1] *Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan*
[2] *Fisheries Science Center, The Hokkaido University Museum, 3-1-1 Minato-cho, Hakodate, Hokkaido 041-8611, Japan*
[3] *Division of Human Genetics, Department of Integrated Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan*
[4] *Division of Science Interpreter Training, Komaba Organization for Education Excellence College of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan*
[5] *Shimonoseki Marine Science Museum, 6-1 Arcaport, Shimonoseki, Yamaguchi 750-0036, Japan*
[6] *Department of Clinical Pharmacology, Division of Basic Clinical Science and Public Health, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan*
[7] *Centre for Forensic Science, The University of Western Australia, Nedlands, WA 6008, Australia*
[8] *Department of Marine Biology, School of Marine Science and Technology, Tokai University, 3-20-1 Orido, Shimizu, Shizuoka 424-8610, Japan*

Correspondence should be addressed to Takashi Shiina; tshiina@is.icc.u-tokai.ac.jp

Received 1 March 2013; Accepted 26 July 2013

Academic Editor: Dietmar Quandt

Hexanchiformes is regarded as a monophyletic taxon, but the morphological and genetic relationships between the five extant species within the order are still uncertain. In this study, we determined the whole mitochondrial DNA (mtDNA) sequences of seven sharks including representatives of the five Hexanchiformes, one squaliform, and one carcharhiniform and inferred the phylogenetic relationships among those species and 12 other Chondrichthyes (cartilaginous fishes) species for which the complete mitogenome is available. The monophyly of Hexanchiformes and its close relation with all other Squaliformes sharks were strongly supported by likelihood and Bayesian phylogenetic analysis of 13,749 aligned nucleotides of 13 protein coding genes and two rRNA genes that were derived from the whole mDNA sequences of the 19 species. The phylogeny suggested that Hexanchiformes is in the superorder Squalomorphi, *Chlamydoselachus anguineus* (frilled shark) is the sister species to all other Hexanchiformes, and the relations within Hexanchiformes are well resolved as *Chlamydoselachus*, (*Notorynchus*, (*Heptranchias*, (*Hexanchus griseus*, *H. nakamurai*))). Based on our phylogeny, we discussed evolutionary scenarios of the jaw suspension mechanism and gill slit numbers that are significant features in the sharks.

1. Introduction

The subdivision Selachii or modern sharks, along with skates and rays, comprises the subclass Neoselachii within the class Chondrichthyes or cartilaginous fishes. Chondrichthyans, including the Neoselachii, chimaeroids, and several fossil forms are defined as jawed fish with skeletons made of prismatic cartilage rather than bone and pelvic claspers in males. The Selachii can be divided into two superorders, the Galeomorphi (339 species) and the Squalomorphi (155 species), and eight extant orders [1].

Among the sharks, Hexanchiformes is regarded as an ancient order of sharks with just five extant species that are characterized by having only one dorsal fin, either six or seven gill clefts, and no nictitating membrane in the eyes

[2]. The Hexanchiformes is usually divided into two families, the Chlamydoselachidae (*Chlamydoselachus anguineus* (*C. anguineus*)) and the Hexanchidae (*Hexanchus griseus* (*H. griseus*), *Hexanchus nakamurai* (*H. nakamurai*), *Notorynchus cepedianus* (*N. cepedianus*), and *Heptranchias perlo* (*H. perlo*)) with the latter family also known as "cow sharks." The frilled shark, *C. anguineus*, is very different from the cow sharks, and its own order of Chlamydoselachiformes was proposed [3]. However, derived features (e.g., the extra gill arch and more-heart valve rows) shared with other Hexanchiformes support its retention within the Hexanchiformes. A third family, Notorynchidae, was also proposed for the *Notorynchus* species because of morphological and behavioral differences from the other members of the family Hexanchidae [3]. Interestingly, the tooth structure and composition of one of the Hexanchiformes, *C. anguineus*, is similar to that of the stem-group fossil shark *Cladoselache* sp., although such features are not observed in the other Hexanchiformes species [4]. Therefore, the accurate placement of Hexanchiformes is essential to understand the evolution of morphology in sharks. However, the lack of available DNA sequence data for most shark species and orders remains a major limitation to obtaining reliable results in molecular phylogenetic studies.

The mitochondrial DNA (mtDNA) has been one of the most widely used molecular markers for diversity and phylogenetic studies in animals because of its size, maternal mode of inheritance, high rate of mutation, and simple genomic structure [5]. Although mtDNA sequences have proved valuable in determining phylogenetic relationships, the choice of a gene as a molecular marker and clock in phylogeny is also important [6, 7]. Recent phylogenetic studies in different taxa suggest that full-length mitochondrial genomic sequences provide an improved resolution for reconstructing a robust phylogeny and for molecular dating of divergence events within a phylogeny [8]. So far, of all the shark species, the complete mtDNA sequences were determined in the subclass Neoselachii of some species, and the sequences were used for elucidating interrelationships between sharks and bony fishes and between sharks and rays [9–14]. Therefore, there were no analyses of the relationships between species within cartilaginous fish orders using whole mtDNA sequences, although intrarelationships were estimated by using partial mtDNA and nuclear DNA sequences [6, 7, 10, 15–20].

In order to obtain more sequence data for different shark species and to allow accurate placement within the order Hexanchiformes, we chose to determine the complete mtDNA sequences of seven shark species including five species of the order Hexanchiformes, along with *Somniosus pacificus* (*S. pacificus*), which is a member of the order Squaliformes and *Pseudotriakis microdon* (*P. microdon*), which is a member of the order Carcharhiniformes. We then analysed the phylogenetic relationships among the five Hexanchiformes species and between the Hexanchiformes and ten other shark species using the complete mitochondrial genomic sequences of 19 cartilaginous fish species. On the basis of the results of our phylogenetic analysis, we propose new scenarios for the evolution of the jaw suspension mechanism and the number of gill clefts that are significant features in sharks.

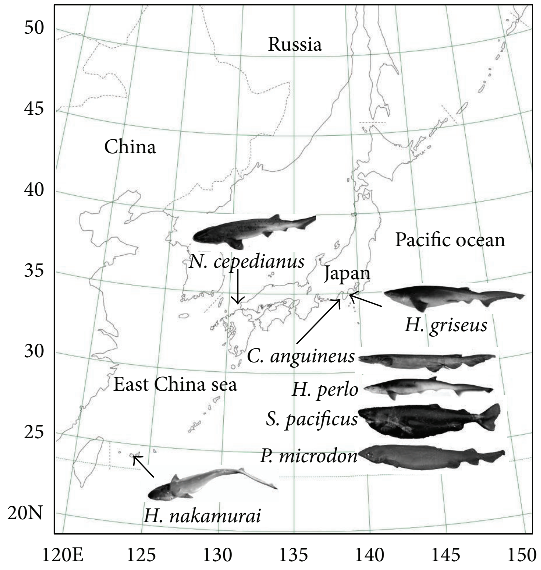


FIGURE 1: Geographic locations of the five Hexanchiformes species and two other species caught off the coast of Japan for nucleotide sequencing in this study. Biological features of the sharks are shown in Table 1.

2. Materials and Methods

2.1. Sample Collection and Isolation of the mtDNA. The shark specimens in this study were all captured off the coast of Japan, *C. anguineus*, *H. perlo, S. pacificus*, and *P. microdon* within Suruga Bay, *H. griseus* within Sagami Bay, *H. nakamurai* near Ishigaki Island, and *N. cepedianus* near Futaoi Island (Figure 1). In this paper, we use the species name *Hexanchus nakamurai* [21] instead of its synonym *Hexanchus vitulus* [22]. The mtDNAs were isolated from the muscle or spleen tissue using the mtDNA Extractor CT Kit (Wako Pure Chemical Industries, Ltd., Osaka, Japan) or standard phenol-chloroform method [23]. The quantity and quality of the isolated DNA samples were measured and estimated, respectively, by the spectral absorbance of the DNA at 260 nm and 280 nm.

2.2. PCR Amplification of the Entire mtDNA Regions. Twenty-two pairs of primers were newly designed by comparing previously published shark mtDNA sequences in the GenBank/EMBL/DDBJ database with assistance from Primer Express v1.0 (Applied Biosystems, CA, USA) for polymerase chain reaction (PCR) amplification of the entire mtDNA regions (eee Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2013/147064). For PCR amplification of the cytochrome b (CYTB) region, the 10 μL amplification reaction contained 10 ng of mtDNA, 1.0 unit of Ex Taq polymerase (TaKaRa Shuzo, Otsu, Japan), 1x PCR buffer, 2.5 mM $MgCl_2$, 2 mM of each dNTP, and 0.5 μM of each primer. The cycling parameters were as follows: an initial denaturation of 96°C for 3 min followed by 30 cycles of 96°C for 30 sec, 50°C for 30 sec, and 72°C for 1 min and followed by a final cycle of 72°C for 4 min. For long-ranged PCR amplifications, the 20 μL amplification reaction contained 10 ng of mtDNA, 0.4 unit of KOD-FX polymerase (TOYOBO, Osaka, Japan), 2x PCR buffer, 2.5 mM $MgCl_2$, 400 μM of each dNTP, and 0.5 μM of each primer. The cycling parameters were as follows: an initial denaturation of 94°C for 2 min followed by 35 cycles of 98°C for 10 sec, 60 or 68°C for 30 sec, and 68°C for 2 to 20 min. PCR reactions were performed by using the thermal cycler GeneAmp PCR system 9700 (Applied Biosystems, CA, USA). The long-ranged PCR size was 6.5 kb on average and ranged from 1,711 bp to 16,800 bp (Supplementary table 1).

2.3. Genomic Sequencing Strategy and Sequence Analysis. The PCR products were subjected to complete and bidirectional shotgun sequencing with an average 7.2x coverage which was sufficient for assembly and analysis of the entire sequence using previously established procedures [26] and direct sequencing using PCR primers as sequencing primers. DNA sequencing was performed by the cycle sequencing method using Ampli*Taq*-DNA polymerase FS and the fluorescently labeled BigDye terminators in a GeneAmp PCR system 9700 (Applied Biosystems, Foster City, CA, USA). A 3130xl Genetic Analyzer was used for automated fluorescent sequencing (Applied Biosystems, Foster City, CA, USA). Individual sequences were minimally edited to remove vector sequences and assembled into contigs using the Sequencher 4.2 software (Gene Codes CO., MI, USA). Remaining gaps or ambiguous nucleotides were determined by the direct sequencing of PCR products obtained with appropriate PCR primers or by nucleotide sequence determination of shotgun clones.

Nucleotide similarities between sequences were calculated by the "Search Homology" tool of GENETYX-MAC ver. 12.0 (Software Development Co. Ltd., Tokyo, Japan), and those with nucleotide sequences in GenBank/EMBL/DDBJ were searched by BLAST program (http://www.ncbi.nlm.nih.gov/BLAST/). The newly determined mtDNA sequences were annotated by comparison with known mtDNA sequence information of other shark species.

2.4. Phylogenetic Analysis. Multiple sequence alignments were created using the ClustalW Sequence Alignment program of the Molecular Evolution Genetics Analysis software 5 (MEGA5; [27]). Nucleotide alignments were separately created for each of the 13 protein coding genes, *ND1, ND2, COX1, COX2, ATP8, ATP6, COX3, ND3, ND4L, ND4, ND5, ND6*, and *CYTB* and two ribosomal RNA (rRNA) genes, *12S* and *16S*. After excluding some gaps, 13,749 nucleotides were aligned (Align_Set_1) for 19 Chondrichthyes species (including *C. monstrosa*), and 13,784 nucleotides were aligned (Align_Set_2) for 18 Selachii and Batoidea species using the ClustalW sequence alignment settings with few manual adjustments (Table 2).

Phylogenetic trees were constructed using the model-based maximum likelihood (ML) and Bayesian inference (BI: MrBayes Ver. 3.2.1 [28]) methods and the distance-based neighbour-joining (NJ) method (MEGA5). For the ML analyses we used "Find best DNA/protein models (ML)" program of MEGA5 to estimate the most likely model of sequence evolution including third codon sites. Based on maximum likelihood values and the Akaike information criterion (AIC) [29], the TN93+G+I model was selected as the most likely model ($\ln L = -118970.445$, AIC = 238464.763 for Align_Set_1, and $\ln L = -110086.677$, AIC = 220253.367 for Align_Set_2) for a nucleotide based ML tree using 10,000 ML-bootstrap replicates. The neighbor-joining tree was constructed using distances corrected according to the Kimura 2-parameter model with 1.0 gamma parameters [30] and assessed using 10,000 bootstrap replicates. For the BI analyses, we used MrAIC Ver. 1.4.4, http://www.abc.se/~nylander/, Nylander unpublished) with PhyML Ver. 3.0 [31] to estimate the most likely model of the sequence evolution. Based on maximum likelihood values and the Bayesian information criterion (BIC), the GTR+G+I model was selected as the most likely model ($\ln L = -118104.860$, BIC = 236638.513 for Align_Set_1 and $\ln L = -109301.198$, BIC = 218688.396 for Align_Set_2) for the Bayesian inference (BI) method. The Bayesian analysis was run using the Metropolis coupled Markov chain Monte Carlo (MCMC) algorithm from randomly generated starting trees for 1,500,000 generations with sampling every 100 generations. The first 100,100 steps of each run were discarded as burn-in. The stabilized burn-in level was assessed using Tracer v1.4 (http://beast.bio.ed.ac.uk/Tracer; Rambaut & Drummond unpublished). Convergence for both runs was examined using the average standard deviation of the split frequencies and through examination of the Markov Chain Monte Carlo chains using Tracer v1.4.

2.5. Estimation of Divergence Times. Divergence times for each node of Hexanchiformes and other sharks from the orders Squalomorphi and Galeomorphi were estimated by the Divergence Time program in MEGA5 software based on the phylogenetic tree of the ML method and on previously estimated divergence times of Selachii and Batoidea (213.4 Mya (203.3~228.8 Mya in the 95% confidence intervals)) [24].

3. Results

3.1. Biological and Genetic Information. The biological and genetic features of the seven shark species sampled for this study are shown in Table 1. Six or seven gill clefts and one dorsal fin were observed in the Hexanchiformes species, whereas five gill clefts and two dorsal fins were observed in *S. pacificus* and *P. microdon* and in all non-Hexanchiformes and Neoselachii, except some skates and myliobatoid rays in which dorsal fins were presumably lost secondarily [32]. Our preliminary sequence analysis of the mtDNA *Cytb* or *12S* genes for six of the seven species, excluding *H. nakamurai*, using a BLAST search (GenBank) primarily showed the highest nucleotide identities (97.0% to 100%) with previously published nucleotide sequences of the target species and secondarily showed lower nucleotide identities (80.2% to 93.1%) with the nucleotide sequences of different species (Table 1). Therefore, this analysis, on the basis of biological and/or genetic features, helped us to cross-check and confirm that all our collected samples were identified correctly as the targeted species.

Table 1: Biological features of seven sharks analyzed in this study.

Scientific name		*C. anguineus*	*N. cepedianus*	*H. perlo*	*H. griseus*	*H. nakamurai*	*S. pacificus*	*P. microdon*
Captured location		Suruga Bay	Futaoi Island	Suruga Bay	Sagami Bay	Ishigaki Island	Suruga Bay	Suruga Bay
Captured date		25 Sep. 2007	5 Feb. 2008	26 Sep. 1996	9 Jul. 2008	5 Sep. 2009	31 Mar. 2009	13 Oct. 2008
Sex		F	M	F	F	M	M	F
Total length (mm)		1,480	1,500	715	4,270	1,300	2,900	2,084
Body length (mm)		1,226	Un	483	Un	Un	2,450	1,685
Body weight (g)		10,400	17,000	1,272	Un	7,000	Un	43,040
Gill cleft number		6	7	7	6	6	5	5
Dorsal fin number		1	1	1	1	1	2	2
Nucleotide identity with previously published mtDNA nucleotide sequences	Primary	99.7% with *C. anguineus* CYTB (D50022)	98.7% with *N. cepedianus* CYTB (M91186)	97.3% with *H. perlo* 12S rRNA (AY14788)	100% with *H. griseus* CYTB (DQ132493)	*H. nakamurai* mtDNA sequences have not been published so far.	100% identity with *S. pacificus* CYTB (EF090957)	97.0% with *P. microdon* CYTB (DQ422078)
	Secondary	80.2% with *M. pelagios* CYTB (U91440)	87.8% with *H. griseus* CYTB (DQ132493)	93.1% with *H. griseus* 12S rRNA (AY147887)	82.0% with *C. anguineus* CYTB (D50022)	90.9% with *H. griseus* CYTB (DQ132493)	84.2% with *S. acanthia* CYTB (Y18134)	89.9% with *G. attenuatus* CYTB (DQ422079)

Un: the feature is not known.

Table 2: List of the 19 mitochondrial genomes analyzed in this study.

Classification	Scientific name	Common name	Accession no. and reference
Chondrichthyes			
Neoselachii			
Selachii			
Squalomorphi			
Hexanchiformes			
Chlamydoselachidae	***Chlamydoselachus anguineus***	**Frilled shark**	**AB560487, this study**
Hexanchidae	***Heptranchias perlo***	**Sharpnose sevengill shark**	**AB560489, this study**
	Hexanchus griseus	**Bluntnose sixgill shark**	**AB560490, this study**
	Hexanchus nakamurai	**Bigeye sixgill shark**	**AB560491, this study**
Notorynchidae	***Notorynchus cepedianus***	**Broadnose sevengill shark**	**AB560488, this study**
Squaliformes			
Squalidae	*Squalus acanthias*	Spiny dogfish	Y18134, Rasmussen and Arnason [11]
Somniosidae	***Somniosus pacificus***	**Pacific sleeper shark**	**AB560492, this study**
Galeomorphi			
Orectolobiformes			
Hemiscylliidae	*Chiloscyllium plagiosum*	Whitespotted bamboo shark	FJ853422, [54] Zhang et al.
	Chiloscyllium griseum	Grey bamboo shark	JQ434458, unpublished
	Chiloscyllium punctatum	Brownbanded bamboo shark	JQ082337, Chen et al. [55]
Carcharhiniformes			
Carcharhinidae	*Scoliodon macrorhynchos*		JQ693102, Chen et al. [56]
Triakidae	*Mustelus manazo*	Starspotted smooth-hound	AB015962, Cao et al. [9]
Scyliorhinidae	*Scyliorhinus canicula*	Small spotted catshark	Y16067, Delarbre et al. [10]
Pseudotriakidae	***Pseudotriakis microdon***	**False catshark**	**AB560493, this study**
Lamniformes			
Mitsukurinidae	*Mitsukurina owstoni*	Goblin shark	EU528659, unpublished
Batoidea			
Rajiformes			
Rajidae	*Okamejei kenojei*	Ocellate spot skate	AY525783, Kim et al. [14]
	Amblyraja radiata	Starry ray	AF106038, Rasmussen and Arnason [12]
Myliobatiformes			
Plesiobatidae	*Plesiobatis daviesi*	Deepwater stingray	AY597334, unpublished
Holocephali			
Chimaeriformes			
Chimaeridae	*Chimaera monstrosa*	Rabbitfish	AJ310140, Arnason et al. [13]

Note-Classifications follow Nelson (2006) and Inoue et al. [8]. Bold letter indicates the mitochondrial genome sequences of shark species determined in this study.

3.2. Genome Structure of mtDNA Sequences in Chondrichthyes Species. Whole mtDNA sequences of the seven sharks were determined by PCR-based shotgun sequencing. Their nucleotide length was 17,314 bp in *C. anguineus*, 16,990 bp in *N. cepedianus*, 18,909 bp in *H. perlo*, 17,223 bp in *H. griseus*, 18,605 bp in *H. nakamurai*, 16,730 bp in *S. pacificus*, and 16,700 bp in *P. microdon* (GenBank/EMBL/DDBJ accession numbers: AB560487 to AB560493), with the mtDNA length variability due to the presence of varying numbers and compositions of tandem repeats in the control region (Supplementary table 2). From a genomic comparison of 19 Chondrichthyes mitogenomes, all were basically composed of two rRNAs (*12S* and *16S*), 22 transfer RNAs (tRNAs), 13 protein coding genes, and the D-loop control region. Species-specific duplications, insertions, and deletions were observed in some species such as the duplication of tRNA-Trp and the deletion of tRNA-Asn and tRNA-Leu2 in *M. manazo* and insertion of tRNA-Thr for the D-loop region in *C. monstrosa* (Supplementary table 2). The gene directions of *ND6* and eight tRNAs, tRNA-Gln, tRNA-Ala, tRNA-Asn, tRNA-Cys, tRNA-Tyr, tRNA-Ser, tRNA-Glu, and tRNA-Pro, were encoded in the mtDNA L chain, and the other genes were encoded in the H chain in all species. The GC contents of the species ranged from 35.0% in *C. anguineus* to 42.4% in *O. kenojei* (Supplementary table 2).

3.3. Phylogeny of the Hexanchiformes Using Complete mtDNA Sequences. Figure 2 shows an ML phylogenetic tree constructed by using the TN93+G+I model of the ML method. The tree was constructed using the 13,749 bp nucleotide

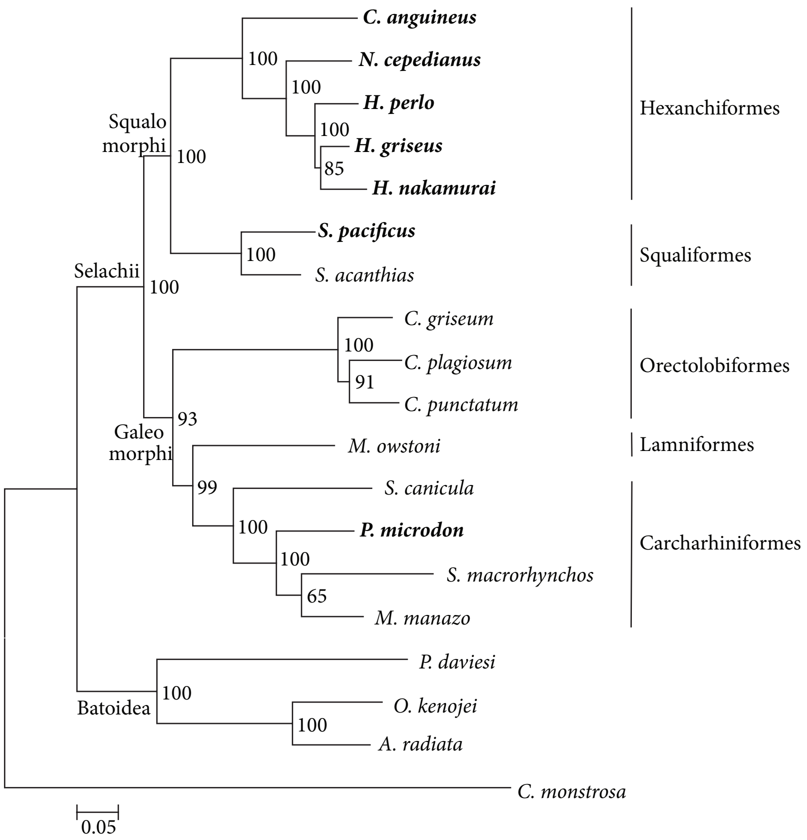


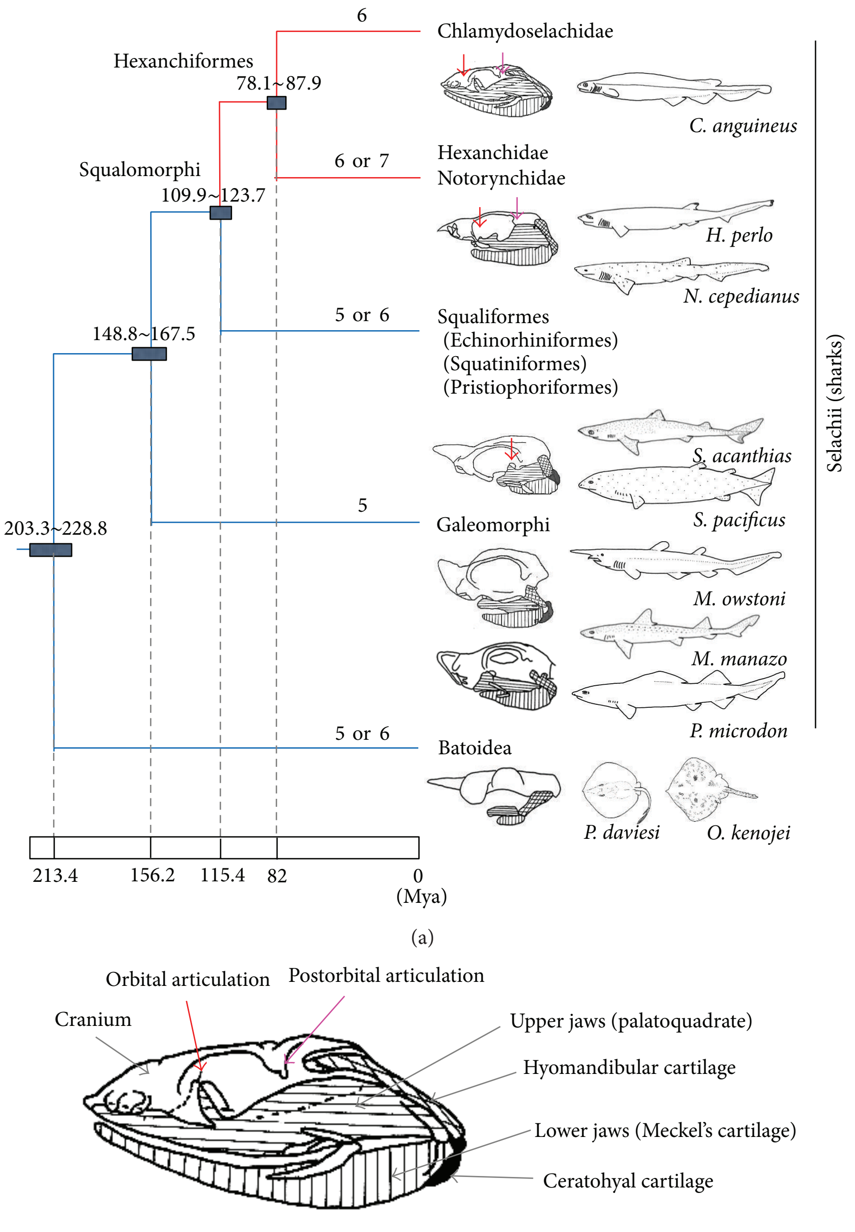
Figure 2: Maximum likelihood phylogeny depicting relationships among 19 Chondrichthyes species inferred from the whole-mtDNA sequences. The 13,749 bp nucleotide alignment (Align_Set_1) was used for the analysis. Numbers on the branches are bootstrap support values. Bold letters indicate the species that were newly sequenced for this study. The Neoselachii subdivisions (Selachii, Batoidea), superorders (Squalomorphi, Galeomorphi) and orders (Hexanchiformes, Squaliformes, Orectolobiformes, Carcharhiniformes, and Lamniformes) are indicated in block letters on the basal branches of the tree. The outgroup species *C. monstrosa* is in the subclass Holocephali.

alignment (Align_Set_1) of the 13 protein coding genes and two rRNA genes from 19 species with *C. monstrosa* selected as the outgroup. In the aligned sequence 48.7% nucleotides (6,693 sites) were constant sites. The similar topology, supported clades, and bootstrap support values or posterior probability were shown by the GTR+G+I model of the BI method and the Kimura 2-parameter model of the NJ method using the same nucleotide alignment as for the ML method (Supplementary figure 1). Moreover, even if we set three Batoidea species as outgroups, the same topology, supported clades, and bootstrap support values or posterior probability were shown by the TN93+G+I model within the ML method, the GTR+G+I model within the BI method and the Kimura 2-parameter model within the NJ method using 13,784 nucleotide alignment (Align_Set_2) of the 13 protein coding genes and two rRNA genes from 18 species (Supplementary figure 2).

Our phylogenetic analyses of nucleotide sequences using the ML, BI, and NJ methods suggest that the Selachii is divided into the two superorders, Squalomorphi and Galeomorphi. The Squalomorphi superorder includes the orders Hexanchiformes and Squaliformes, and the Galeomorphi superorder includes the orders Orectolobiformes, Lamniformes, and Carcharhiniformes (Figure 2). In Figure 2 tree, the Hexanchiformes clade is monophyletic and separates from the Squaliformes clade in the Squalomorphi order that is separated from the Galeomorphi lineage. The Hexanchiformes and Squaliformes lineages showed clades with extremely high bootstrap support values (100) and posterior probabilities (100) (Figure 2, Supplementary figures 1 and 2). The mitogenomic data also show good resolution within the Galeomorphi with Orectolobiformes separated from “Lamniformes plus Carcharhiniformes” with high bootstrap support values (61~93) and posterior probabilities (100) (Figure 2, Supplementary figures 1 and 2).

Within the Hexanchiformes, *C. anguineus* is sister of all the others with the longest branch, so, assuming clock like evolution, it could be the oldest extant shark lineage. The *N. cepedianus* lineage emerged next. By comparison, the terminal branch lengths of *H. griseus*, *H. nakamurai*, and *H. perlo* are relatively short.

Assuming the divergence time of Selachii and Batoidea was 213.4 Mya (203.3 Mya~228.8 Mya) [24] and based on divergence rates in Figure 2, then the divergence times for each node of Squalomorphi and Galeomorphi and Squaliformes and Hexanchiformes are estimated as 156.2 Mya (148.8 Mya~167.5 Mya) and 115.4 Mya (109.9 Mya~123.7 Mya), respectively (Figure 3(a)). Of the Hexanchiformes, the



(a)

(b)

Figure 3: Morphological character evolution mapped onto the phylogenetic tree derived in this study (a) and structure of jaw suspension of *C. anguineus* (b). (a) Divergence time for each node of Selachii was estimated by the divergence time of Selachii and Batoidea (213.4 Mya (203.3~228.8 Mya in the 95% confidence intervals)) [24]. The red line extending from the circled node indicates the evolutionary time (115.4 Mya) of the divergence of Hexanchiformes from the other Squalomorphi lineages. Numbers above the branches indicate the gill cleft numbers, and numbers above the nodes indicate range of the divergence time. The schematic diagrams of jaw suspension were based on previous reports [3, 13, 25]. (b) Vertical, horizontal, and grid-lined areas and black and white areas indicate lower jaws (Meckel's cartilage), upper jaws (palatoquadrate), hyomandibular and ceratohyal cartilages, and cranium, respectively. Red and orange arrows in the schematic diagram of the jaw suspension apparatus indicate orbital and postorbital articulations, respectively.

divergence time of *C. anguineus* is estimated as 82.0 Mya (78.1 Mya~87.9 Mya). This estimation is largely consistent with the known fossil record of Chlamydoselachiformes (85 Mya) [33, 34].

4. Discussion

4.1. Phylogeny of Cartilaginous Fishes. In the case of Batoidea, Carcharhiniformes, Hexanchiformes, and Squaliformes, our phylogeny supports, in a number of respects, the previous findings derived from partial mitochondrial genomes, genes *COX1, NADH2, NADH4, CYTB, 12S, 16S*, and/or tRNAs and nuclear genome genes *5.8S, 18S, 28S*, and *RAG1* [6, 7, 15–19, 35]. Recently, Naylor et al. [7] published the most comprehensive phylogeny of sharks using a total of 595 shark species representing eight orders and 159 genera and 56 families, but mainly using a single mitochondrial gene

(*NADH2*) as the molecular marker. Although we did not have any examples of the Echinorhiniformes, Pristiophoriformes, and Squatiniformes in our study, the remaining orders within Squalomorphi were generally similar to the relationships reported by Vélez-Zuazo [6] and by Naylor et al. [7] of Hexanchiformes and Squaliformes within Squalomorphi and Lamniformes, Orectolobiformes, and Carcharhiniformes within Galeomorphi. Recent myological studies on Hexanchiformes also support the inclusion of the Chlamydoselachidae and Hexanchidae in the Squalomorphi [36].

Our analysis, though limited to just one lamniform representative, found strong support for Lamniformes as the sister order of Carcharhiniformes [2, 19] instead of the Orectolobiformes [6, 35]. In addition, our phylogeny is inconsistent with some previously reported morphology-based phylogenies such as the hypnosqualean hypothesis that places the batoids within sharks [2, 37], the tooth structure-based tree that places Hexanchiformes outside of batoids [38, 39], the jaw protrusion, and feeding-based trees [40] and a Bayesian analysis based on the CYTB gene that supported the position of Hexanchiformes as the sister of all other shark orders [20]. The full-length mitochondrial genomic sequences provide strong statistical support and an improved resolution for reconstructing a robust phylogeny in the cartilaginous fish [8]. Therefore, our phylogeny appears to be a reasonable reconstruction of the evolutionary process dividing the Selachii into the two superorders, Squalomorphi and Galeomorphi. However, the divergence time for a node of Squaliformes and Hexanchiformes was estimated to be 115.4 Mya (109.9 Mya~123.7 Mya). This is not largely consistent with the known fossil record of Hexanchiformes (190 Mya) [33]. In this regard, in case of setting the divergence time of Hexanchiformes for 190 Mya, the divergence times for each node of Selachii/Batoidea and Chlamydoselachidae were estimated as 351.3 Mya and 134.9 Mya, respectively. The divergence time of Chlamydoselachidae is consistent with the known fossil record of Chlamydoselachidae (85 Mya) [33] but that of Selachii/Batoidea is not consistent with the previously estimated divergence time of Selachii and Batoidea (213.4 Mya (203.3~228.8 Mya in the 95% confidence intervals)) [24]. In this study we used only 15 Selachii and three Batoidea species for the phylogenetic analysis, but some Selachii species that show ambiguous classification such as Echinorhiniformes, Pristiophoriformes, and Squatiniformes were not included in this study. Therefore, detailed phylogenetic analysis based on full-length mitochondrial genomic sequences using additional Selachii species that are thought to be diverged on evolutionary important positions, will be necessary for estimation of the precise divergence times in future.

4.2. Phylogenetic Relationships in Hexanchiformes. Vélez-Zuazo and Agnarsson [6] reported that the Hexanchidae within Hexanchiformes was paraphyletic because it also contained the only species of Chlamydoselachidae. Their Bayesian analysis of Hexanchiformes, in using only a small portion (15%) of a single nucleotide sequence composed of five genes, *COX1, NADH2, CYTB, 16S*, and *Rag1*, may be compromised by too little data. They showed low support with a posterior probability of 45% at the node splitting the *C. anguineus* and *N. maculates* taxa. In addition, although they found that *N. maculates* was a sister group to the *N. cepedianus*, *N. maculates* is in fact a synonym of *N. cepedianus* and therefore the same species [6]. In our analysis, *N. cepedianus* is clearly the sister of the *Heptranchias-Hexanchus* clade with which it forms a clade separate from *Chlamydoselachus* with all nodes strongly supported (Figure 2). The sequencing of the complete mitochondrial genome of *N. maculates,* if it is a separate species from *N. cepedianus,* should help to resolve this issue.

The frilled shark, *C. anguineus*, in our phylogenetic analysis, was found to be the sister group to all the other Hexanchiformes, similar to the results on their anatomical studies [2, 3, 36]. Some systematists have proposed a separate order for the frilled shark [3]. Recently, a new Chlamydoselachidae species (*Chlamydoselachus africana; C. africana*) was discovered around southern Africa [41]. Although it is not known if the mtDNA of *C. africana* will group phylogenetically with *C. anguineus,* our phylogeny supports the retention of *C. anguineus* within the Hexanchiformes rather than within its own order. The frilled shark, *C. anguineus,* is placed in its own family, Chlamydoselachidae, because of its many unusual features such as elongated and eel-like body; its low dorsal fin; blunt snout, long jaws that are narrower at the tip than at the corners; terminal mouth, similar upper and lower teeth with three prong-like cusps; and gill clefts with frilly margins and the first gill slit continuous across the throat [42].

The position of *N. cepedianus* in our phylogenetic tree favours its own family name of Notorynchidae rather than being placed into the Hexanchidae family with the one *Heptranchias* species and the two *Hexanchus* species. Incidentally, *H. nakamurai* [21] in our study corresponds to *H. viulus* [22] in the Vélez-Zuazo and Agnarsson study [6], as the latter species name is junior to the former name [42, 43]. However, Naylor et al. [44] suggest that both *H. nakamurai* and *H. viulus* may actually represent valid species.

4.3. New Insights on the Morphological Evolution of Sharks. Our phylogenetic analysis based on mitochondrial genomic sequences is useful for comparing morphological features to phylogenetic relationships among the sharks. For example, the Hexanchiformes species have six or seven gill clefts and one dorsal fin, whereas most of other Selachii species have five or six gill clefts and mostly two dorsal fins (Table 1). Most fossils of Agnathans and some fossils of acanthodian (stem Chondrichthyes or stem osteichthyes) support the presence of multiple gill clefts [45], and the sea lamprey, which is in the class Petromyzontida, has seven gill clefts. However, the rabbitfish, which is in the subclass Holocephali, is from the sister clade of sharks and Batoidea, and it has one gill cleft. In this regard, the phylogenetic tree suggests that the multiple gill clefts have been maintained by species-specific increases and decreases in both the Petromyzontida and Neoselachii lineages (Figure 3(a)), but holocephalians could easily have reduced gill slits to one.

The phylogenetic placement of Hexanchiformes in our study suggests a clearer scenario for the evolution of the upper jaw suspension in sharks. Upper jaws (palatoquadrate) of sharks are "suspended" from the cranium by the hyomandibular cartilage and several articulations between the cranium and upper jaw (Figure 3(b)) and shark jaws evolved with a general trend towards jaw shortening, increased kinesis of upper jaw suspension, and protrusion [46]. Many researchers have focused on the evolution of the jaw suspension mechanism in order to clarify the phylogenetic relationship within sharks [47–49]. Because the jaw suspension mechanism is strongly coupled with the jaw protrusion capability, the evolution of jaw protrusion behaviour has been reconstructed based on the evolution of jaw suspension mechanisms [46]. However, the position of Hexanchiformes as the sister to Squaliformes (Figure 3) suggests that the Hexanchiformes represents an evolved state of the jaw suspension mechanism independent of the other shark orders. It can be expected that investigation of fossils from non-Neoselachii Chondrichthyes such as the hybodonts and cladodonts will help to elucidate the polarization of the evolution of the jaw suspension mechanism.

The Hexanchiformes species have two characteristic articulations between their cranium and palatoquadrate. One is the orbital articulation, which is between the orbital process of the palatoquadrate and the orbital walls of the cranium, and the other is the postorbital articulation, which is between the otic process of the palatoquadrate and the postorbital process of the cranium [47, 48]. Due to the postorbital articulation, jaw protrusion capabilities of the Hexanchiformes species are strongly restricted. On the other hand, Squaliformes species lack the postorbital articulations, and Galeomorphi species lack both the orbital articulations and the postorbital articulation [47, 48]. Because the postorbital and/or orbital articulations are absent from these two clades, in contrast to Hexanchiformes species, they have a greater capability of jaw protrusion, flexibility, and maneuverability [46]. Extensive upper jaw protrusion in modern sharks was found to allow faster closure of jaws to gouge or cut smaller pieces of prey to fit into the mouth [50]. Therefore, based on the differences in jaw suspension mechanisms among Hexanchiformes, Squaliformes, and Galeomorphi, the following evolutionary scenario can be proposed. First, the orbital process and postorbital articulation were missing from the common ancestor of all shark orders. Second, the orbital articulation was gained at the base of Squaliformes + Hexanchiformes clade, and this modification restricted the evolved sharks to protrude their jaws. Third, the postorbital articulation was gained at the base of Hexanchiformes clade, and this modification might have further restricted the Hexanchiformes from the capability of upper jaw protrusion.

Previous research suggested that the presence of the orbital process in the palatoquadrate is one of the strong morphological characteristics indicating the monophyly of Hexanchiformes and Squaliformes. Thus, Maisey named the Hexanchiformes + Squaliformes clade as "orbitostylic sharks" [25, 47–49]. Our result suggests that the presence of the orbital process is an evolved character of the Hexanchiformes + Squaliformes clades and not a plesiomorphic character of the whole shark clade (Figure 3(a)).

In summary, we sequenced and analysed the complete mtDNA sequences of five Hexanchiformes species. Our phylogeny and the known morphological features of sharks resolved interrelationships of major Hexanchiformes species. Further insights into phylogeny of the mtDNA sequences were provided by comparative analyses using other shark and nonshark species. A similar approach using the whole mtDNA genome of sharks in the other orders should help to resolve the intraspecific and interspecific relationships within Chondrichthyes (cartilaginous fishes). The Hexanchiformes are observed mainly in deep-sea areas (<1000 m) all over the world. However, they rarely occur in the continental shelf shallower than 200 m [51]. This may be a relic of a past behavioral habit, when their ancestors had inhabited the coastal shelf [52, 53]. Although some Hexanchiformes specific morphological features such as tooth, gill cleft numbers, jaw suspension, and no nictitating membrane in the eyes. have been reported so far, the other features such as distribution of the living areas, physiology, reproduction and genetic diversity are unknown. Therefore, it will be necessary to compare between those features and phylogenetic relationships derived from nucleotide sequences to comprehensively understand evolution of the sharks including Hexanchiformes.

Conflict of Interests

The authors have no conflict of interests.

Acknowledgments

The authors gratefully acknowledge the scientific communication of Gavin J. P. Naylor at the the Florida State University. The authors thank Minoru Toda, Toshiaki Ishibashi, Shinsaku Ochiai, and Toshikazu Yano who provided and collected samples. This work was supported by the Scientific Research on Priority Areas "Comparative Genomics" (20017023) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT).

References

[1] L. J. V. Compagno, D. A. Didier, and G. H. Burgess, "Classification of chondrichthyan fish," in *Sharks, Rays and Chimaera: The Status of the Chondrichthyan Fishes Status Survey IUCN/SSC Shark Specialist Group*, S. L. Fowler, R. D. Cavanagh, M. Camhi et al., Eds., Information Press, Oxford, UK, 2005.

[2] M. R. de Carvalho, "Higher-level elasmobranch phylogeny, basal squaleans, and paraphyly," in *Interrelationships of Fishes*, M. L. J. Stassny, L. R. Parenti, and G. D. Johnson, Eds., Academic Press, San Diego, Calif, USA, 1996.

[3] S. Shirai, "Phylogenetic interrelationships of Neoselachians (Chondrichthyes: Euselachii)," in *Interrelationships of Fishes*, M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson, Eds., Academic Press, San Diego, Calif, USA, 1996.

[4] M. Goto and I. Hashimoto, "Studies on the teeth of *Chlamydoselachus anguineus*, a living archaic fish. I. On the morphology, structure and composition of the teeth," *Japanese Journal of Oral Biology*, vol. 18, no. 3, pp. 362–377, 1976.

[5] J. C. Avise, *Molecular Markers, Natural History, and Evolution*, Sinauer Associates, Sunderland, Mass, USA, 2nd edition, 2004.

[6] X. Vélez-Zuazo and I. Agnarsson, "Shark tales: a molecular species-level phylogeny of sharks (Selachimorpha, Chondrichthyes)," *Molecular Phylogenetics and Evolution*, vol. 58, no. 2, pp. 207–217, 2011.

[7] G. J. P. Naylor, J. N. Kirsten, K. A. M. Rosana, N. Straube, and C. Lakner, "Elasmobranch phylogeny: a mitochondrial estimate based on 595 species," in *Biology of Sharks and Their Relatives*, C. Jeffrey, J. A. M. Carrier, and M. R. Heithans, Eds., pp. 31–56, CRC Press, Boca Raton, Fla, USA, 2012.

[8] J. G. Inoue, M. Miya, K. Lam et al., "Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective," *Molecular Biology and Evolution*, vol. 27, no. 11, pp. 2576–2586, 2010.

[9] Y. Cao, P. J. Waddell, N. Okada, and M. Hasegawa, "The complete mitochondrial DNA sequence of the shark Mustelus manazo: evaluating rooting contradictions to living bony vertebrates," *Molecular Biology and Evolution*, vol. 15, no. 12, pp. 1637–1646, 1998.

[10] C. Delarbre, N. Spruyt, C. Delmarre et al., "The complete nucleotide sequence of the mitochondrial DNA of the dogfish, Scyliorhinus canicula," *Genetics*, vol. 150, no. 1, pp. 331–344, 1998.

[11] A. Rasmussen and U. Arnason, "Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fishes within the tree of bony fishes," *Journal of Molecular Evolution*, vol. 48, no. 1, pp. 118–123, 1999.

[12] A. Rasmussen and U. Arnason, "Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 5, pp. 2177–2182, 1999.

[13] U. Arnason, A. Gullberg, and A. Janke, "Molecular phylogenetics of gnathostomous (jawed) fishes: old bones, new cartilage," *Zoologica Scripta*, vol. 30, no. 4, pp. 249–255, 2001.

[14] I. Kim, S. Jung, Y. Lee, C. Lee, J. Park, and J. Lee, "The complete mitochondrial genome of the rayfish Raja porosa (Chondrichthyes, Rajidae)," *DNA Sequence*, vol. 16, no. 3, pp. 187–194, 2005.

[15] K. Dunn and J. Morrissey :, "Molecular phylogeny of elasmobranchs," *Copeia*, vol. 3, pp. 526–531, 1995.

[16] C. J. Douady, M. Dosay, M. S. Shivji, and M. J. Stanhope, "Molecular phylogenetic evidence refuting the hypothesis of Batoidea (rays and skates) as derived sharks," *Molecular Phylogenetics and Evolution*, vol. 26, no. 2, pp. 215–221, 2003.

[17] C. J. Winchell, A. P. Martin, and J. Mallatt, "Phylogeny of elasmobranchs based on LSU and SSU ribosomal RNA genes," *Molecular Phylogenetics and Evolution*, vol. 31, no. 1, pp. 214–224, 2004.

[18] M. P. Heinicke, G. J. S. Naylor, and S. B. Hedges, "Cartilaginous fishes (Chondrichthyes)," in *The Timetree of Life*, S. B. Hedges and S. Kumar, Eds., Oxford University Press, New York, NY, USA, 2009.

[19] G. J. P. Naylor, J. A. Ryburn, O. Fedrigo, and A. Lopez, "Phylogenetic relationships among the major lineages of modern elasmobranchs," in *Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids, and Chimaeras*, W. C. Hamlett, Ed., Science Publishers, Enfield, UK, 2005.

[20] B. A. Human, E. P. Owen, L. J. V. Compagno, and E. H. Harley, "Testing morphologically based phylogenetic theories within the cartilaginous fishes with molecular data, with special reference to the catshark family (Chondrichthyes; Scyliorhinidae) and the interrelationships within them," *Molecular Phylogenetics and Evolution*, vol. 39, no. 2, pp. 384–391, 2006.

[21] H. Teng, *Classification and Distribution of the Chondrichthys of Taiwan*, Ogawa Press, Tokyo, Japan, 1962.

[22] Springer, W. Stewart, and A. Richard, "Hexanchus vitulus, a new sixgill shark from the Bahamas," *Bulletin of Marine Science*, vol. 19, no. 1, pp. 159–174, 1969.

[23] N. Blin and D. W. Stafford, "A general method for isolation of high molecular weight DNA from eukaryotes," *Nucleic Acids Research*, vol. 3, no. 9, pp. 2303–2308, 1976.

[24] N. C. Aschliman, M. Nishida, M. Miya, J. G. Inoue, K. M. Rosana, and G. J. P. Naylor, "Body plan convergence in the evolution of skates and rays (Chondrichthyes: Batoidea)," *Molecular Phylogenetics and Evolution*, vol. 63, no. 1, pp. 28–42, 2012.

[25] M. R. de Carvalho and J. G. Maisey, "Phylogenetic relationships of the Late Jurassic shark Protospinax WOODWARD 1919 (Chondrichthyes: Elasmobranchii)," in *Mezoic Fishes: Systematics and Ecology*, G. Arratia and V. G. München, Eds., Verlag Dr. Friedrich Pfeil, München, Germany, 1996.

[26] T. Shiina, G. Tamiya, A. Oka et al., "Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 23, pp. 13282–13287, 1999.

[27] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[28] F. Ronquist, M. Teslenko, P. Van Der Mark et al., "Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space," *Systematic Biology*, vol. 61, no. 3, pp. 539–542, 2012.

[29] H. Akaike, "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki, Eds., pp. 267–281, Akademiai Kiado, Budapest, Hungary, 1973.

[30] Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods," *Journal of Molecular Evolution*, vol. 39, no. 3, pp. 306–314, 1994.

[31] S. Guindon, J. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0," *Systematic Biology*, vol. 59, no. 3, pp. 307–321, 2010.

[32] L. Compagno, "Phyletic relationships of living sharks and rays," *Integrative and Comparative Biology*, vol. 17, no. 2, pp. 303–322, 1977.

[33] J. G. Maisey, "What is an "elasmobranch"? The impact of palaeontology in understanding elasmobranch phylogeny and evolution," *Journal of Fish Biology*, vol. 80, no. 5, pp. 918–951, 2012.

[34] H. Cappetta :, "Chondrichthyes II, mesozoic and cenozoic elasmobranchii," in *Handbook of Paleoichthyology*, p. 193, Gustav Fischer, New York, NY, USA, 1987.

[35] J. Mallatt and C. J. Winchell, "Ribosomal RNA genes and deuterostome phylogeny revisited: more cyclostomes, elasmobranchs, reptiles, and a brittle star," *Molecular Phylogenetics and Evolution*, vol. 43, no. 3, pp. 1005–1022, 2007.

[36] M. C. Soares and M. R. de Carvalho, "Comparative myology of the mandibular and hyoid arches of sharks of the order

hexanchiformes and their bearing on its monophyly and phylogenetic relationships (Chondrichthyes: Elasmobranchii)," *Journal of Morphology*, vol. 274, no. 2, pp. 203–214, 2013.

[37] S. Shirai, *Squalean Phylogeny: A New Framework of Squaloid Sharks and Related Taxa*, Hokkaido University Press, Sapporo, Japan, 1992.

[38] D. Thies and W. E. Reif, "Phylogeny and evolutionary ecology of Mesozoic Neoselachii," *Neues Jahrbuch für Geologie und Paläontologie*, vol. 169, pp. 333–361, 1985.

[39] B. Seret, "Classification et phylogenese des chondrichthyens," *Oceanis*, vol. 12, pp. 161–180, 1986.

[40] E. H. Wu, "Kinematic analysis of jaw protrusion in orectolobiform sharks: a new mechanism for jaw protrusion in Elasmobranchs," *Journal of Molphology*, vol. 222, no. 2, pp. 175–190, 1994.

[41] D. A. Ebert and L. J. V. Compagno, "*Chlamydoselachus africana*, a new species of frilled shark from southern Africa (Chondrichthyes, Hexanchiformes, Chlamydoselachidae)," *Zootaxa*, no. 2173, pp. 1–18, 2009.

[42] K. E. Carpenter, *The Living Marine Resources of the Western Central Atlantic. Volume 1: Introduction, Molluscs, Crustaceans, HagFishes, Sharks, Batoid Fishes and Chimaeras*, FAO Library, Rome, Italy, 2002.

[43] T. Taniuchi and H. Tachikawa, "*Hexanchus nakamurai*, a senior synonym of H. vitulus (Elasmobranchii), with notes on its occurrence in Japan," *Japanese Journal of Ichthyology*, vol. 38, no. 1, pp. 57–60, 1991.

[44] G. J. P. Naylor, J. N. Caira, K. Jensen, K. A. M. Rosana, W. T. White, and P. R. Last, "A DNA sequence-based approach to the identification of shark and ray species and its implications for global elasmobranch diversity and parasitology," *Bulletin of the American Museum of Natural History*, vol. 367, pp. 1–262, 2012.

[45] P. Janvier, *Early Vertebrates*, Clarendon Press, Oxford, UK, 1996.

[46] C. D. Wilga, P. J. Motta, and C. P. Sanford, "Evolution and ecology of feeding in elasmobranchs," *Integrative and Comparative Biology*, vol. 47, no. 1, pp. 55–69, 2007.

[47] J. G. Maisey, "An evaluation of jaw suspension in sharks," *American Museum Novitates*, vol. 2706, pp. 1–17, 1980.

[48] J. G. Maisey, "Chondrichthyan phylogeny: a look at the evidence," *Journal of Vertebrate Paleontology*, vol. 4, no. 3, pp. 359–371, 1984.

[49] S. Shirai, "Identity of extra branchial arches of Hexanchiformes (Pisces, Elasmobranchii)," *Bulletin of the Faculty of Fisheries Hokkaido University*, vol. 43, pp. 24–32, 1992.

[50] T. C. Tricas and J. E. McCosker, "Predatory behavior of the white shark (Carcharodon carcharias), with notes on its biology," *Proceedings of the California Academy of Sciences*, vol. 43, no. 14, pp. 221–238, 1984.

[51] A. Barnett, J. M. Braccini, C. A. Awruch, and D. A. Ebert, "An overview on the role of Hexanchiformes in marine ecosystems: biology, ecology and conservation status of a primitive order of modern sharks," *Journal of Fish Biology*, vol. 80, no. 5, pp. 966–990, 2012.

[52] M. Goto, "The Japanese club for fossil shark tooth research: tooth remains chlamydoselachian sharks from Japan and their phylogeny and paleoecology," *Earth Science*, vol. 58, pp. 361–374, 2004.

[53] T. Tomita and T. Oji, "Habitat reconstruction of oligocene elasmobranchs from Yamaga Formation, Ashiya group, Western Japan," *Paleontological Research*, vol. 14, no. 1, pp. 69–80, 2010.

[54] J. Zhang, Y. Liu, X. Zhang et al., "The identification of microRNAs in the whitespotted bamboo shark (Chiloscyllium plagiosum) liver by Illumina sequencing," *Gene*, vol. 527, no. 1, pp. 259–265, 2013.

[55] X. Chen, Z. Zhou, S. Pichai, X. Huang, and H. Zhang, "Complete mitochondrial genome of the brownbanded bamboo shark Chiloscyllium punctatum," *Mitochondrial DNA*, 2013.

[56] X. Chen, X. Peng, P. Zhang, S. Yang, and M. Liu, "Complete mitochondrial genome of the spadenose shark (Scoliodon macrorhynchos)," *Mitochondrial DNA*, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 390354, 9 pages
http://dx.doi.org/10.1155/2013/390354

Research Article

Dynamic of Mutational Events in Variable Number Tandem Repeats of *Escherichia coli* O157:H7

A. V. Bustamante,[1,2] A. M. Sanso,[1,2] D. O. Segura,[1] A. E. Parma,[1] and P. M. A. Lucchesi[1,2]

[1] *Laboratorio de Inmunoquímica y Biotecnología, Centro de Investigación Veterinaria de Tandil (CIVETAN), Facultad de Ciencias Veterinarias, UNCPBA, Campus Universitario, Paraje Arroyo Seco S/N, 7000 Tandil, Argentina*
[2] *CONICET, Argentina*

Correspondence should be addressed to A. V. Bustamante; avbustaman@vet.unicen.edu.ar

Received 18 February 2013; Accepted 22 July 2013

Academic Editor: Dietmar Quandt

VNTRs regions have been successfully used for bacterial subtyping; however, the hypervariability in VNTR loci is problematic when trying to predict the relationships among isolates. Since few studies have examined the mutation rate of these markers, our aim was to estimate mutation rates of VNTRs specific for verotoxigenic *E. coli* O157:H7. The knowledge of VNTR mutational rates and the factors affecting them would make MLVA more effective for epidemiological or microbial forensic investigations. For this purpose, we analyzed nine loci performing parallel, serial passage experiments (PSPEs) on 9 O157:H7 strains. The combined 9 PSPE population rates for the 8 mutating loci ranged from 4.4×10^{-05} to 1.8×10^{-03} mutations/generation, and the combined 8-loci mutation rate was of 2.5×10^{-03} mutations/generation. Mutations involved complete repeat units, with only one point mutation detected. A similar proportion between single and multiple repeat changes was detected. Of the 56 repeat mutations, 59% were insertions and 41% were deletions, and 72% of the mutation events corresponded to O157-10 locus. For alleles with up to 13 UR, a constant and low mutation rate was observed; meanwhile longer alleles were associated with higher and variable mutation rates. Our results are useful to interpret data from microevolution and population epidemiology studies and particularly point out that the inclusion or not of O157-10 locus or, alternatively, a differential weighting data according to the mutation rates of loci must be evaluated in relation with the objectives of the proposed study.

1. Introduction

Repetitive DNA, which occurs in large quantities in eukaryotic cells, has been increasingly identified in prokaryotes [1]. Repeats organized in tandem, representing a single locus and showing interindividual unit number variability, are designated variable-number tandem repeat (VNTR) loci [2]. Repeat copy number variation at these loci is the consequence of mutations resulting in the gain or loss of a certain number of tandem repeats (TRs) and can lead to a very large number of alleles [1]. These mutations are thought to occur via an intramolecular slipped-strand mispairing (SSM), which may occur in combination with inadequate DNA mismatch repair pathways [3]. Recombination may also play a role, especially in mutations involving large numbers of repeat units [4], although SSM is thought to be the predominant mutational mechanism [5, 6].

The VNTR regions have been successfully used for subtyping purposes [1, 7], and multiple-locus VNTR analysis (MLVA) involves determination of the number of repeat copy units in multiple loci [8]. Molecular epidemiology, the integration of molecular typing and conventional epidemiological studies, likely adds significant value to analyses of infections caused by pathogenic bacteria [9]. MLVA has led to the development of a highly effective typing system for use in molecular epidemiology and forensic analyses [10, 11]. In recent emergent pathogens, TRs are a source of very informative markers for strain genotyping [12, 13]. However, any change in the rate of hypervariability in these TRs is problematic when trying to predict the phylogenetic relationships among isolates [14]. Noller et al. [15] observed single locus variants during some outbreaks, and, therefore, one potential concern is that some VNTRs evolve so rapidly that multiple

MLVA profiles would emerge during an outbreak initially caused by a single clone. In relation to it, Hyytiä-Trees et al. [8] proposed that during an outbreak two isolates differing in one repeat unit at one or two loci could be considered as related isolates. This is supported by the observation that isolates with five or more repeat differences at one or more loci were epidemiologically unrelated. Therefore, care has to be taken to select those repeat regions that show adequate stability over the timeframe in which the epidemiological investigations take place [16].

The necessity of more information on mutations and their rates in different loci to develop appropriate models to interpret MLVA data has been pointed out in several studies [8, 17–19]. Understanding of VNTR mutational rates and the factors affecting them would make MLVA more effective for epidemiological or microbial forensic investigations. Mutation rate data are valuable resources for probabilistic modeling of genetic relatedness, assessing the significance of finding genotype matches, near-matches, and mismatches [18, 20]. This information can be useful to determine the source of an outbreak, to asses true versus fortuitous disease clusters and to identify sites of exposure for human infection, among other purposes [20, 21].

In the present work our aim was to estimate mutation rates of *Escherichia coli* O157:H7 specific VNTR loci. For this purpose, we analyzed nine loci performing parallel, serial passage experiments (PSPEs) on nine verotoxigenic *E. coli* (VTEC) O157:H7 strains.

2. Materials and Methods

2.1. Strains. VTEC strains used in this study are listed in Table 1. The strains belong to O157:H7 serotype and were selected from those previously studied by Bustamante et al. [24] because each one of them represents a unique MLVA profile. Strains were stored at −80°C.

Table 1: *E. coli* O157:H7 strains.

Strain	Original source	MLVA profile*
EDL 933	Reference strain	10, 8, 11, 3, 5, 6, 13, 23, 7
HT2-15	Ground beef	9, 8, 70, 70, 5, 8, 4, 11, 7
FB3	Feedlot cattle	8, 11, 8, 2, 5, 7, 12, 40, 6
FB22	Feedlot cattle	8, 9, 6, 2, 6, 9, 6, 43, 6
FB81	Feedlot cattle	8, 9, 6, 2, 6, 10, 6, 25, 6
FC O157	Feedlot cattle	10, 8, 70, 1, 4, 9, 7, 23, 7
665p	Grazing cattle	9, 11, 6, 2, 5, 9, 7, 27, 6
Gal26	Human	6, 70, 12, 3, 3, 8, 3, 70, 9
Mat 167/6	Human	9, 15, 70, 1, 4, 8, 8, 28, 7

*String order: Vhec2, Vhec4, Vhec7, TR3, TR4, TR7, O157-3, O157-10, and O157-37. The value represents the number of TRs at each locus. Vhec loci are from Lindstedt et al. [22], TR loci from Noller et al. [15], and O157-*n* loci from Keys et al. [23]. Null alleles were designated as 70.

2.2. PSPE and Detection of Mutations. Parallel, serial passage experiments (PSPEs) have been previously described for *Yersinia pestis* [25] and for *Escherichia coli* O157:H7 [20]. We performed similar PSPEs on 9 *E. coli* O157:H7 strains (Table 1), generating a set of *in vitro* populations representing ~211.500 total generations (23.5 generations/colony × 100 lineages × 10 passages × 9 PSPEs). In our work, each PSPE consisted of ~100 independent clonal lineages that were each serially passaged 10 times. For each PSPE, a single isolated colony of each strain ($T = 0$) was used to start ~100 independent clonal lineages by streaking for single colonies on LB agar plates. All cultures were grown at 37°C for 24 h before the next passage. Each lineage was then serially passaged 10 times by streaking from a single colony from the previous passage. DNA was extracted from all ~100 lineages at $T = 10$ (day 10) in the 9 PSPEs by lysis of a single colony. Mutational events for each strain were visualized by MLVA, as previously described by Bustamante et al. [24]. Each lineage was tested by MLVA before and after all the passages. In order to know if the mutations were due to changes in repeat copy number or to point mutation events, the mutational products and parental alleles were sequenced. Mutational events and products for the nine strains were determined from an analysis of the $T = 10$ populations.

2.3. Mutation Rate Calculations and Genetic Diversity Analysis after the PSPE. Single-locus and combined 9-loci mutation rates (μ) were estimated for each strain by dividing the observed number of mutations by the number of total generations (GT) [20]. GT was calculated as the average number of generations per colony × the number of lineages with usable data (n) × the number of transfers. The number of generations per colony was determined to be 23.5 using an average of viable plate counts. In the combined single locus mutation rates, n was based on an average of n across all nine strains for each locus. To calculate the mutation rate in the combined 9 loci, n was based on an average of n across all 9 loci for each individual strain calculation and an average of n across all 9 loci and all nine strains for the combined strain calculations.

In order to analyze the relationship between mutation rate and repeat number, progressive linear regression analyses (PROC REG, Statistical Analysis System, Version 9.2) were performed to detect the range of repeat numbers where the slope of the regression line was not significantly different from zero.

A dendrogram showing the genetic variability cumulated during the PSPE, taking into account each strain at $T = 0$ with its derived mutated lineages, was constructed by UPGMA clustering using START version 1.0.5 [26].

3. Results and Discussion

Mutation is the major mechanism for generating diversity in clonal organisms, and VNTR loci are ones of the fastest mutation sequences that allow detecting genetic differences. In this study, we report the mutation dynamics of nine VNTR loci used for VTEC O157 subtyping. Eight of the nine VNTRs of this MLVA protocol have repeat units (RU) of six or seven bp and the remaining one (Vhec2), 18 bp. The genomic regions susceptible to intramolecular slipped-strand mispairing (SSM) are those that contain a short, contiguous homogenous or heterogeneous repetitive DNA sequence of 6 bp or less [1]. Repeat motifs longer than eight base pairs show typically less unit number variation but are more likely to have point mutations [1]. In this study, from the nine VNTRs analyzed, eight mutated across all PSPEs; meanwhile the Vhec2 locus did not exhibit changes in repeat number, in accordance with data from other authors [14, 17, 20]. Noller et al. [17], using a different experimental model, analyzed seven of the 9 VNTRs studied here but they only observed mutation in three VNTRs (Vhec4, O157-3, and O157-10).

In order to detect rare occurrence of mutation before $T = 10$, in some PSPEs (5 out 9) DNA was also extracted at $T = 5$ (day 5); however, we did not detect any mutations at this earlier time point.

In the nine *in vitro* PSPE populations, we observed 57 mutational events corresponding to a combined 8-loci mutation rate of 2.5×10^{-03} mutations/generation, and noticeably, 41 (72%) of the mutation events corresponded to locus O157-10 (Table 2). Comparison between mutated and parental allele sequences confirmed that mutations were changes involving complete repeat units, with only one exception at a single locus in one PSPE (Vhec4, FC O157) in which the change was due to a point mutation (Table 2). Of the 56 repeat mutations, 33 (59%) were insertions and 23 (41%) were deletions (Table 2). This proportion of insertion/deletion of repeats is similar to that observed by Vogler et al. [20], 62% insertions and 38% deletions, respectively. They explained the insertion bias as due to the influence of a single locus, O157-10, in a single PSPE, but in our study, on the contrary, when O157-10 was excluded from the analysis the proportion of insertions increased to 67%.

Several single and multiple repeat events could be observed (Figures 1 and 2), 29 (52%) mutation events were changes due to single repeats (either insertion or deletion), and the remaining 27 (48%) involved multiple repeat changes (Table 2). This similar proportion between single and multiple repeat changes differs with one of the general properties of the mutation model for bacterial VNTRs proposed by Vogler et al. [18], who indicated that the majority of mutations involve single repeats, although a certain percentage of events consist of multiple repeats. In agreement with the results of Vogler et al. [20], the frequency of mutations followed a decreasing pattern from 1 to 3 RU (or to 5 RU if only the deletions are taken into account), and the frequency of deletions of more than 5 RU did not seem to follow any pattern (Figure 2). In the multiple repeat events the changes involved up to 23 RU, and 14 of the 27 multiple repeat mutations were "large repeat copy number" events. This last term was defined by Vogler et al. [20] in order to name mutational events involving greater than four repeat units. The occurrence of "large repeat copy number" events observed in this study (14/27 = 52%) was markedly higher than the one obtained by Vogler et al. [20] (12/47 = 26%). Interestingly, all "large repeat copy number" events detected by us were at locus O157-10, with insertions and deletions occurring at nearly the same frequency. Noticeably, two of the deletions comprised the complete VNTR; meanwhile in all the insertion events, 5 RU were inserted.

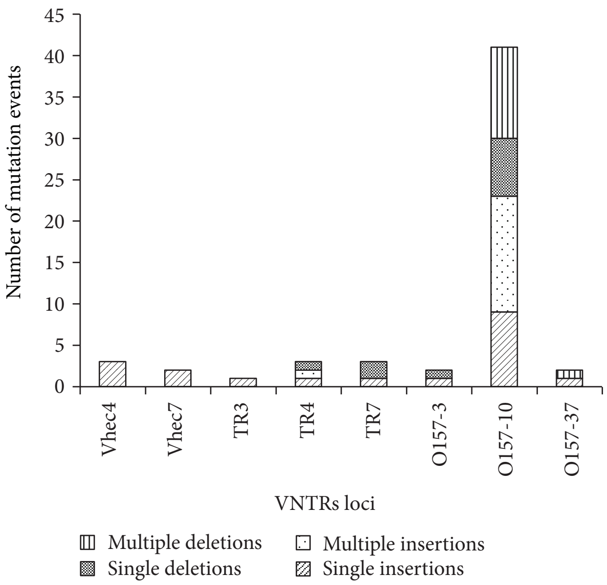


Figure 1: Number and kind of mutational events in each locus.

Long microsatellites are likely to mutate to shorter ones, and short microsatellites are likely to mutate to longer ones [27]. Considering the $T = 0$ MLVA profile of the different strains analyzed, the largest repeat copy numbers were present at locus O157-10, with 7 alleles comprising more than 22 RU, while in the remaining analyzed loci the largest allele was of 15 RU. VNTR alleles of very large size could be detrimental or just highly unstable, leading to the observation of a greater number of deletion events [20], but in our study, although we observed that the VNTR O157-10 was the most unstable, we did not observe a trend towards the deletions (23 insertions versus 18 deletions).

The combined 9 PSPE population rates for the 8 mutating loci ranged from 4.4×10^{-05} to 1.8×10^{-03} mutations/generation (Table 2). Our experimental design allowed us to detect mutation rates equal or higher than 4.4×10^{-05} mutations/generation. No mutation was observed at Vhec2 suggesting that its mutation rate is near to or less than that value, which is the detection limit given the number of generations in the PSPE population.

The mutation rates at locus O157-10 ranged from 1.3×10^{-04} to 4.8×10^{-04} mutations/generation among the nine PSPE strains (Table 2), and this locus also presented the highest mutation rate when the results of all the strains were taken into account (1.8×10^{-03} mutations/generation) (Table 2). Forty-one of the 57 (72%) observed mutations occurred at this locus, and when mutations associated to this locus were removed from the combined data set, the mutation rate (excluding O157-10) decreased noticeably (6.9×10^{-04} mutations/generation). Thus, our results, as those from Vogler et al. [20], showed that allele sizes at locus O157-10 were the most variable and therefore impacted strongly on the combined locus mutation rates.

Table 2: Mutation rates and products for 9 *E. coli* O157:H7 strains.

Strains and loci	RU (bp)	n[a]	Total no. of mutations	Mutation rate	No. of insertions		No. of deletions	
					S[b]	M[c]	S[b]	M[c]
HT2-15								
TR4	6	99	1	4.3×10^{-05}	0	1	0	0
O157-10	6	96	5	2.2×10^{-04}	4	1	0	0
Total			**6**	**2.6×10^{-04}**	**6**		**0**	
EDL933								
Vhec7	7	99	2	8.6×10^{-05}	2	0	0	0
TR3	6	97	1	4.4×10^{-05}	1	0	0	0
TR4	6	97	1	4.4×10^{-05}	1	0	0	0
TR7	6	97	1	4.4×10^{-05}	1	0	0	0
O157-3	6	100	2	8.5×10^{-05}	1	0	1	0
O157-10	6	100	3	1.3×10^{-04}	0	2	0	1
Total			**10**	**4.3×10^{-04}**	**8**		**2**	
FB3								
O157-10	6	90	8	3.8×10^{-04}	1	2	3	2
Total			**8**	**3.8×10^{-04}**	**3**		**5**	
Mat167/6								
TR4	6	100	1	4.3×10^{-05}	0	0	1	0
O157-10	6	100	6	2.5×10^{-04}	0	5	1[d]	0
O157-37	6	100	1	4.3×10^{-05}	1	0	0	0
Total			**8**	**3.4×10^{-04}**	**6**		**2**	
Gal 26								
O157-37	6	100	1	4.3×10^{-05}	0	0	0	1
Total			**1**	**4.3×10^{-05}**	**0**		**1**	
FC O157								
Vhec4	6	100	1[e]	4.3×10^{-05}	0	0	0	0
TR7	6	100	2	8.5×10^{-05}	0	0	2	0
O157-10	6	89	10	4.8×10^{-04}	1	3	0	6
Total			**13**	**7.1×10^{-04}**	**4**		**8**	
FB22								
O157-10	6	92	5	2.3×10^{-04}	1	1	3	0
Total			**5**	**2.3×10^{-04}**	**2**		**3**	
FB81								
Vhec4	6	92	1	4.6×10^{-05}	1	0	0	0
Total			**1**	**4.6×10^{-05}**	**1**		**0**	
665p								
Vhec4	6	100	1	4.3×10^{-05}	1	0	0	0
O157-10	6	100	4	1.7×10^{-04}	2	0	0	2
Total			**5**	**2.1×10^{-04}**	**3**		**2**	
Combined data for the nine PSPEs					Insertions		Deletions	
					S[b]	M[c]	S[b]	M[c]
Vhec4		97.3	3	1.3×10^{-04}	2	0	0	0
Vhec7		99	2	8.6×10^{-05}	2	0	0	0
TR3		97	1	4.4×10^{-05}	1	0	0	0
TR4		98.7	3	1.3×10^{-04}	1	1	1	0
TR7		98.5	3	1.3×10^{-04}	1	0	2	0
O157-3		100	2	8.5×10^{-05}	1	0	1	0

Table 2: Continued.

Strains and loci	RU (bp)	*n*[a]	Total no. of mutations	Mutation rate	No. of insertions		No. of deletions	
					S[b]	M[c]	S[b]	M[c]
O157-10		95.3	41	1.8×10^{-03}	9	14	7	11
O157-37		100	2	8.5×10^{-05}	1	0	0	1
					18	15	11	12
Total			57	2.5×10^{-03}	33		23	

[a] Number of lineages with usable data.

[b] Single repeat mutation.

[c] Multiple repeat mutation.

[d] Deletion of one TR and insertion of 36 pb aleatory sequence.

[e] Point mutation.

It has been described in different bacterial species that mutation rates correlate with the number of repeats present in the VNTRs [18, 20, 28]. Taking into account the present results, the distribution of the mutation rates versus the number of RU present in the corresponding strain at $T = 0$ is shown in Figure 3. It can be observed that for alleles with up to 13 UR a more or less constant and low mutation rate corresponds to them; meanwhile longer alleles are associated with higher and variable mutation rates (values higher than 1×10^{-04}). Furthermore, multiple repeat changes predominated among this second group of alleles, which correspond mainly to the locus O157-10. When we compared our results with those obtained by other authors who analysed VNTRs from VTEC O157:H7, *Yersinia pestis,* and *Vibrio parahaemolyticus* [18, 20, 28], we noticed that the distribution was similar, although they did not discuss this and instead focused on the correlation between mutation rates and repeat copy numbers. However, as O157-10 was the only locus which presented large alleles, its influence cannot be ruled out, and therefore we cannot make generalized conclusions about allele size and mutation rate.

In the present study, when results were compared among strains, strain EDL 933 noticeably showed mutations in 6 VNTRs, followed by Mat167/6 and FC O157 which presented variation in only 3 loci. On the other hand, two strains showed repeat number variation in only one locus (O157-10). Mutation rates were not so different among strains, but when they were recalculated without the inclusion of O157-10 locus, EDL 933 strain showed the highest rate followed by FC O157. On the contrary, Vogler et al. [20] did not observe a pure “strain” effect on mutation rate and indicated that instead allele size at individual loci did affect the mutation rate. Furthermore, the MLVA profile reported for the EDL 933 strain differed among several studies [8, 22, 23, 29] and also with an *in silico* analysis with the GenBank sequence (accession no. AE005174). These differences are probably due to repeated propagation of the strain and could also reflect a high mutation degree of the EDL 933 strain.

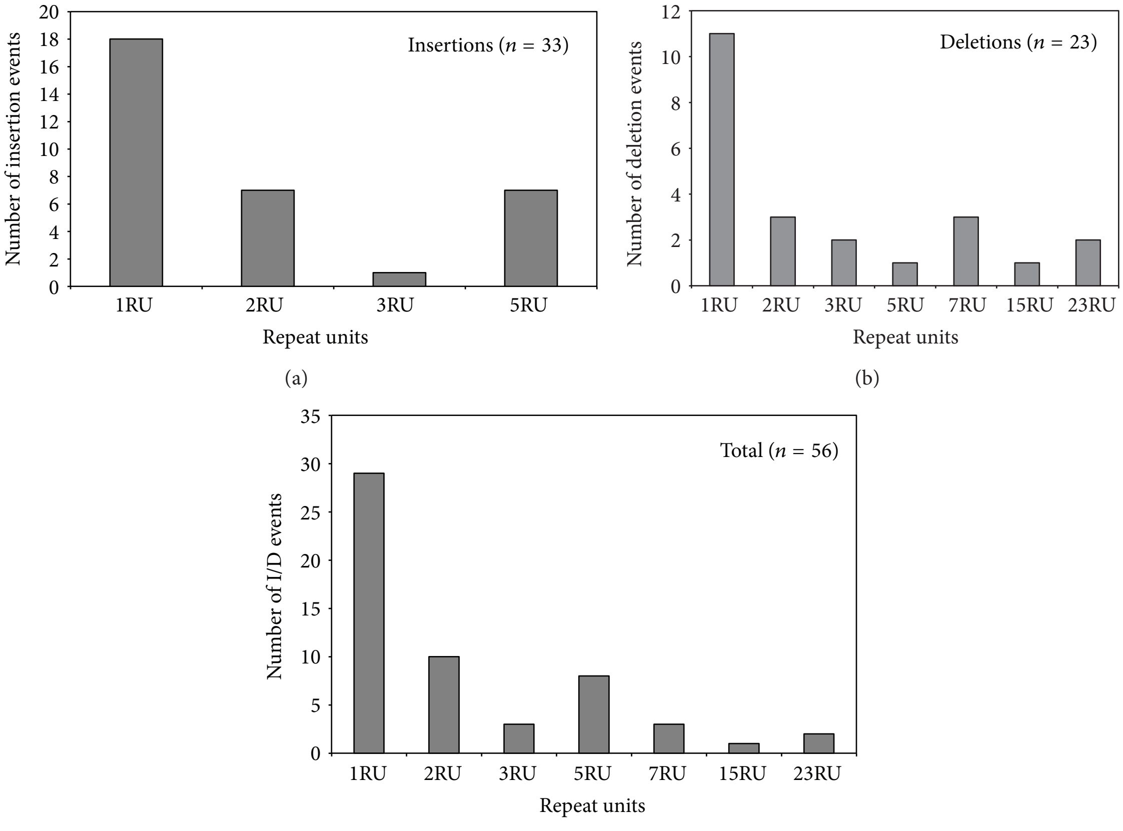
MLVA is applied to study the relationships among different isolates and as an aid to determine the presence of an outbreak. The analysis of our *in vitro* PSPE population showed that each parental strain and its derived lineages became clustered evidencing their clonal relationship. The dendrogram showing the original MLVA profiles (at $T =$ 0) of the strains and those of the mutated lineages is shown in Figure 4.

Cooley et al. [14] studied mutations in stressed bacteria and concluded that conditions such as elevated temperature and starvation are associated with mutations. They observed that tandem repeats were stable after limited replication, but in our study the strains were not subjected to selection pressure, and, however, we observed mutations in several VNTRs. Interestingly, their results showed that one locus (O157-10) mutated in all the conditions tested. Several studies recognized O157-10 as the most polymorphic locus [15, 19, 24, 30–33]. Furthermore, when the Pulse Net USA subtyping network recently established a standardized protocol in order to characterize verotoxigenic *Escherichia coli* O157 by MLVA, this locus was excluded because it showed too much variability during outbreaks, generating data that confounded epidemiological investigations [33]. However, Keys et al. [23] previously suggested that markers with high diversities are crucial in discriminating between closely related isolates, and they highlighted the importance of including O157-10, as an example of this kind of crucial markers. In relation with it, Urdahl et al. [19] observed that removing this locus from the analysis resulted in an MLVA that did not discriminate between closely related strains.

In summary, taking into account different points of view of several authors and our own results, care must be taken to select loci in order to be included in MLVA and to interpret data obtained from highly discriminating loci. Particularly, some aspects such as the inclusion or not of O157-10 locus in the MLVA or alternatively, a differential weighting data according to the mutation rates, have to be considered in relation with the objectives of a certain study.

4. Conclusions

Our set of *in vitro* populations showed that mutation rates detected for nine VNTR loci used for VTEC O157 subtyping were according to those reported previously for other bacterial VNTRs, with a similar proportion between single and multiple repeat changes. A markedly high proportion of “large repeat copy number” events was observed in this study.



(a)

(b)

(c)

Figure 2: Frequency distributions of mutation products implicating complete repeat units. (a) Insertions; (b) deletions; (c) total.

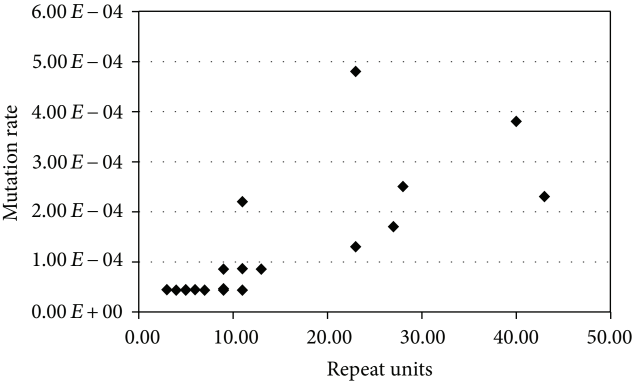


Figure 3: Distribution of mutation rates in relation to the number of repeats in each mutating locus. Progressive linear regression analysis (PROC REG, Statistical Analysis System, Version 9.2) showed that for alleles containing up to 13 UR the slope of the regression line was not significantly different from zero ($P > 0.07$).

Long alleles were associated with high and variable mutation rates and corresponded to locus O157-10, which presented an outstanding high mutation rate.

Our results are useful to interpret data from microevolution and population epidemiology studies and suggest, together with previously published ones, that a differential weighting of VNTR data should be applied according to the available mutation rates of loci in order to get more accurate phylogenetic relationships.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

The authors thank María Rosa Ortiz for technical assistance and Juan Passucci and Edgardo Rodríguez for statistical analysis. This work was supported by grants from the CONICET, FONCYT, and SECAT-UNICEN.

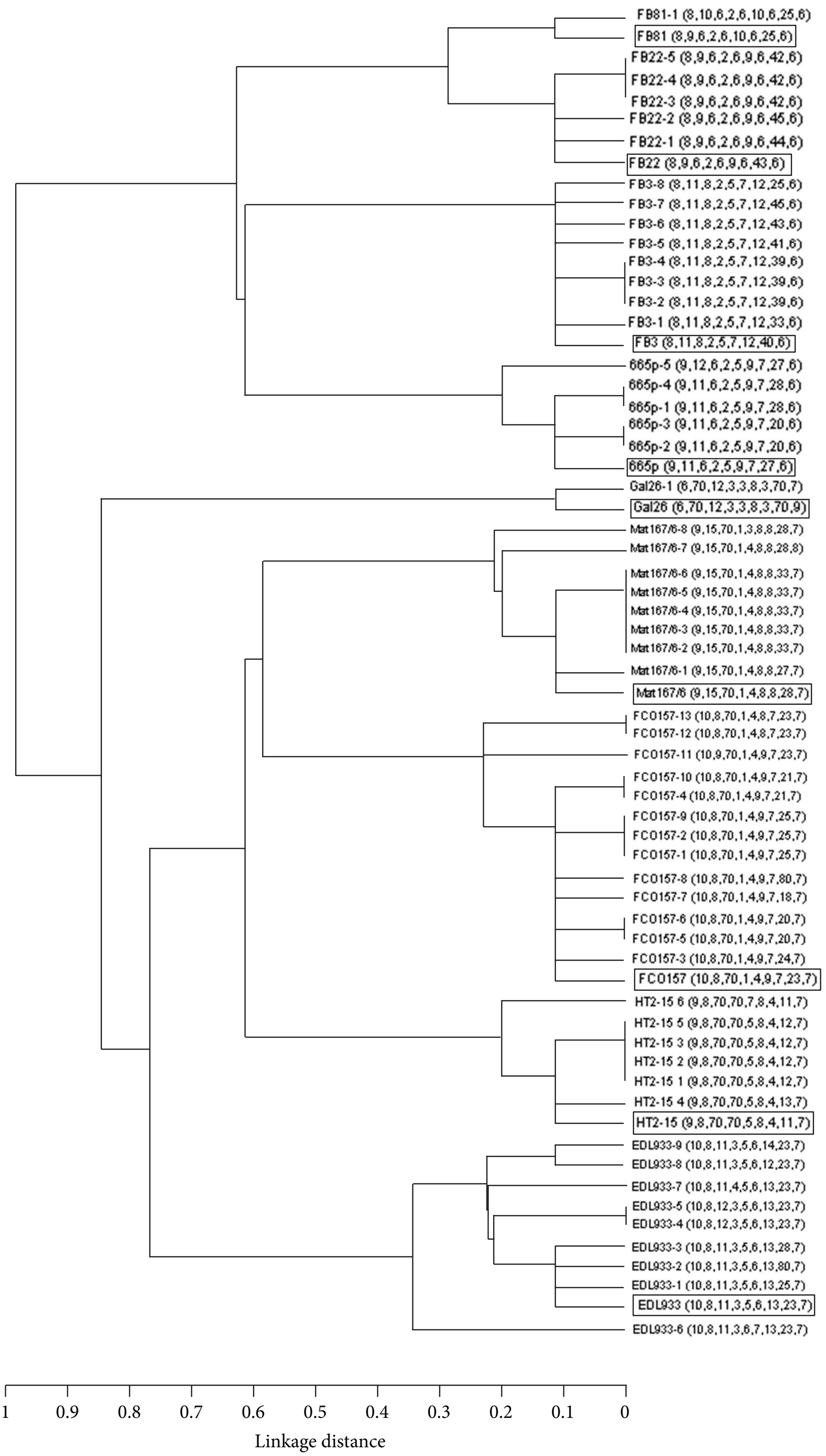


Figure 4: Dendrogram showing the genetic variability cumulated during the PSPE, taking into account each strain at $T = 0$ with its derived mutated lineages. String order: Vhec2, Vhec4, Vhec7, TR3, TR4, TR7, O157-3, O157-10, and O157-37. Boxed profiles correspond to the parental strains. Null alleles are indicated with 70 and with 80 when an amplification product was detected but the VNTR was missing.

References

[1] A. Van Belkum, S. Scherer, L. Van Alphen, and H. Verbrugh, "Short-sequence DNA repeats in prokaryotic genomes," *Microbiology and Molecular Biology Reviews*, vol. 62, no. 2, pp. 275–293, 1998.

[2] Y. Nakamura, M. Leppert, and P. O'Connell, "Variable number of tandem repeat (VNTR) markers for human gene mapping," *Science*, vol. 235, no. 4796, pp. 1616–1622, 1987.

[3] S. Strand, T. A. Prolla, R. M. Liskay, and T. D. Petes, "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair," *Nature*, vol. 365, pp. 274–276, 1993.

[4] G. Levinson and G. A. Gutman, "Slipped-strand mispairing: a major mechanism for DNA sequence evolution," *Molecular Biology and Evolution*, vol. 4, no. 3, pp. 203–221, 1987.

[5] H. Ellegren, "Microsatellite mutations in the germline: implications for evolutionary inference," *Trends in Genetics*, vol. 16, no. 12, pp. 551–558, 2000.

[6] C. Schlötterer, "Evolutionary dynamics of microsatellite DNA," *Chromosoma*, vol. 109, pp. 365–371, 2000.

[7] Y. Kashi, D. King, and M. Soller, "Simple sequence repeats as a source of quantitative genetic variation," *Trends in Genetics*, vol. 13, no. 2, pp. 74–78, 1997.

[8] E. Hyytiä-Trees, S. C. Smole, P. A. Fields, B. Swaminathan, and E. M. Ribot, "Second generation subtyping: a proposed PulseNet protocol for multiple-locus variable-number tandem repeat analysis of Shiga toxin-producing *Escherichia coli* O157 (STEC O157)," *Foodborne Pathogens and Disease*, vol. 3, no. 1, pp. 118–131, 2006.

[9] F. Denœud and G. Vergnaud, "Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource," *BMC Bioinformatics*, vol. 5, article 4, 2004.

[10] J. M. Doll, P. S. Zeitz, P. Ettestad, A. L. Bucholtz, T. Davis, and K. Gage, "Cat-transmitted fatal pneumonic plague in a person who traveled from Colorado to Arizona," *American Journal of Tropical Medicine and Hygiene*, vol. 51, no. 1, pp. 109–114, 1994.

[11] A. M. Klevytska, L. B. Price, J. M. Schupp, P. L. Worsham, J. Wong, and P. Keim, "Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome," *Journal of Clinical Microbiology*, vol. 39, no. 9, pp. 3179–3185, 2001.

[12] A. Van Belkum, S. Scherer, W. Van Leeuwen, D. Willemse, L. Van Alphen, and H. Verbrugh, "Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*," *Infection and Immunity*, vol. 65, no. 12, pp. 5017–5027, 1997.

[13] P. Le Flèche, M. Fabre, F. Denoeud, J.-L. Koeck, and G. Vergnaud, "High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing," *BMC Microbiology*, vol. 2, no. 1, article 37, 2002.

[14] M. B. Cooley, D. Carychao, K. Nguyen, L. Whitehand, and R. Mandrell, "Effects of environmental stress on stability of tandem repeats in *Escherichia coli* O157:H7," *Applied and Environmental Microbiology*, vol. 76, no. 10, pp. 3398–3400, 2010.

[15] A. C. Noller, M. C. McEllistrem, A. G. F. Pacheco, D. J. Boxrud, and L. H. Harrison, "Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates," *Journal of Clinical Microbiology*, vol. 41, pp. 5389–5397, 2003.

[16] A. Van Belkum, "Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA)," *FEMS Immunology and Medical Microbiology*, vol. 49, no. 1, pp. 22–27, 2007.

[17] A. C. Noller, M. C. McEllistrem, K. A. Shutt, and L. H. Harrison, "Locus-specific mutational events in a multilocus variable-number tandem repeat analysis of *Escherichia coli* O157:H7," *Journal of Clinical Microbiology*, vol. 44, no. 2, pp. 374–377, 2006.

[18] A. J. Vogler, C. E. Keys, C. Allender et al., "Mutations, mutation rates, and evolution at the hypervariable VNTR loci of *Yersinia pestis*," *Mutation Research*, vol. 616, no. 1-2, pp. 145–158, 2007.

[19] A. M. Urdahl, N. J. C. Strachan, Y. Wasteson, M. MacRae, and I. D. Ogden, "Diversity of *Escherichia coli* O157 in a longitudinal farm study using multiple-locus variable-number tandem-repeat analysis," *Journal of Applied Microbiology*, vol. 105, no. 5, pp. 1344–1353, 2008.

[20] A. J. Vogler, C. Keys, Y. Nemoto, R. E. Colman, Z. Jay, and P. Keim, "Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7," *Journal of Bacteriology*, vol. 188, no. 12, pp. 4253–4263, 2006.

[21] J. L. Lowell, D. M. Wagner, B. Atshabar et al., "Identifying sources of human exposure to plague," *Journal of Clinical Microbiology*, vol. 43, pp. 650–656, 2005.

[22] B.-A. Lindstedt, E. Heir, E. Gjernes, T. Vardund, and G. Kapperud, "DNA fingerprinting of shiga-toxin producing *Escherichia coli* O157 based on multiple-locus variable-number tandem-repeats analysis (MLVA)," *Annals of Clinical Microbiology and Antimicrobials*, vol. 2, article 12, 2003.

[23] C. Keys, S. Kemper, and P. Keim, "Highly diverse variable number tandem repeat loci in the *E. coli* O157:H7 and O55:H7 genomes for high-resolution molecular typing," *Journal of Applied Microbiology*, vol. 98, no. 4, pp. 928–940, 2005.

[24] A. V. Bustamante, P. M. A. Lucchesi, and A. E. Parma, "Molecular characterization of verocytotoxigenic *Escherichia coli* O157:H7 isolates from Argentina by multiple-loci VNTR analysis (MLVA)," *Brazilian Journal of Microbiology*, vol. 40, no. 4, pp. 927–932, 2009.

[25] J. M. Girard, D. M. Wagner, A. J. Vogler et al., "Differential plague-transmission dynamics determine *Yersinia pestis* population genetic structure on local, regional, and global scales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 22, pp. 8408–8413, 2004.

[26] K. A. Jolley, E. J. Feil, M.-S. Chan, and M. C. J. Maiden, "Sequence type analysis and recombinational tests (START)," *Bioinformatics*, vol. 17, no. 12, pp. 1230–1231, 2002.

[27] Y. Lai and F. Sun, "The relationship between microsatellite slippage mutation rate and the number of repeat units," *Molecular Biology and Evolution*, vol. 20, no. 12, pp. 2123–2131, 2003.

[28] K. García, R. G. Gavilán, M. G. Höfle, J. Martínez-Urtaza, and R. T. Espejo, "Microevolution of pandemic *Vibrio parahaemolyticus* assessed by the number of repeat units in short sequence tandem repeat regions," *PLoS One*, vol. 7, no. 1, Article ID e30823, 2012.

[29] L. Yun, Y. Gu, L. Zha, F. Zhu, and Y. Hou, "Utility of multilocus variable number tandem repeat analysis as a microbial forensic tool for subtyping Chinese *Escherichia coli* O157: H7 strains," *Forensic Science International*, vol. 3, no. 1, supplement 3, pp. e293–e294, 2011.

[30] B.-A. Lindstedt, T. Vardund, and G. Kapperud, "Multiple-locus variable- number tandem-repeats analysis of *Escherichia coli* O157 using PCR multiplexing and multi-colored capillary electrophoresis," *Journal of Microbiological Methods*, vol. 58, no. 2, pp. 213–222, 2004.

[31] F. Kawamori, M. Hiroi, T. Harada et al., “Molecular typing of Japanese *Escherichia coli* O157:H7 isolates from clinical specimens by multilocus variable-number tandem repeat analysis and PFGE,” *Journal of Medical Microbiology*, vol. 57, no. 1, pp. 58–63, 2008.

[32] M. Murphy, D. Minihan, J. F. Buckley, M. O’mahony, P. Whyte, and S. Fanning, “Multiple-Locus Variable number of tandem repeat Analysis (MLVA) of Irish verocytotoxigenic *Escherichia coli* O157 from feedlot cattle: uncovering strain dissemination routes,” *BMC Veterinary Research*, vol. 4, article 2, 2008.

[33] E. Hyytiä-Trees, P. Lafon, P. Vauterin, and E. M. Ribot, “Multilaboratory validation study of standardized multiple-locus variable-number tandem repeat analysis protocol for Shiga toxin-producing *Escherichia coli* O157: a novel approach to normalize fragment size data between capillary electrophoresis platforms,” *Foodborne Pathogens and Disease*, vol. 7, no. 2, pp. 129–136, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 413450, 6 pages
http://dx.doi.org/10.1155/2013/413450

Research Article

Transcription Regulation of Plastid Genes Involved in Sulfate Transport in Viridiplantae

Vassily A. Lyubetsky, Alexander V. Seliverstov, and Oleg A. Zverkov

Institute for Information Transmission Problems (Kharkevich Institute), The Russian Academy of Sciences, Moscow 127994, Russia

Correspondence should be addressed to Vassily A. Lyubetsky; lyubetsk@iitp.ru

Received 1 April 2013; Accepted 12 June 2013

Academic Editor: William Piel

This study considers transcription regulation of plastid genes involved in sulfate transport in the parasites of invertebrate (*Helicosporidium* sp.) and other species of the Viridiplantae. A one-box conserved motif with the consensus TAAWATGATT is found near promoters upstream the *cysT* and *cysA* genes in many species. In certain cases, the motif is repeated two or three times.

1. Introduction

This study focuses on selected species of the Viridiplantae, particularly, the genus *Helicosporidium* sp. (class Trebouxiophyceae), which comprises green algae parasitizing flies of the species *Simulium jonesi* [1–3]. Plastids of these parasites are a good target for antibiotic treatment, as earlier was shown for apicomplexan parasites of vertebrates (*Toxoplasma gondii* and *Plasmodium* spp. [4]).

The plastome of *Helicosporidium* sp. is relatively small, about 37 kb. Most of the plastome genes encode tRNA, rRNA, ribosomal proteins, and subunits of the bacterial-type RNA polymerase. One of two nonhousekeeping proteins is the CysT subunit of a sulfate ABC transporter.

Sulfate ABC transporters in cyanobacteria and proteobacteria consist of two identical ATP-binding CysA proteins, two transmembrane proteins (CysT and CysW), and a sulfate-binding protein SbpA. In the cyanobacteria *Synechocystis* sp. PCC 6803 [5], genes encoding the sulfate transporter subunits are arranged in a single operon *sbpA-ssr2439-cysT-cysW-cysA*. In cyanobacteria, no data on expression is available for this operon; however, in *Escherichia coli* and some other proteobacteria, genes of the sulfate transporter subunits are known to be regulated in the single operon *cysPTWAM* (further details are given in Discussion).

Plastomes of vascular plants lack genes of the sulfate transport system except for rare instances of *cysT* and *cysA*. However, the green alga *Helicosporidium* sp. retains *cysT*. Plastomes of the rhodophyte *Cyanidium caldarium* and *Cyanidioschyzon merolae* and the cyanelle genome of *Cyanophora paradoxa* lack *cysT* homologues but possess distant homologues of *cysA* presumably involved in the transport of zinc or manganese (further details are given in Results). Similar proteins are involved in the transport of molybdenum, zinc, and manganese and belong to a large family of transporters of ions, sugars, peptides, and more complex organic molecules. For example, the transcription regulation of the *ziaA* gene (encoding a polypeptide similar to a P-type ATPases involved in transporting heavy metals) is described in the cyanobacterium *Synechocystis PCC 6803* [6].

The sulfate transport in plastids is necessary for the synthesis of many sulfur-containing compounds. For example, in *Spinacia oleracea*, the lack of sulfates leads to considerable changes in the expression of cysteine synthesis genes [7]. Also, plastids of many algae synthesize thiamine and other sulfur-containing compounds. For example, the lipoic acid is synthesized in plastids of apicomplexan parasites [8].

In this paper, we consider the expression regulation of *cysT* and *cysA* in Viridiplantae, in particular, *Helicosporidium* sp. and *Pycnococcus provasolii*, where *cysT* is present and *cysA* is absent.

In proteobacteria, the regulation mechanism of transcription initiation of *cysA* and *cysT* is known. The CysB protein is a transcription factor of the LysR family and acts as a tetramer. This protein binds DNA upstream the −35 box of a promoter and activates transcription initiation of

Table 1: CysA and CysT proteins encoded in plastids of the Viridiplantae and in the cyanobacterium *Synechocystis* sp. PCC 6803.

Species	DNA locus	GC, %	CysA	CysT
Synechocystis sp. PCC 6803	NC_017277.1	49	YP_005652904.1	YP_005652902.1
Nephroselmis olivacea	NC_000927.1	42	NP_050872.1	NP_050928.1
Pycnococcus provasolii	NC_012097.1	40	None	YP_002600832.1
Bryopsis hypnoides	NC_013359.1	33	YP_003227066.1	YP_003227041.1
Chlorella variabilis	NC_015359.1	34	YP_004347745.1	YP_004347762.1
Chlorella vulgaris	NC_001865.1	32	NP_045832.1	NP_045890.1
Coccomyxa subellipsoidea	NC_015084.1	51	YP_004222028.1	YP_004221988.1
Helicosporidium sp.	NC_008100.1	27	None	YP_635918.1
Leptosira terrestris	NC_009681.1	27	YP_001382174.1	YP_001382135.1
Parachlorella kessleri	NC_012978.1	30	YP_003058285.1	YP_003058340.1
Mesostigma viride	NC_002186.1	30	NP_038429.1	NP_038441.1
Chlorokybus atmophyticus	NC_008822.1	36	YP_001019096.1	YP_001019170.1
Zygnema circumcarinatum	NC_008117.1	31	YP_636486.1	YP_636569.1
Anthoceros formosae	NC_004543.1	33	NP_777407.1	NP_777462.1
Aneura mirabilis	NC_010359.1	41	Pseudogene	Pseudogene
Marchantia polymorpha	NC_001319.1	29	NP_039293.1	NP_039346.1
Ptilidium pulcherrimum	NC_015402.1	33	Pseudogene	Pseudogene

many operons. For proteobacteria *Salmonella typhimurium*, *Escherichia coli* [9], and *Klebsiella aerogenes* [10], the CysB activation of *cysPTWAM*, *cysK*, *cysJIH*, *cysDNC*, *sbp*, and L-cysteine transport genes, as well as CysB autorepression is described in detail. Binding of CysB to DNA is not directly dependent on sulfate concentration but requires high concentrations of acetylserine. Also, proteobacteria lack a distinct binding motif for CysB.

2. Materials and Methods

The list of species is given in Table 1. Genomes were obtained from GenBank, NCBI. Clustering of proteins was performed using the method described in [11, 12].

An original algorithm from [13, 14] was employed to search for bacterial-type promoters. Relevant promoter sequences and the evolutionary impact of DNA point mutations on polymerase binding affinity are described in [15, 16]; experimental evidence was obtained using the *psbA* promoter of *Sinapis alba* [17].

A novel method based on clique search in a multipartite graph [18] was used to identify conserved motifs. In current modification of the method, the nucleotide similarity was estimated accounting for the GC content in plastid DNA. Namely, if the average GC-rate was p, the additive contribution for a mismatch at any position in the calculation of the distance between two words of equal lengths was $(1-p)$ for a A-T pair, p for a C-G pair, and $1/2$ for a S-W pair, where S = {C,G} and W = {A,T}. A large-scale search for binding sites was performed using formulas from [19]. Protein alignments were constructed with ClustalW v. 2.0.3 [20].

3. Results

3.1. Analysis of the Domain Structure and Phylogenetic Classification of Proteins. The *cysT* gene is present in the following species of Viridiplantae: Chlorophyta (*Bryopsis hypnoides, Nephroselmi solivacea, Pycnococcus provasolii, Chlorella variabilis, Chlorella vulgaris, Coccomyxa subellipsoidea C-169, Helicosporidium* sp. ex *Simulium jonesii, Leptosira terrestris,* and *Parachlorella kessleri*), Streptophyta (*Chlorokybus atmophyticus, Mesostigma viride,* and *Zygnema circumcarinatum*), Bryophyta (*Anthoceros formosae, Marchantia polymorpha,* and *Aneura mirabilis*), and *Ptilidium pulcherrimum* (*A.m.* and *P.p.* are pseudogenes) [21].

CysT proteins are conserved across green algae, cyanobacteria, and proteobacteria and function as the transmembrane domain of the ABC transporter (Figure 1). Short N-terminus regions of these proteins can vary. Another exception is CysT in *Bryopsis hypnoides* and *Leptosira terrestris* that have truncated C-termini.

Cyanobacteria possess strong potential orthologs of *cysT* in the Viridiplantae. Cyanobacterial CysT is considered to be ancestral.

As CysA and CysT functions are linked, *cysA* and *cysT* normally either coexist or are absent in plastids of all Viridiplantae with the exception of *Helicosporidium* sp. and *Pycnococcus provasolii*. (Note that the *cysA* ortholog in *Marchantia polymorpha* is named *mbpX*.) Viridiplantae species lacking *cysA* and *cysT* are closely related to the species that contain both genes [11, 22]. Unexpectedly, *cysA* and *cysT* are present in Bryophyta while they are absent in many highly organized algae close to land plants (*Chaetosphaeridium globosum*, *Chara vulgaris*, and *Staurastrum punctulatum*). They are also absent in plastomes of *Physcomitrella patens* and all vascular plants. In green algae, these genes are mainly present in the class Trebouxiophyceae (genera *Chlorella*, *Coccomyxa*, *Helicosporidium*, *Leptosira*, and *Parachlorella*).

Sequences of plastid CysA and their orthologs from cyanobacterium are well aligned (Figure 2). CysA in all Viridiplantae has a highly conservative N-terminus domain which is typical for the ATP-binding cassette

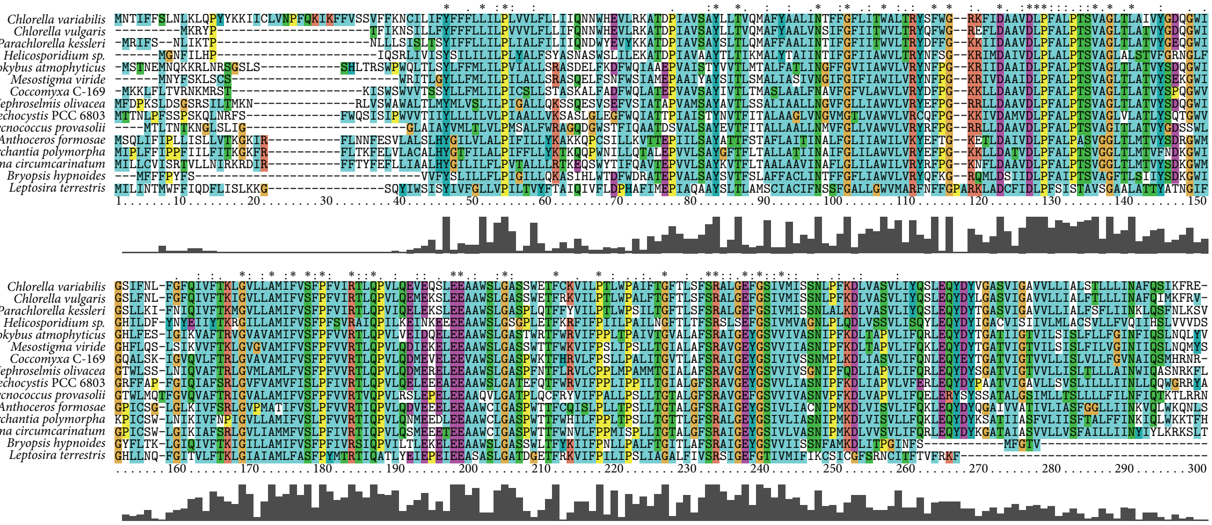


Figure 1: A multiple alignment of CysT orthologs from the cyanobacterium *Synechocystis* sp. PCC 6803 and plastids of the Viridiplantae. Conservativity is shown above and below the columns.

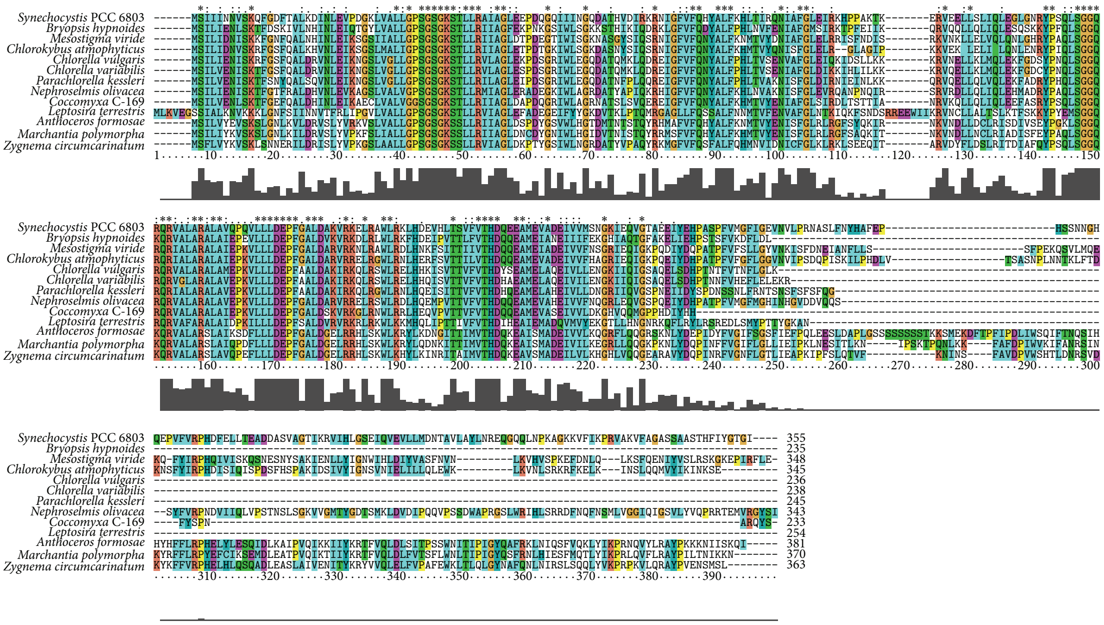


Figure 2: A multiple alignment of CysA orthologs from the cyanobacterium *Synechocystis* sp. PCC 6803 and plastids of the Viridiplantae.

of ABC transporters. In all studied Chlorophyta, except for *Nephroselmis olivacea,* this protein possesses a short C-terminus. Conversely, in the Streptophyta, *Nephroselmis oli-vacea*, and cyanobacteria, C-termini are long and conservative. In *Mesostigma viride* and *Chlorokybus atmophyticus*, this domain is homologous to the TOBE domain involved in sulfate binding [23]. According to the Pfam 26.0 database [24], *e*-value for this domain is 0.0017 in *M. viride* and 0.00007 in *Ch. atmophyticus.* Other plastid-encoded proteins, although also being well conserved, have a lower score for this domain. There is no sulfate-binding CysP (SbpA) subunit in plastids which could indicate that *cysP* is located in the nucleus.

Outside of the Viridiplantae group, CysA orthologs are encoded in plastids of *Cyanidium caldarium* (NP_045139.1), *Cyanidioschyzon merolae* (NP_848950.1), and *Cyanophora paradoxa* (NP_043273.1). Interestingly, cyanobacteria CysA orthologs differ from plastid CysA orthologs. The NP043273.1 protein of *C. paradoxa* is an ortholog for the substrate-binding subunit of a zinc or manganese transporter.

3.2. Analysis of the Genomic Context. Genes upstream and downstream *cysT* and *cysA* are listed in Table 2. The *rpl32* gene located upstream *cysT* encodes the ribosomal protein L32 and in most cases belongs to the same DNA strand as *cyst*; refer to Table 2. *Pycnococcus provasolii*, *Bryopsis hypnoides*,

Table 2: The genomic context of the *cysA* and *cysT* genes in plastids of the Viridiplantae. The symbol "&" designated the lack of a bacterial-type promoter, "!" means that the intergenic region is very short, "*P*" designates the presence of a bacterial type promoter, "∗" designates a pseudogene, "()" marks an opposite direction, and "#" designates prediction of a conserved site.

Species	*cysA*	*cysT*
Nephroselmis olivacea	(*trnE*)-&-*cysA*-(*psbZ*)	*rpl32*-&-*cysT*-*ycf1*
Pycnococcus provasolii	None	*trnP*-&-*cysT*-*ycf1*
Bryopsis hypnoides	*ccsA*-&!-*cysA*-*psbB*	*rpl12*-&-*cysT*-(*rpoA*)
Chlorella variabilis	*accD*-#*P*-*cysA*-(*trnT1*)	*rpl32*-#*P*-*cysT*-*ycf1*
Chlorella vulgaris	*accD*-&-*cysA*-(*trnT*)	*rpl32*-##*P*-*cysT*-*orf819*
Coccomyxa subellipsoidea	*accD*-###*P*-*cysA*-(*trnN*)	*rpl32*-##*P*-*cysT*-*ycf1*
Helicosporidium sp.	None	*ftsH*-##*P*-*cysT*-*ycf1*
Leptosira terrestris	*orf96*-&-*cysA*-*rbcL*	*orf67*-&-*cysT*-(*trnC*)
Parachlorella kessleri	*accD*-#*P*-#*P*-*cysA*-(*trnT*)	*rpl32*-#*P*-#*P*-*cysT*-*ycf1*
Mesostigma viride	(*trnE*)-#*P*-*cysA*-(*trnT*)	*rpl32*-#*P*-#*P*-*cysT*-*ycf1*
Chlorokybus atmophyticus	(*trnR*)-&-*cysA*-(*trnT*)	(*rpl32*)-*P*-*cysT*-*ycf1*
Zygnema circumcarinatum	(*trnE*)-#*P*-*cysA*-*trnT*	*trnV*-&-*cysT*-(*rpl21*)
Anthoceros formosae	*trnE*-#*P*-#*P*-#*P*-*cysA*-*trnT*	*rpl32*-#*P*-*cysT*-(*trnP*)
Aneura mirabilis	(*trnE*)-*cysA*∗-*trnT*	*rpl32*-&-*cysT*∗-(*trnP*)
Marchantia polymorpha	(*trnE*)-*P*-*mbpX*-*trnT*	*rpl32*-*P*-*cysT*-*trnP*∗
Ptilidium pulcherrimum	(*trnE*)-*cysA*∗-*trnT*	*rpl32*-*cysT*∗-*trnL*

Helicosporidium sp., *Leptosira terrestris*, and *Zygnema circumcarinatum* have different gene configurations of this loci; refer to Table 2. Only in *Chlorokybus atmophyticus*, *rpl32* is both upstream of *cysT* and belongs to a different strand than *cysT*. In most cases, the intergenic region upstream the *cysT* gene is quite long. The *ycf1* gene is present downstream *cysT* in most algae. A few other genes are found downstream *cysT*: tRNA in Bryophyta and the alga *Leptosira terrestris*; *rpl21* (L21 protein) in the alga *Zygnema circumcarinatum*; *rpoA* (alpha subunit of bacterial-type RNA polymerase) in *Bryopsis hypnoides*; refer to Table 2.

The *accD* gene is located upstream *cysA* and belongs to the same DNA strand in Trebouxiophyceae algae, except for *Leptosira terrestris*. In *Nephroselmis olivacea*, however, a tRNA gene upstream *cysA* is on the complementary strand. In *Bryopsis hypnoides, ccsA* is upstream *cysA,* and the intergenic region is very short. In Streptophyta, *cysA* is surrounded by tRNA genes, and they often belong to the complementary strand which suggests the presence of a promoter directly in the upstream region of *cysA*.

3.3. Searching for Bacterial Type Promoters. Only two candidate bacterial-type promoters are found in 5′-leader regions of the considered genes; refer to Table 2. The exception is the *cysA* gene in *Anthoceros formosae*, for which we detect three potential promoters of similar quality. In *Chlorella vulgaris,* the single promoter candidate is located upstream *cysA* and has the unusual −35 box, AAGAAA, which was the reason for its rejection. However, in *Ch. variabilis,* a good potential promoter is detected in the upstream region of this gene with a TG-extension of the −10 box. Promoters were not found in the upstream regions of either *cysA* or *cysT* in *Nephroselmis olivacea*, *Pycnococcus provasolii*, *Bryopsis hypnoides*, *Leptosira terrestris*, *Aneura mirabilis,* and *Ptilidium pulcherrimum.* Promoters were not found in the upstream region of *cysA* in *Chlorokybus atmophyticus* and in the upstream region of *cysT* in *Zygnema circumcarinatum.* We speculate that in these cases these genes are transcribed as a part of an operon or by an RNA polymerase of the phage type.

3.4. Searching for the Conservative Motif. Transcription regulation of plastid genes involved in the sulfate transport in the parasites of invertebrate (*Helicosporidium* sp.) and in other species of the Viridiplantae is considered. A one-box conserved motif with the consensus TAAWATGATT is found near the promoters in the upstream regions of the *cysT* and *cysA* genes in many species. In some cases, the motif is repeated two or three times. In the upstream region of the *cysA* promoter in alga *C. subellipsoidea* C-169, however, the entire motif is repeated twice and is supplemented with its partial repeat at the 5′-terminus. The motif is not present near the promoters in *Chlorokybus atmophyticus* and *Marchantia polymorpha*. Deviations from the motif consensus are often the same in the same taxonomic lineage, which may reflect the variability of the transcription factor between lineages. The consensus was obtained from multiple alignments of 28 regions upstream two genes in 9 species (*Coccomyxa subellipsoidea*, *Chlorella variabilis*, *Chlorella vulgaris*, *Helicosporidium*, *Parachlorella kessleri*, *Mesostigma viride*, *Chlorokybus atmophyticus*, *Zygnema circumcarinatum*, and *Anthoceros formosae*). The LOGO profile of this motif is shown in Figure 3.

In most species, the motif is found upstream the −35 box or is overlapping the promoter. In the *cysA* upstream region in *Zygnema circumcarinatum* and *Anthoceros formosae*, the motif is detected between −35 and −10 boxes or is overlapping the −10 box of the promoter.

4. Discussion

We believe that the found motif represents binding sites of a transcription factor because of its positional linking with the promoter. The variable distance between the motif and the promoters suggests a repressor role of a putative transcription factor. Repeating of the motif is typically associated with a cooperative factor binding. This cooperativity can compensate for the motif variability, which is the case of *Coccomyxa subellipsoidea*.

The motif sequence confirms that *Helicosporidium* sp. belongs to the class Trebouxiophyceae. Its conservativity emphasizes the importance of *cysT* in plastids of parasites and suggests its key importance in understanding the role of the plastids in virulence. Indeed, plastids often synthesize many chemicals, which are usually provided by mitochondria [7, 25].

In *Leptosira terrestris, cysA* and *cysT* are not predicted to have the regulatory sites, unlike other Trebouxiophyceae and their close relatives, which suggests a shift in the transporter (consisting of CysA, CysT, and CysP subunits) specificity in *Leptosira*. This observation conforms with considerable changes in the CysA sequence in *Leptosira*. The lack of regulatory sites in *Nephroselmis olivacea*, *Pycnococcus provasolii*, *Bryopsis hypnoides*, and *Marchantia polymorpha* may suggest a demising role of the protein, which is consistent with the loss or pseudogene nature of *cysA* and *cysT* in *Aneura mirabilis* and *Ptilidium pulcherrimum*. The absence of bacterial-type promoters upstream *cysA* and *cysT* is often associated with changes in genes order on the chromosome. This effect may be explained by the *de novo* formation of phage-type promoters (possibly activated by another factor), or the inclusion of *cysA* or *cysT* in another operons. In general, the sulfate transport can be regulated by changing the expression level of a nuclear encoded sulfate-binding domain CysP (SbpA).

CysB binding sites in proteobacteria [9, 10] differ considerably from binding sites of a putative factor that we predicted for *cysT* and *cysA*. However, the two most conserved motif positions in plastids coincide with the two conserved positions in experimentally characterized sites upstream *cysPTWAM*, *cysK*, *cysJIH*, *cysDNC*, *sbp*, and *cysB* in proteobacteria. This evidence is however insufficient to establish the identity of our predicted motif and the CysB binding sites in proteobacteria. In *E. coli,* both proteins CysT and CysW consist of transmembrane domains that are very close to each other. Their genes belong to the sulfate transport operon *cysPTWAM*. But the CysW protein is absent in plastids. We hypothesize that CysW is replaced by the second CysT copy in plastids. The *cysT* and *cysA* genes do not form an operon in plastids, so we assume the CysT protein's double abundance over the CysA protein. It conforms to our hypothesis.

In *E. coli,* CysT and CysW are transmembrane domains with similar structure. Their genes belong to the same operon with the sulfate transport subunit *cysPTWAM*; however, *cysW* is absent in plastids. We hypothesize that in plastids the CysW subunit is functionally replaced by another copy of CysT, and the *cysT* mRNA concentration is twice as high compared to *cysA* mRNA. This hypothesis is indirectly supported by the fact that *cysT* and *cysA* are not included in one operon in plastids, and thus their mRNA expression levels may differ considerably.

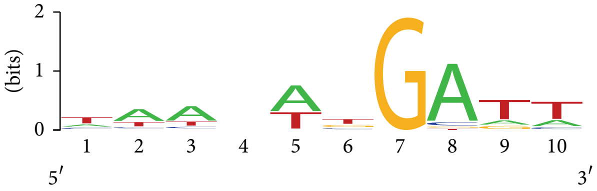


Figure 3: A LOGO nucleotide frequency profile for 10 positions of the found motif.

Acknowledgments

The authors are deeply grateful to L. Yu. Rusin, A. V. Lyubetskaya and S. A. Korolev for a valuable discussion of the results and methods, helping in their preparation. This work was partly supported by the Ministry of Education and Science of the Russian Federation (State Contract 14.740.11.1053 and Agreement 8481).

References

[1] A. Tartar, D. G. Boucias, B. J. Adams, and J. J. Becnel, "Phylogenetic analysis identifies the invertebrate pathogen *Helicosporidium* sp. as a green alga (Chlorophyta)," *International Journal of Systematic and Evolutionary Microbiology*, vol. 52, no. 1, pp. 273–279, 2002.

[2] A. P. de Koning and P. J. Keeling, "The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured," *BMC Biology*, vol. 4, article 12, 2006.

[3] J.-F. Pombert and P. J. Keeling, "The mitochondrial genome of the entomoparasitic green alga *Helicosporidium*," *PloS ONE*, vol. 5, no. 1, Article ID e8954, 2010.

[4] T. A. Sadovskaya and A. V. Seliverstov, "Analysis of the 5°-Leader regions of several plastid genes in protozoa of the phylum apicomplexa and red algae," *Molecular Biology*, vol. 43, no. 4, pp. 552–556, 2009.

[5] N. Tajima, S. Sato, F. Maruyama et al., "Genomic structure of the cyanobacterium *synechocystis* sp. PCC 6803 strain GT-S," *DNA Research*, vol. 18, no. 5, pp. 393–399, 2011.

[6] C. Thelwell, N. J. Robinson, and J. S. Turner-Cavet, "An SmtB-like repressor from *synechocystis* PCC 6803 regulates a zinc exporter," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 18, pp. 10728–10733, 1998.

[7] H. Takahashi and K. Saito, "Subcellular localization of spinach cysteine synthase isoforms and regulation of their gene expression by nitrogen and sulfur," *Plant Physiology*, vol. 112, no. 1, pp. 273–280, 1996.

[8] N. Thomsen-Zieger, J. Schachtner, and F. Seeber, "Apicomplexan parasites contain a single lipoic acid synthase located in the plastid," *FEBS Letters*, vol. 547, no. 1–3, pp. 80–86, 2003.

[9] N. M. Kredich, "The molecular basis for positive regulation of cys promoters in *Salmonella typhimurium* and *Escherichia coli*," *Molecular Microbiology*, vol. 6, no. 19, pp. 2747–2753, 1992.

[10] A. S. Lynch, R. Tyrrell, S. J. Smerdon, G. S. Briggs, and A. J. Wilkinson, "Characterization of the CysB protein of *Klebsiella aerogenes*: direct evidence that N-acetylserine rather than O-acetylserine serves as the inducer of the cysteine regulon," *Biochemical Journal*, vol. 299, no. 1, pp. 129–136, 1994.

[11] O. A. Zverkov, L. Rusin Yu, A. V. Seliverstov, and V. A. Lyubetsky, "Study of direct repeats in micro evolution of plant

mitochondria and plastids based on protein clustering," *Moscow University Biological Sciences Bulletin*, vol. 68, no. 2, pp. 58–62, 2013.

[12] O. A. Zverkov, A. V. Seliverstov, and V. A. Lyubetsky, "Plastid-encoded protein families specific for narrow taxonomic groups of algae and protozoa," *Molecular Biology*, vol. 46, no. 5, pp. 717–726, 2012.

[13] A. V. Seliverstov, E. A. Lysenko, and V. A. Lyubetsky, "Rapid evolution of promoters for the plastome gene *ndhF* in flowering plants," *Russian Journal of Plant Physiology*, vol. 56, no. 6, pp. 838–845, 2009.

[14] V. A. Lyubetsky, L. I. Rubanov, and A. V. Seliverstov, "Lack of conservation of bacterial type promoters in plastids of Streptophyta," *Biology Direct*, vol. 5, article 34, 2010.

[15] V. A. Lyubetsky, O. A. Zverkov, L. I. Rubanov, and A. V. Seliverstov, "Modeling RNA polymerase competition: the effect of σ-subunit knockout and heat shock on gene transcription level," *Biology Direct*, vol. 6, article 3, 2011.

[16] V. A. Lyubetsky, O. A. Zverkov, S. A. Pirogov, L. I. Rubanov, and A. V. Seliverstov, "Modeling RNA polymerase interaction in mitochondria of chordates," *Biology Direct*, vol. 7, article 26, 2012.

[17] A. Homann and G. Link, "DNA-binding and transcription characteristics of three cloned sigma factors from mustard (*Sinapis alba* L.) suggest overlapping and distinct roles in plastid gene expression," *European Journal of Biochemistry*, vol. 270, no. 6, pp. 1288–1300, 2003.

[18] V. A. Lyubetsky and A. V. Seliverstov, "Some algorithms related to finite groups," *Information Processes*, vol. 3, no. 1, pp. 39–46, 2003 (Russian).

[19] Z. Su, V. Olman, F. Mao, and Y. Xu, "Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis," *Nucleic Acids Research*, vol. 33, no. 16, pp. 5156–5171, 2005.

[20] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Research*, vol. 25, no. 24, pp. 4876–4882, 1997.

[21] N. J. Wickett, L. L. Forrest, J. M. Budke, B. Shaw, and B. Goffinet, "Frequent pseudogenization and loss of the plastid-encoded sulfate-transport gene cysA throughout the evolution of liverworts," *American Journal of Botany*, vol. 98, no. 8, pp. 1263–1275, 2011.

[22] V. A. Lyubetsky, A. V. Seliverstov, and O. A. Zverkov, "Elaboration of the homologous plastid-encoded protein families that separate paralogs in magnoliophytes," *Mathematical Biology and Bioinformatics*, vol. 8, no. 1, pp. 225–233, 2013 (Russian).

[23] E. V. Koonin, Y. I. Wolf, and L. Aravind, "Protein fold recognition using sequence profiles and its application in structural genomics," *Advances in Protein Chemistry*, vol. 54, pp. 245–275, 2000.

[24] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, Database Issue 40, pp. D290–D301, 2012.

[25] R. J. M. Wilson, K. Rangachari, J. W. Saldanha et al., "Parasite plastids: maintenance and functions," *Philosophical Transactions of the Royal Society B*, vol. 358, no. 1429, pp. 155–164, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 503193, 7 pages
http://dx.doi.org/10.1155/2013/503193

Research Article

Inferring Phylogenetic Networks from Gene Order Data

Alexey Anatolievich Morozov, Yuri Pavlovich Galachyants, and Yelena Valentinovna Likhoshway

Limnological Institute of the Siberian Branch of the Russian Academy of Sciences, 3 Ulan-Batorskaya Street, Irkutsk 664033, Russia

Correspondence should be addressed to Yuri Pavlovich Galachyants; yuri.galachyants@lin.irk.ru

Received 19 March 2013; Revised 4 July 2013; Accepted 20 July 2013

Academic Editor: Vassily Lyubetsky

Existing algorithms allow us to infer phylogenetic networks from sequences (DNA, protein or binary), sets of trees, and distance matrices, but there are no methods to build them using the gene order data as an input. Here we describe several methods to build split networks from the gene order data, perform simulation studies, and use our methods for analyzing and interpreting different real gene order datasets. All proposed methods are based on intermediate data, which can be generated from genome structures under study and used as an input for network construction algorithms. Three intermediates are used: set of jackknife trees, distance matrix, and binary encoding. According to simulations and case studies, the best intermediates are jackknife trees and distance matrix (when used with Neighbor-Net algorithm). Binary encoding can also be useful, but only when the methods mentioned above cannot be used.

1. Introduction

Gene order data gain increasing popularity in the phylogenetic community because of several advantages they have, compared with gene sequences. First, in most cases the genome structure evolves slower than DNA or protein sequence, allowing the inference about ancient events with less noise level [1]. Second, like all phylogenomic studies, the analysis of genomic rearrangements is not hampered by conflicts between gene trees and species tree. One can expect the rearrangements-based inference and phylogenomics in general to become more and more widespread as DNA sequencing cost continues to decline and new computational tools are developed to deal with this kind of data.

As with any analysis, it is based upon several assumptions. First, every genome is represented as a permutation of homologous markers, which are usually directed. All genomes in the dataset should contain the same set of markers. Though in the majority of studies these markers are genes (hereinafter referred to as “genes”), they can be contiguous parts of a chromosome of any reasonable length. Such a permutation is traditionally represented as a sequence of signed numbers with absolute values being identifiers of elements and a sign denoting direction. Second, a set of operations on a permutation is limited to some subset of all actually possible evolutionary events. Inversions are the most common, but there can be also translocations, double-cut-and-join operations, and several other events [1].

In the majority of phylogenetic studies, the evolutionary relationships of taxa are represented by phylogenetic trees. Despite their usefulness for biology, a phylogenetic tree by definition is able to display only one divergence-based scenario. Real evolution, on the other hand, is not limited to diverging taxa: recombination and horizontal gene transfer occur in all major groups of organisms. Even in cases when a tree-like evolution is safe to assume, sometimes it is useful to simultaneously visualize all conflicting scenarios supported by the dataset [2]. Reticulate evolutionary events are currently beyond the scope of the gene order analysis because a similar set of genes in all genomes is assumed. On the other hand, ambiguous data do not often allow the choice of a single tree. To solve these problems, *phylogenetic networks*, that is, nontree graphs describing evolutionary relationships, were proposed. Many types of phylogenetic networks exist [2], but in this study we are interested only in *split networks* [3].

An idea of the split network is based upon one crucial observation: every branch of the phylogenetic tree defines a split or bifurcation of taxa set, that is, separates it into two parts. Split is called trivial if one of its parts contains only one taxon. Such a split corresponds to a leaf branch on the phylogenetic tree. A set of splits is called compatible if it can be represented by an unrooted phylogenetic tree. Phylogenetic inference in this framework is reduced to generating a set of nontrivial splits and representing them as a graph.

A split network is a generalization of mathematical concept of a phylogenetic tree. It is able to represent both compatible and incompatible split sets. Unlike unrooted tree, a split network may use several parallel edges to represent any given split. Deletion of all edges corresponding to a split divides the network in exactly two connected components, one containing all taxa from one part of the split, and another containing taxa from the other. The edge lengths are defined by split weight, which can be of different sense depending on the algorithm used to generate a split set [2].

2. Materials and Methods

2.1. Building Phylogenetic Networks. The main idea of this work is to apply existing network-building algorithms to intermediate data, which were generated according to the structure of studied genomes. We used a set of phylogenetic trees (consensus network algorithm), inversion distance matrix (split decomposition and Neighbor-Net algorithms), and binary encoding (parsimony-splits algorithm) as intermediate data. In all cases, we considered rearrangements in unichromosomal genomes consisting of directed markers and limited the evolutionary process to inversions. The latter assumption is not applied to binary encoding (see details below).

We did not filter the splits with any of the algorithms and intermediates. The aim of the additional filtering is to obtain a relatively simple set of splits (e.g., cyclic or weakly compatible). Such a set of splits corresponds to a network with simple topology, which can be easily drawn on a flat surface. However, such a simplification causes removal of some splits from the resulting network, which potentially leads to loss of important data. We decided to sacrifice simplicity of the network for the sake of its accuracy.

The data transformations, jackknife analysis, and simulations were done by custom Perl scripts, which are available upon request. All network construction algorithms are implemented in SplitsTree4 software by Huson and Bryant [4]. Computations were done at Irkutsk High-Performance Computer Center (ISDCT SB RAS, Irkutsk).

2.2. Jackknife Trees. Since the function of split networks is to represent conflicting trees, the most obvious solution is to generate a bunch of trees and sum them up into a network. To obtain a set of trees, we conducted a jackknife procedure: 40% of genes were chosen randomly and removed from all permutations, preserving the order of the remaining ones. We generated 100 replicates and built phylogenetic trees using COGNAC package by Kang et al. [5].

Phylogenetic network was built by consensus network algorithm [6]. This algorithm uses all splits which are present in all input trees to build a network. The algorithm was set up to include a split into the network if it is present in at least 10% of the input trees. In this case, the split weights and therefore the edge lengths in the resulting network are equal to a jackknife support of the corresponding bifurcation, that is, a proportion of input trees that include this bifurcation.

2.3. Distance Matrix. Distance matrix is a common kind of intermediate data which allows building phylogenetic trees and networks from different raw data using the same algorithms. GRAPPA package [7] was used to generate the matrix of pair-wise inversion distances. This distance metric designates minimum number of inversion operations necessary to transform an initial gene permutation into a target one. Two algorithms were applied to distance matrices: Neighbor-Net [8] and split decomposition [2].

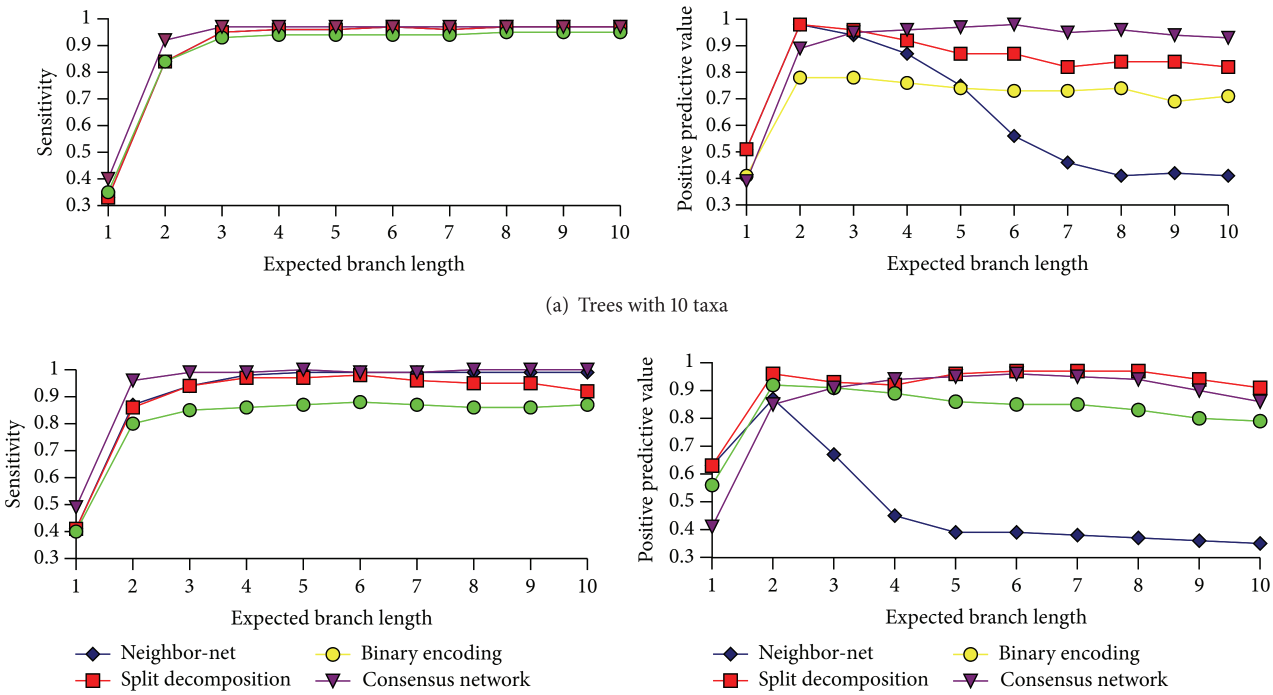
2.4. Binary Encoding. Binary encoding (BE) represents a set of permutations as a matrix of binary characters [9]. The rows of this matrix correspond to permutations. Columns are the pairs of genes in all four relative orientations (+N +M, +M +N, +N −M, and −N +M). Every element of matrix equals 1 if two genes are adjacent in genome in given directions; otherwise, it is equal to 0. Such an approach allows us to promptly analyze large datasets. It can use existing software and does not make explicit assumption about the nature of rearrangement process. BE matrices were processed by parsimony-splits algorithm [10].

The median network algorithm is also popular in the analysis of binary sequences. However, a crucial step of the analysis is to reconstruct ancestral states. In case of BE, the algorithm does not account for actual evolution of underlying gene order, therefore the reconstruction of ancestral sequences will be incorrect. Thus, we did not use this approach in our work.

2.5. Simulations. For simulations, we generated Yule trees for 10 and 20 taxa with the branch lengths sampled from Poisson distribution. The expected branch lengths λ ranged from 1 to 10 inversions. For each combination of the taxa amount and λ, we generated 100 random trees. The gene orders evolved according to topology of these trees by random inversions starting from the identity permutation of 100 genes at the root of the tree. Permutations observed at the tree leaves were used as an input for methods described above.

The resulting networks were compared with the true underlying tree. We assessed only a network topology, that is, a set of splits. We counted nontrivial splits that are either present in the true tree, but not in the network (false negative (FN)), or *vice versa* (false positive (FP)). Obviously, all trivial splits are always present in both tree and network, therefore they were excluded from the analysis.

Two values were obtained from the FN and FP counts to compare performance of the methods. First value is sensitivity, that is, probability for any split in the true tree to be included into the phylogenetic network. Another value is a positive predictive value (PPV), which has an opposite meaning: probability that a network split belongs to the true tree. We calculated neither specificity nor negative predictive



(a) Trees with 10 taxa

(b) Trees with 20 taxa

Figure 1: Simulation results.

value because the true negative count, which is the amount of all possible splits on given taxa set minus the amount of splits in the true tree, is always several orders of magnitude larger than that of FN and FP.

Split weights were not taken into account because their meaning is different in the network-building approaches used. This is an important flaw of the simulation procedure because once a split is present in the network, it will increase sensitivity (or decrease PPV) of the method, even if the weight of this split is vanishingly small.

2.6. Case Studies. We used two real datasets in the case studies. The first is Campanulaceae dataset, which is commonly used to evaluate performance of the rearrangement-based phylogenetic inference software [9, 11, 12], provided as test data with the Badger package [12]. This dataset is poorly resolved: all trees in the aforementioned papers contain multifurcations and, in some cases, support values are not provided (see Figure 3(e) for example).

Another dataset consists of six chloroplast genomes of diatoms and two genomes of diatom-derived chloroplasts of dinoflagellates. All these are circular genomes 120–130 kbp long, containing from 154 to 159 genes, including tRNAs. Each chloroplast DNA bears a long inverted repeat that contains genes for rRNAs and several proteins. Phylogenetic relationships of taxa are quite well established [13], and our tree inferred from the order of genes in chloroplast genomes supports the conventional scenario (Figure 2(e)).

We removed one of the copies of inverted repeat from all diatom genomes, thus transforming sequences from circular to linear. Then all common genes were assigned numbers and marked with signs depending on their orientation. The resulting dataset consisted of eight permutations of 149 genes each. Both real datasets were analyzed in the same way as the simulated data.

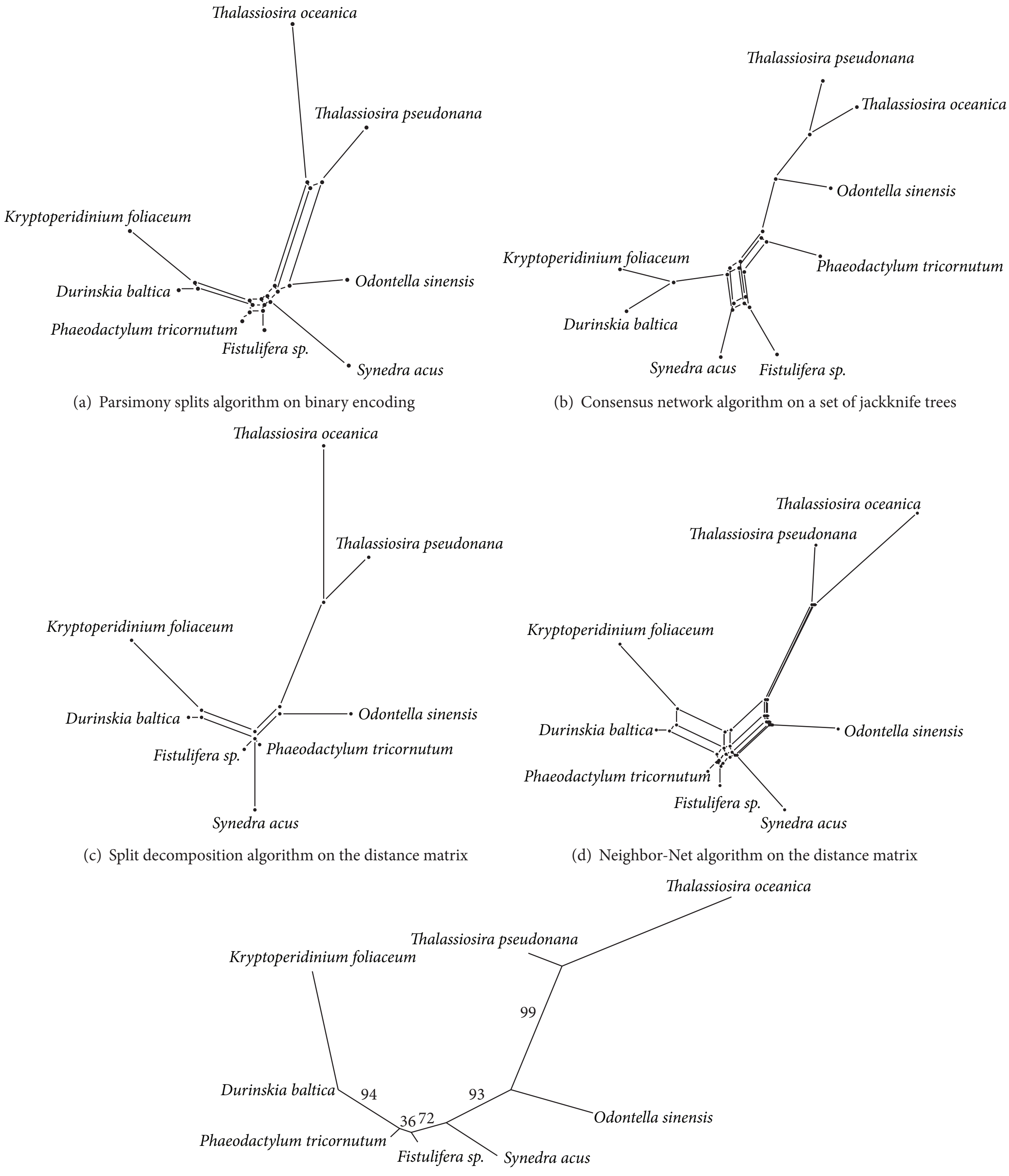
3. Results and Discussion

3.1. Simulations. According to simulation studies (Figure 1), the analysis of a set of the jackknife trees with the consensus network algorithm appeared to be the best way to build split networks. Networks generated by this method had the highest sensitivity and PPV in most tests. However, the reconstruction of a set of trees is significantly more CPU-intensive than computation of either distance matrix or binary encoding.

The split decomposition algorithm is slightly outperformed by the consensus network approach in terms of both sensitivity and PPV. The distance matrix can be computed significantly faster than a set of trees. It took less than a minute on a desktop computer for all tests performed. Additionally, this method guarantees to produce a weakly compatible set of splits, ensuring less complicated network.

PPV of Neighbor-Net algorithm significantly decreases with increasing lengths of tree branches. On the other hand, sensitivity is similar to that of other methods.

The parsimony-splits algorithm applied to the binary encoding is different from other methods. It does not analyze permutations, but processes binary matrices built from them. This may be useful, if the evolutionary process is not assumed to be based mainly on inversions. However, in our



(a) Parsimony splits algorithm on binary encoding

(b) Consensus network algorithm on a set of jackknife trees

(c) Split decomposition algorithm on the distance matrix

(d) Neighbor-Net algorithm on the distance matrix

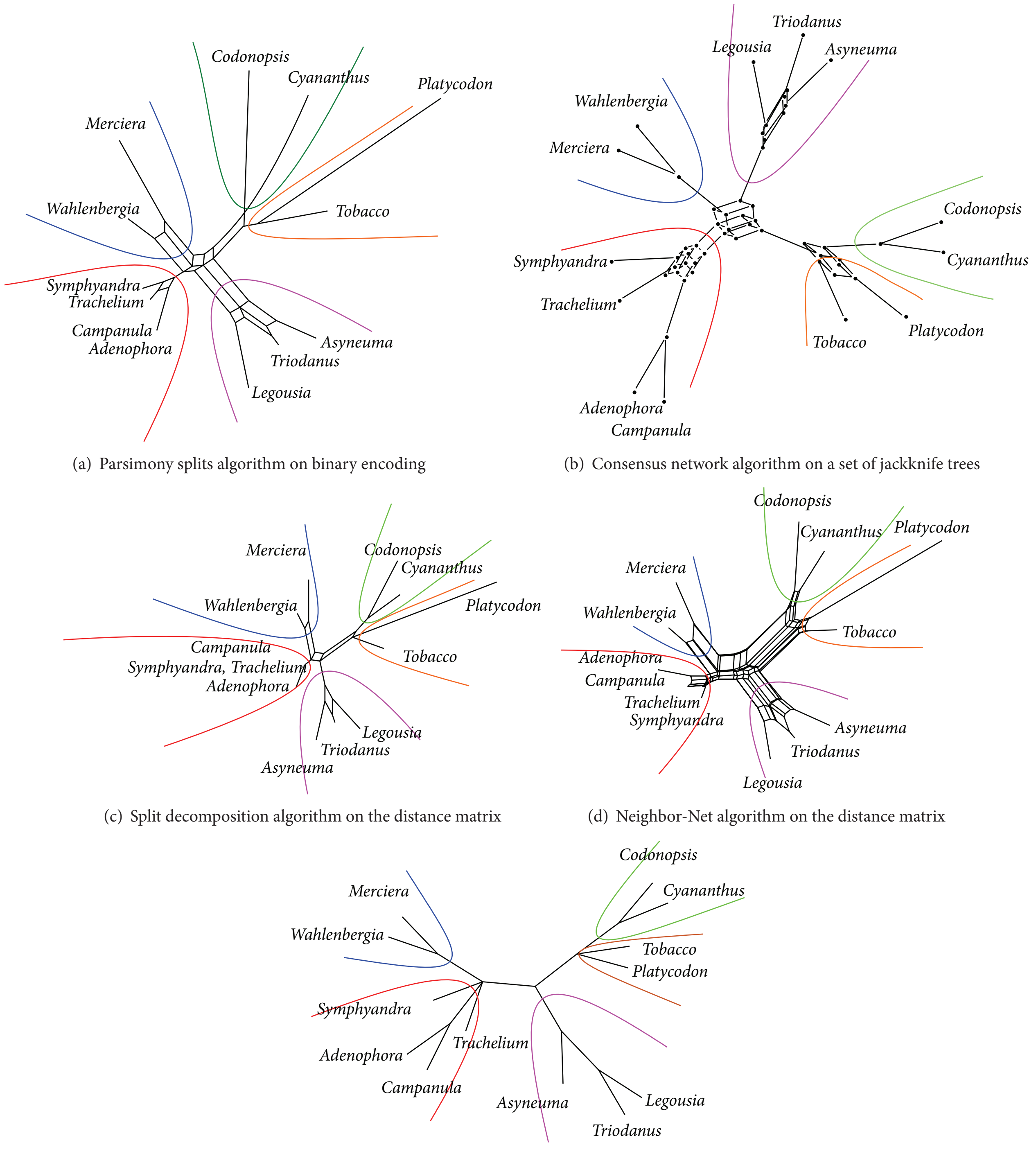
(e) Phylogenetic tree built by MGR package

Figure 2: Split networks and reference tree for the Bacillariophyta dataset.

"inversions-only" simulations, the networks built with this method have the lowest sensitivity and the second lowest PPV.

3.2. Case Studies. Obviously, one can never know exactly which tree is actually true for a real dataset. Therefore, by analyzing real data: one can only assess relative complexity of a network, that is, how many splits are included, and whether it is congruent with the trees obtained on the same dataset with other methods.

On Bacillariophyta dataset, we first built a phylogenetic tree using MGR package (Figure 2(e)) [11]. This tree is



(a) Parsimony splits algorithm on binary encoding

(b) Consensus network algorithm on a set of jackknife trees

(c) Split decomposition algorithm on the distance matrix

(d) Neighbor-Net algorithm on the distance matrix

(e) Tree built by BADGER package (redrawn from [12])

Figure 3: Split networks and reference tree for the Campanulaceae dataset.

congruent to the trees obtained using molecular phylogenetic analysis of several diatom genes [13]. Therefore, we used topology of this tree as a reference to assess quality of split networks.

The Neighbor-Net algorithm has produced the most complex network (Figure 2(d)) with the largest number of splits. Most of these splits, however, have low weights, making the best scenario clearly visible. Binary encoding-based network (Figure 2(a)) is smaller in terms of splits. Its topology is also the closest to the MGR tree. Two other networks contain significant flaws. Consensus network gives the highest support to positions of *Synedra*, *Phaeodactylum*, and *Fistulifera* that contradict our MGR tree (Figure 2(b)). It is also the only network that contains a 3-dimensional structure, which makes it much harder to read. Network built by split decomposition algorithm has a small number of additional splits, but it also lacks a few crucial ones, leaving the *Synedra*/*Phaeodactylum*/*Fistulifera* relationships completely unresolved (Figure 2(c)).

Campanulaceae dataset was confirmed to be ambiguous. All methods support five relatively well-supported monophyletic clusters (marked by colors in Figure 3), but neither positions of taxa inside them nor relations of the clusters are resolved. The same conflict is present in consensus trees built with other algorithms ([10, 12, 13], see Figure 3(e)). Networks clearly show that some groups can be reliably separated from each other, yet the complete reconstruction cannot be done based solely on this dataset.

4. Conclusions

In this study, we propose several methods to build split networks using the gene order data via generating the intermediate datasets. We used a set of jackknife trees, an inversion distance matrix, and a binary encoding of the gene order as intermediate data. The performance of these methods is shown to vary depending on input data. Furthermore, the suggested methods are different in assumptions and mathematical approaches behind them. Below we summarize *pro et contra* of every method.

A set of jackknife trees is useful in most cases. In simulations, it performs well in terms of both sensitivity and positive predictive value. Moreover, it produces networks with bifurcation support as a split weight, which is very useful when comparing the reliability of different scenarios. Since, in the absence of additional split filtering, consensus network approach does not limit the produced split set to weakly compatible or circular, it can create networks of very complex, hard to read topology.

Networks are computationally cheaper to build with the distance matrix as an intermediate dataset. This matrix can be analyzed by Neighbor-Net and split decomposition algorithms. When comparing these two algorithms, it is necessary to take into account that PPV is much less important than sensitivity. If the data clearly support only one scenario, it would not be obscured by addition of several low-weight splits represented by barely visible edges. On the other hand, if several contradictory trees are supported, a resulting network must include splits from all of them. In this case, PPV will decrease. However, the use of a network instead of a tree is aimed at representing this contradiction.

Unlike split decomposition algorithm which generates multifurcations, Neighbor-Net tends to add a lot of low-weight splits into network. Moreover, it has slightly higher sensitivity. These two features seem to be advantageous to apply the Neighbor-Net approach. However, the split network generated via Neighbor-Net, which is always producing a circular set of splits, may lack some splits versus the network derived with split decomposition algorithm. For detailed example, see archaeal chaperonins dataset [8]. This problem only appears for very contradictory scenarios, so in the majority of cases Neighbor-Net is preferable.

Analysis of binary encoded genome structures by parsimony-splits algorithm has lower sensitivity and PPV than the rest of methods. Still, it can be useful for very large datasets, when other approaches are computationally expensive. The fact that evolutionary process is not assumed to consist of some limited set of operations is also advantageous when no such set can be proposed. However, in most cases, the consensus network or Neighbor-Net approach would be more reliable.

Acknowledgment

This work was financially supported by the Program of the Presidium of the Russian Academy of Sciences "Molecular and Cell Biology" (project no. 6.9).

References

[1] G. Fertin, A. Labarre, I. Rusu, E. Tannier, and S. Vialette, *Combinatorics of Genome Rearrangements*, The MIT Press, Boston, Mass, USA, 2009.

[2] D. D. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge University Press, Cambridge, UK, 2011.

[3] H. J. Bandelt and A. W. M. Dress, "A canonical decomposition theory for metrics on a finite set," *Advances in Mathematics*, vol. 92, no. 1, pp. 47–105, 1992.

[4] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, 2006.

[5] S. Kang, J. Tang, S. W. Schaeffer, and D. A. Bader, "Rec-DCM-Eigen: reconstructing a less parsimonious but more accurate tree in shorter time," *PLoS ONE*, vol. 6, no. 8, Article ID e22483, 2011.

[6] B. R. Holland, K. T. Huber, V. Moulton, and P. J. Lockhart, "Using consensus networks to visualize contradictory evidence for species phylogeny," *Molecular Biology and Evolution*, vol. 21, no. 7, pp. 1459–1461, 2004.

[7] J. Tang and B. M. E. Moret, "Scaling up accurate phylogenetic reconstruction from gene-order data," *Bioinformatics*, vol. 19, pp. 305–312, 2003.

[8] D. Bryant and V. Moulton, "Neighbor-net: an agglomerative method for the construction of phylogenetic networks," *Molecular Biology and Evolution*, vol. 21, no. 2, pp. 255–265, 2004.

[9] M. E. Cosner, R. K. Jansen, B. M. E. Moret et al., "An empirical comparison of phylogenetic methods on chloroplast gene order data in campanulaceae," in *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, D. Sankoff and J. H. Nadeau, Eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[10] H. J. Bandelt and A. Dress, "A relational approach to split decomposition," in *Information and Classification*, O. Opitz, B. Lausen, and R. Klar, Eds., pp. 123–131, Springer, New York, NY, USA, 1993.

[11] G. Bourque and P. A. Pevzner, "Genome-scale evolution: reconstructing gene orders in the ancestral species," *Genome Research*, vol. 12, no. 1, pp. 26–36, 2002.

[12] B. Larget, J. B. Kadane, and D. L. Simon, "A Bayesian approach to the estimation of ancestral genome arrangements," *Molecular Phylogenetics and Evolution*, vol. 36, no. 2, pp. 214–223, 2005.

[13] E. C. Theriot, M. Ashworth, E. Ruck, T. Nakov, and R. K. Jansen, "A preliminary multi gene phylogeny of the diatoms (Bacillariophyta): challenges for future research," *Plant Ecology and Evolution*, vol. 143, no. 3, pp. 278–296, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 989410, 5 pages
http://dx.doi.org/10.1155/2013/989410

Research Article

Comparative Analysis of Context-Dependent Mutagenesis Using Human and Mouse Models

Sofya A. Medvedeva,[1] Alexander Y. Panchin,[2] Andrey V. Alexeevski,[1,3,4]
Sergey A. Spirin,[1,3,4] and Yuri V. Panchin[2,3]

[1] *Department of Bioengineering and Bioinformatics, Moscow State University, Vorbyevy Gory 1-73, Moscow 119992, Russia*
[2] *Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny Pereulok 19-1, Moscow 127994, Russia*
[3] *Department of Mathematical Methods in Biology, Belozersky Institute, Moscow State University, Vorbyevy Gory 1-40, Moscow 119991, Russia*
[4] *Department of Mathematics, Scientific Research Institute for System Studies, Russian Academy of Sciences, Nakhimovskii Prospekt 36-1, Moscow 117218, Russia*

Correspondence should be addressed to Alexander Y. Panchin; alexpanchin@yahoo.com

Received 18 April 2013; Accepted 19 July 2013

Academic Editor: Vassily Lyubetsky

Substitution rates strongly depend on their nucleotide context. One of the most studied examples is the excess of C > T mutations in the CG context in various groups of organisms, including vertebrates. Studies on the molecular mechanisms underlying this mutation regularity have provided insights into evolution, mutagenesis, and cancer development. Recently several other hypermutable motifs were identified in the human genome. There is an increased frequency of T > C mutations in the second position of the words ATTG and ATAG and an increased frequency of A > C mutations in the first position of the word ACAA. For a better understanding of evolution, it is of interest whether these mutation regularities are human specific or present in other vertebrates, as their presence might affect the validity of currently used substitution models and molecular clocks. A comprehensive analysis of mutagenesis in 4 bp mutation contexts requires a vast amount of mutation data. Such data may be derived from the comparisons of individual genomes or from single nucleotide polymorphism (SNP) databases. Using this approach, we performed a systematical comparison of mutation regularities within 2–4 bp contexts in *Mus musculus* and *Homo sapiens* and uncovered that even closely related organisms may have notable differences in context-dependent mutation regularities.

1. Introduction

Estimates of the average point mutation rates in eukaryotic genomes usually vary between 10^{-7} and 10^{-10} mutations per nucleotide per generation [1, 2]. However, mutation rates may be dramatically altered by their genomic context. For example, there is an increased frequency of C > T mutations in the word CG in humans (and other vertebrates). This is currently attributed to the methylation of cytosines by context specific DNA methyltransferases [3]. Many other examples of context-related factors that affect mutation rates have been reported and reviewed [4–8]. Substitution rates are known to be affected by local G + C content [9], CpG density [10], recombination rates [11], proximity to small insertions or deletions [12], distance from the centromeres or telomeres [13], and the chromosome itself (e.g., the human Y chromosome has higher divergence rates than autosomes) [14]. Some of these factors might be related to each other. The study of context-dependent changes in mutation frequencies may shed light on the molecular mechanisms involved in mutagenesis [15]. Also, it is important to understand how context affects mutation rates when working in the field of molecular phylogenetics. For example, accounting for the hypermutability of certain motifs may improve the accuracy of our estimates of the divergence time between two homologous sequences [16].

Recently, it was reported that there is an increased rate of T > C mutations in the second position of the words ATTG and ATAG and an increased rate of A > C mutations in the first position of the word ACAA in the *Homo sapiens* genome [17]. This result was achieved by calculating the values called "minimal contrast" and "mutation bias" for 2–4 bp mutation contexts to evaluate if the addition of specific nucleotides to the $5'$ or $3'$ end of 1–3 bp words increases the probability of observing certain mutations in fixed positions. Mutation bias indicates the total excess (or deficiency) of mutations within a given mutation context. Minimal contrast indicates the excess (or deficiency) of mutations within a given context that cannot be explained by the excess (or deficiency) of mutations in one of its subcontexts.

The analysis of mutation rates for 4 bp contexts analysis requires large amounts of mutation data (millions of inferred mutations) to provide statistically significant and biologically meaningful results. Sufficient SNP data for the analysis of context-dependent mutagenesis in *H. sapiens* was available for a long time. More recently multiple whole genome sequences of *Mus musculus* were presented [18, 19]. The comparison of these genomes provides essential data on genetic divergence and context-dependent variance between mouse genetic sequences similar to that provided by human SNP analysis. We used a systematical comparison of mutation regularities within 2–4 bp contexts in *M. musculus* and *H. sapiens*, evaluated by calculating mutation bias and minimal contrasts for the contexts and uncovered a number of notable differences in context-dependent mutation regularities. Namely, we found that the aforementioned hypermutable human mutation contexts except for the excess of C > T mutations in the CG context are not hypermutable in *M. musculus*. Also, several mutation contexts are hypermutable in *M. musculus* but not in *H. sapiens*.

2. Methods

2.1. Mutation Data. We used SNP data from 17 strains of mice, available from [18] http://www.sanger.ac.uk/resources/mouse/genomes/. To reduce the possible effects of selection on protein-coding genes, we excluded SNPs present within 1000 bp of known mouse genes (UCSC genes, as in UCSC genome browser [20]). SNPs with low-coverage sequencing, near simple repeats or indels, were excluded, according to [21].

We reconstructed the ancestral states of SNPs by using the genome of SPRET/EiJ mouse as an outgroup. This is justified because this strain is the most divergent from the rest [21]. We determined the direction of mutations that happened in the remaining 16 mouse strains by comparing the observed alleles with the corresponding outgroup sequence. Only those cases were considered, when two genetic variants were present in the 16 mouse strains and one of them was present in the SPRET/EiJ strain. Further analysis was done as in [17]. A total of 12.8 million mouse SNPS were included in the analysis.

2.2. Mutation Context and Subcontext. We denote the mutation context of mutation *mut* in position *pos* of the word W as $\{mut \mid pos, W\}$. For example, $\{C > T \mid 1, CG\}$ represents a C > T mutation in the first position of the word CG. Mutation context $\{mut \mid pos', W'\}$ is called a subcontext of the context $\{mut \mid pos, W\}$ if W' is a subword of W and any mutation *mut* occurring in position *pos* of the word W is at the same time a mutation occurring in position pos' of the word W'. For example, $\{C > T \mid 1, CG\}$ is a subcontext of $\{C > T \mid 2, ACG\}$. We do not study discontiguous contexts.

Table 1: The fractions of basic types of directed mutations, inferred from SNP data.

Mutation	Fraction	
	H. sapiens	*M. musculus*
A > T	0.031	0.034
T > A	0.031	0.034
A > C	0.037	0.029
T > G	0.038	0.029
C > G	0.051	0.035
G > C	0.051	0.035
G > T	0.058	0.059
C > A	0.058	0.059
T > C	0.118	0.097
A > G	0.118	0.097
C > T	0.204	0.247
G > A	0.204	0.247
Transversions	0.355	0.312
Transitions	0.645	0.688

2.3. Contrast. For each pair of context $\{mut \mid pos, W\}$ and its subcontext $\{mut \mid pos', W'\}$, the value of contrast is given by the formula

$$\text{Contrast}\left(\{mut \mid pos, W\}, \{mut \mid pos', W'\}\right) = \frac{P\{mut \mid pos, W\}}{P\{mut \mid pos', W'\}}. \tag{1}$$

Here, $P\{mut \mid pos, W\}$ and $P\{mut \mid pos', W'\}$ are the conditional probabilities of observing mutation *mut* in the position *pos* of the word W and in the position pos' of word W', respectively. Although these probabilities cannot be explicitly calculated without assumptions of the general probability of mutation per nucleotide in the genome, their ratio can be estimated by the following formula:

$$\frac{P\{mut \mid pos, W\}}{P\{mut \mid pos', W'\}} = \frac{N\{mut \mid pos, W\}/P_W}{N\{mut \mid pos', W'\}/P_{W'}}. \tag{2}$$

Here, P_W and $P_{W'}$ are the observed frequencies of words W and W', respectively, among all words of the same length.

The ratio $P_W/P_{W'}$ estimates the probability for W' to be extended to W. This ratio coincides with the expected ratio $N\{mut \mid pos, W\}/N\{mut \mid pos', W'\}$ under the hypothesis that mutations rates are the same in the context $\{mut \mid pos, W\}$ and its subcontext $\{mut \mid pos', W'\}$. Therefore, if Contrast $(\{mut \mid pos, W\}, \{mut \mid pos', W'\})$ is greater than 1,

Table 2: Top 5 40 bp mutation contexts by minimal contrast in *H. sapiens* and *M. musculus*. The provided subcontext is the context with the most similar to the contexts mutation bias value and is the one used for the minimal contrast calculation. Also reverse contexts are provided (contexts with the reverse mutation) with their minimal contrast and mutation bias values.

Context $\{mut \mid pos, W\}$	Minimal contrast	Mutation bias	Subcontext $\{mut \mid pos', W'\}$	Reverse context	Minimal contrast	Mutation bias
			H. sapiens			
{T > C|2, ATTG}	2.12	3.46	{T > C|1, TTG}	{C > T|2, ACTG}	0.86	0.75
{A > C|1, ACAA}	1.89	3.43	{A > C|1, ACA}	{C > A|1, CCAA}	1.01	1.20
{T > C|2, ATAG}	1.78	3.29	{T > C|2, ATA}	{C > T|2, ACAG}	0.95	0.81
{G > C|3, TCGA}	1.43	1.98	{G > C|3, TCG}	{C > G|3, TCCA}	0.59	0.37
{T > G|4, CGGT}	1.42	2.64	{T > G|3, GGT}	{G > T|4, CGGG}	0.84	0.84
			M. musculus			
{G > T|1, GCGA}	1.83	3.00	{G > T|1, GCG}	{T > G|1, TCGA}	1.19	1.19
{T > A|3, TTTA}	1.60	2.31	{T > A|2, TTA}	{A > T|3, TTAA}	1.47	2.55
{T > C|2, ATTG}	1.59	2.25	{T > C|2, AT}	{C > T|2, ACTG}	1.01	1.01
{G > A|4, CGCG}	1.54	1.54	{G > A|1, G}	{A > G|4, CGCA}	0.68	0.68
{T > A|2, TTAA}	1.47	2.55	{T > A|1, TAA}	{A > T|2, TAAA}	1.60	2.31

it indicates an increased mutation rate in the context $\{mut \mid pos, W\}$ compared with the subcontext $\{mut \mid pos', W'\}$. Analogously, if Contrast($\{mut \mid pos, W\}$, $\{mut \mid pos', W'\}$) is less than 1, it indicates a decreased mutation rate.

2.4. Minimal Contrast. For a given context $\{mut \mid pos, W\}$, let us consider all of its subcontexts $\{mut \mid pos', W'\}$. The minimal contrast is the value MC = Contrast($\{mut \mid pos, W\}$, $\{mut \mid pos', W'\}$) such that the absolute difference $|MC - 1|$ is the lowest among all subcontexts $\{mut \mid pos', W'\}$.

2.5. Mutation Bias. For any context $\{mut \mid pos, W\}$, there exists only one subcontext $\{mut \mid pos', W'\}$ such that the length of W' is equal to 1 (i.e., W' is the one-letter word, consisting of the mutated letter). The mutation bias is the contrast of the given context and this subcontext.

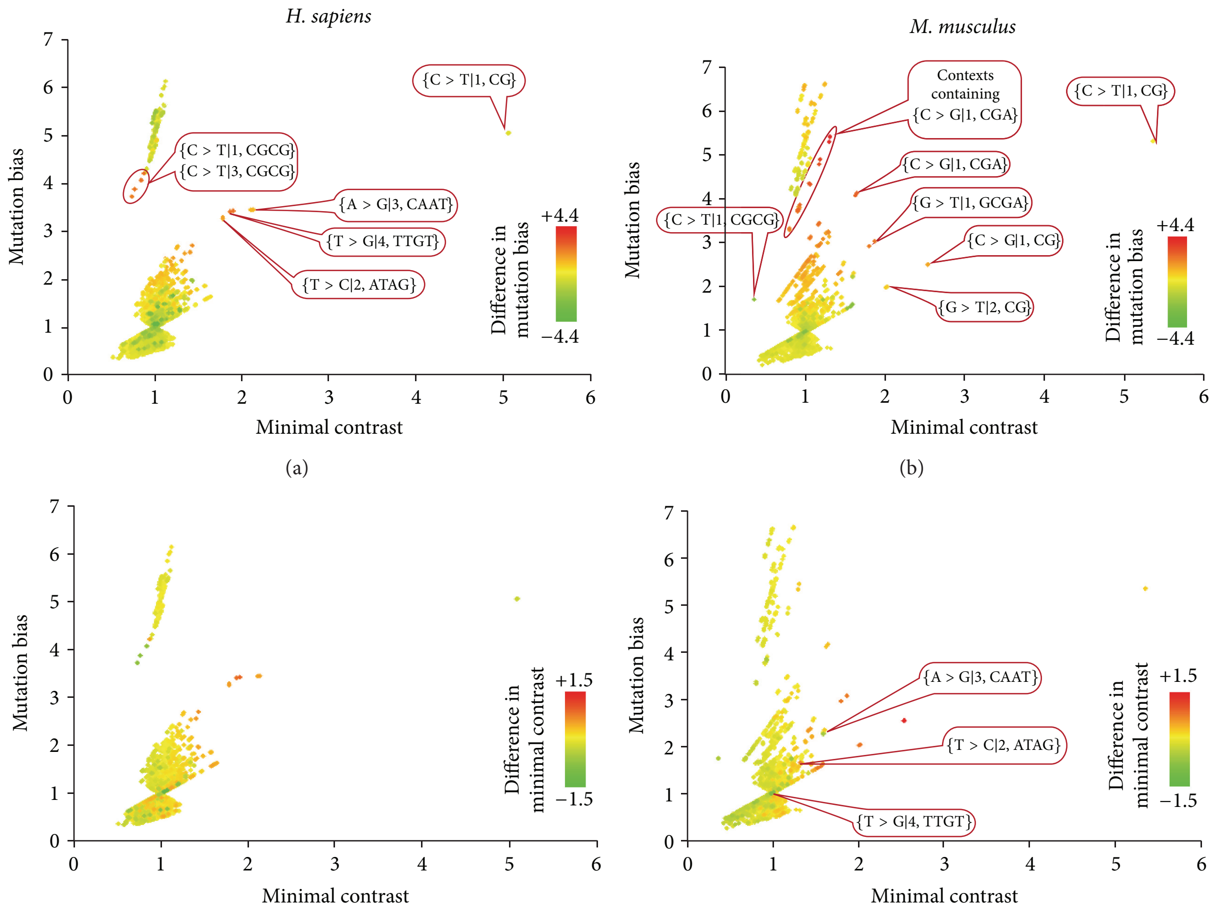
2.6. Word Frequencies. We estimated word frequencies (the fraction of a specific word in all amount of the words of the same length) in the mouse genome using $[-10, -5]$ and $[+5, +10]$ intervals surrounding the mouse SNPs included in our study. We used the reference mouse genome sequence for this purpose. These word frequencies were used in our calculations of mutation bias and minimal contrast for mutation contexts in *M. musculus*.

2.7. Statistical Significance. For a given pair of context and subcontext, let $P = P_W / P_{W'}$ be the expected probability of success in a Bernoulli trial, with the number of trials $N = N\{mut \mid pos', W'\}$ and the number of successes $K = N\{mut \mid pos, W\}$. We assume that the mutation rate for context $\{mut \mid pos, W\}$ is significantly different from the mutation rate of its subcontext $\{mut \mid pos', W'\}$ if the probability to observe K or a more extreme number of successes out of N trials with the probability of success P is lower than a predetermined significance level. Due to large sample sizes, all obtained P values for context/subcontext comparisons are highly significant ($P < 10^{-15}$) for all observations mentioned in our study. This remains true after correcting for multiple comparisons using the Bonferroni correction. For example, there are 1293 observed mutations for the *M. musculus* context $\{G > C \mid 3, TCGA\}$ and 3723 mutations for its closest (with the most similar mutation bias value) subcontext $\{G > C \mid 3, TCG\}$. $P_W / P_{W'}$ for this pair is 0.081. The P value is much less than 10^{-15}.

3. Results and Discussion

As shown in Table 1, among the directed mutations in *M. musculus* C > G and G > C transversions are underrepresented, compared to the fractions of such mutations among all point mutations in *H. sapiens*. Instead, C > T and G > A transitions are overrepresented in *M. musculus*. This might be due to GC-biased gene conversion being weaker in rodents [22]. Gene conversion is the transfer of genetic information between two homologous chromosomes carrying different allele variants during which one allele becomes substituted for the other. It has been shown that in mammals this process is biased in the direction that increases GC content [23]. If during recombination an S-W (where S is a C or G nucleotide and W is an A or T nucleotide) mismatched pair forms between two homologous DNA strands, the more probable scenario is that W will be converted into S. If gene conversion becomes weaker or less biased, then C > T and G > A transitions should become more frequent in observations. This is consistent with the observations of both a decrease in GC content of GC-rich isochores and an increase in GC-poor isochores in rodents [24].

Previously several hypermutable 4 bp mutation contexts were identified in *H. sapiens* [17], as shown in Table 2. We checked if these mutation regularities can be found in *M. musculus*. As shown in Table 2, only the $\{T > C \mid 2, ATTG\}$ mutation context that is hypermutable in *H. sapiens* is also somewhat hypermutable in *M. musculus* compared to its



(a)

(b)

(c)

(d)

Figure 1: Comparison of mutation bias and minimal contrasts for all 2–4 bp mutations contexts in *H. sapiens* and *M. musculus*. Each dot represents a mutation context. The x-axis of each plot represents the contexts minimal contrast values, and the y-axis represents the contexts mutation bias. The values of mutation bias and minimal contrast are given for *H. sapiens* (plots (a) and (c)) or *M. musculus* (plots (b) and (d)). The color scheme indicates the difference between mutation biases (plots (a) and (b)) and minimal contrasts (plots (c) and (d)). Thus red dots on (a) and (c) represent contexts that are hypermutable in *H. sapiens* compared to *M. musculus*, while green dots represent contexts that are hypermutable in *M. musculus* compared to *H. sapiens*. This color scheme is reversed for (b) and (d). Note that many dots are situated in pairs; this is because complimentary mutation contexts have very similar mutation bias and minimal contrast values.

other 4 bp contexts (among all 4 bp contexts in *M. musculus* this context is the third by minimal contrast values). However, even for this context, the observed values of mutation bias and minimal contrast are much lower than those in *H. sapiens*, indicating that context-dependent mutation regularities are very different between *H. sapiens* and *M. musculus* even at the 4 bp scale. One of our reviewers made an interesting observation that the reverse-complement image of the highly mutable *M. musculus* context {T > A | 3, TTTA} is {A > T | 2, TAAA} which is the reverse context for another highly mutable *M. musculus* context {T > A | 2, TTAA} (see Table 2). We checked if other highly mutable contexts have highly mutable reverse contexts, but this does not seem to be a general trend. Minimal contrast and mutation bias values for reverse contexts are also provided in Table 2.

We would like to explain why we make emphasis on minimal contrast and not on mutation bias, when presenting Table 2. If we sort contexts by mutation, bias all the highest ranking contexts in both *H. sapiens* and *M. musculus* will be 4 bp contexts containing the {C > G | 1, CG} context. However, most of the increase in their mutation rates is explained by the high mutation bias of the {C > G | 1, CG} context itself. Among multiple 3-4 bp contexts containing the {C > G | 1, CG} context some will inevitably have higher mutation bias than {C > G | 1, CG}, and some will have a lower mutation bias, but as long as the difference is small, these contexts are unlikely to provide interesting information about mutation regularities. Thus, we believe that minimal contrast is more informative when searching for biologically meaningful contexts.

A more detailed analysis of mutation regularities is presented, in Figure 1. Previously we found it helpful to plot mutation bias versus minimal contrast for 2–4 bp contexts to identify mutation regularities with large effects. Context-dependent mutation regularities are very different between *H. sapiens* and *M. musculus*. While both species share the mutation regularity of increased C > T mutation frequency in the CG word, three hypermutable 4 bp contexts previously identified in *H. sapiens* (Figure 1(a)) are not strikingly hypermutable in *M. musculus* (Figure 1(d)). In *M. musculus* comparing to *H. sapiens*, there is also a notable increase of both mutation bias and minimal contrast values for C > G mutations in the first position of the word CGA and in contexts that include this context as a subcontext; G > T mutations in the first position of the word GCGA; C > G and G > T mutations in CG dinucleotides (Figure 1(b)). These differences in mutation patterns might reflect differences in biological mechanisms involved in primate and rodent mutagenesis.

4. Conclusions

We have found a number of substantial differences in context-dependent mutation regularities of *Mus musculus* and *Homo sapiens*. These differences include the reduced mutation bias and minimal contrasts for mutation contexts {T > C | 2, ATTG}, {A > C | 1, ACAA}, and {T > C | 2, ATAG} in *M. musculus* when compared to *H. sapiens*. These mutation contexts are hypermutable in *H. sapiens*. Only {T > C | 2, ATTG} is hypermutable in *M. Musculus*, but to a smaller extent than in *H. sapiens*. Mutation bias and minimal contrasts are instead increased for {C > G | 1, CGA}, {C > G | 1, CG}, {G > T | 2, CG}, and {G > T | 1, GCGA} mutation contexts in *M. musculus* when compared to *H. sapiens*.

Acknowledgments

This work was supported by the Russian Ministry of Science and Education State Contracts 8494, 8100, and 14.740.11.1202 of the Federal Special Program "Scientific and Educational Human Resources of Innovative Russia" for 2009–2013 and the Russian Foundation for Basic Research Grants 12-04-91334, 11-04-91340, 13-07-00969, 12-04-31071, and 11-04-01511.

References

[1] C. F. Baer, M. M. Miyamoto, and D. R. Denver, "Mutation rate variation in multicellular eukaryotes: causes and consequences," *Nature Reviews Genetics*, vol. 8, no. 8, pp. 619–631, 2007.

[2] A. Kong, M. L. Frigge, G. Masson et al., "Rate of *de novo* mutations and the importance of father's age to disease risk," *Nature*, vol. 488, no. 7412, pp. 471–475, 2012.

[3] D. N. Cooper and M. Krawczak, "Cytosine methylation and the fate of CpG dinucleotides in vertebrates genomes," *Human Genetics*, vol. 83, no. 2, pp. 181–188, 1989.

[4] N. Arnheim and P. Calabrese, "Understanding what determines the frequency and pattern of human germline mutations," *Nature Reviews Genetics*, vol. 10, no. 7, pp. 478–488, 2009.

[5] A. Hodgkinson, E. Ladoukakis, and A. Eyre-Walker, "Cryptic variation in the human mutation rate," *PLoS Biology*, vol. 7, no. 2, Article ID e1000027, 2009.

[6] R. D. Blake, S. T. Hess, and J. Nicholson-Tuell, "The influence of nearest neighbors on the rate and pattern of spontaneous point mutations," *Journal of Molecular Evolution*, vol. 34, no. 3, pp. 189–200, 1992.

[7] D. G. Hwang and P. Green, "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 39, pp. 13994–14001, 2004.

[8] N. D. Singh, P. F. Arndt, A. G. Clark, and C. F. Aquadro, "Strong evidence for lineage and sequence specificity of substitution rates and patterns in Drosophila," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1591–1605, 2009.

[9] N. G. C. Smith, M. T. Webster, and H. Ellegren, "Deterministic mutation rate variation in the human genome," *Genome Research*, vol. 12, no. 9, pp. 1350–1356, 2002.

[10] J. C. Walser, L. Ponger, and A. V. Furano, "CpG dinucleotides and the mutation rate of non-CpG DNA," *Genome Research*, vol. 18, no. 9, pp. 1403–1414, 2008.

[11] M. J. Lercher and L. D. Hurst, "Human SNP variability and mutation rate are higher in regions of high recombination," *Trends in Genetics*, vol. 18, no. 7, pp. 337–340, 2002.

[12] D. Tian, Q. Wang, P. Zhang et al., "Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes," *Nature*, vol. 455, no. 7209, pp. 105–108, 2008.

[13] I. Hellmann, K. Prüfer, H. Ji, M. C. Zody, S. Pääbo, and S. E. Ptak, "Why do human diversity levels vary at a megabase scale?" *Genome Research*, vol. 15, no. 9, pp. 1222–1231, 2005.

[14] K. D. Makova and W. H. Li, "Strong male-driven evolution of DNA sequences in humans and apes," *Nature*, vol. 416, no. 6881, pp. 624–626, 2002.

[15] I. B. Rogozin, B. A. Malyarchuk, Y. I. Pavlov, and L. Milanesi, "From context-dependence of mutations to molecular mechanisms of mutagenesis," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 409–420, January 2005.

[16] M. Falconnet and S. Behrens, "Accurate estimations of evolutionary times in the context of strong CpG hypermutability," *Journal of Computational Biology*, vol. 19, no. 5, pp. 519–531, 2012.

[17] A. Y. Panchin, S. I. Mitrofanov, A. V. Alexeevski, S. A. Spirin, and Y. V. Panchin, "New words in human mutagenesis," *BMC Bioinformatics*, vol. 12, article 268, 2011.

[18] B. Yalcin, D. J. Adams, J. Flint, and T. M. Keane, "Next-generation sequencing of experimental mouse strains," *Mammalian Genome*, vol. 23, no. 9-10, pp. 490–498, 2012.

[19] K. Wong, S. Bumpstead, L. van der Weyden et al., "Sequencing and characterization of the FVB/NJ mouse genome," *Genome Biology*, vol. 13, no. 8, article R72, 2012.

[20] W. J. Kent, C. W. Sugnet, T. S. Furey et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[21] T. M. Keane, L. Goodstadt, P. Danecek et al., "Mouse genomic variation and its effect on phenotypes and gene regulation," *Nature*, vol. 477, no. 7364, pp. 289–294, 2011.

[22] Y. Clément and P. F. Arndt, "Substitution patterns are under different influences in primates and rodents," *Genome Biology and Evolution*, vol. 3, no. 1, pp. 236–245, 2011.

[23] N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret, "GC-content evolution in mammalian genomes: the biased gene conversion hypothesis," *Genetics*, vol. 159, no. 2, pp. 907–911, 2001.

[24] L. Duret, M. Semon, G. Piganeau, D. Mouchiroud, and N. Galtier, "Vanishing GC-rich isochores in mammalian genomes," *Genetics*, vol. 162, no. 4, pp. 1837–1847, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 902467, 10 pages
http://dx.doi.org/10.1155/2013/902467

Research Article

miR-1279, miR-548j, miR-548m, and miR-548d-5p Binding Sites in CDSs of Paralogous and Orthologous *PTPN12*, *MSH6*, and *ZEB1* Genes

Anatoliy T. Ivashchenko, Assel S. Issabekova, and Olga A. Berillo

National Nanotechnology Laboratory, Al-Farabi Kazakh National University, Almaty 050038, Kazakhstan

Correspondence should be addressed to Anatoliy T. Ivashchenko; a_ivashchenko@mail.ru

Received 27 February 2013; Revised 14 May 2013; Accepted 28 May 2013

Academic Editor: Vassily Lyubetsky

Only *PTPN12, MSH6*, and *ZEB1* have significant miR-1279 binding sites among paralogous genes of human tyrosine phosphatase family, DNA mismatch repair family, and zinc finger family, respectively. All miRNA binding sites are located within CDSs of studied mRNAs. Nucleotide sequences of hsa-miR-1279 binding sites with mRNAs of human *PTPN12, MSH6*, and *ZEB1* genes encode TKEQYE, EGSSDE, and GEKPYE oligopeptides, respectively. The conservation of miRNA binding sites encoding oligopeptides has been revealed. MRNAs of many paralogs of zinc finger gene family have from 1 to 12 binding sites coding the same GEKPYE hexapeptide. MRNAs of *PTPN12, MSH6*, and *ZEB1* orthologous genes from different animal species have binding sites for hsa-miR-1279 which consist of homologous oligonucleotides encoding similar human oligopeptides TKEQYE, EGSSDE, and GEKPYE. MiR-548j, miR-548m, and miR-548d-5p have homologous binding sites in the mRNA of *PTPN12* orthologous genes which encode PRTRSC, TEATDI, and STASAT oligopeptides, respectively. All regions of miRNA are important for binding with the mRNA.

1. Introduction

Posttranscriptionally miRNAs regulate the expression of genes involved in organism development [1], metabolism [2], cell cycle, apoptosis [3], carcinogenesis [4], protein-protein interactions [5], and so forth. The specific functions of most identified miRNAs remain unknown, although some targets have been established. Computational algorithms have been developed to predict miRNA targets [6]. For animal miRNAs, target prediction is an issue because miRNAs can bind to target genes without perfect complementarity [7]. Only some predicted miRNA binding sites had been experimentally verified [8].

Many pieces of software initially used complementarity to identify potential binding sites. Subsequent filtering steps were based on thermodynamics, binding site structures, and site conservation across species [9]. The false-positive rate of predicting miRNA targets has resulted in the development of several algorithms: miRanda (http://www.microrna.org/microrna/home.do) [10], TargetScan (http://www.targetscan.org/) [11], and PicTar (http://pictar.mdc-berlin.de/) [12]. Other algorithms apply thermodynamics as the initial requirement for selection of interaction sites between miRNAs and their targets. DIANA-microT (http://diana.cslab.ece.ntua.gr/) [13] and RNAhybrid (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) [14] belong to this type of software. We have used a local RNAHybrid software to identify miRNA target sites and have attempted to explain the benefits of this software [15]. The software allows the determination of the most energetically favourable interactions of small RNAs to mRNAs in all regions: $5'$untranslated regions ($5'$UTR), coding domain sequences (CDS), and $3'$untranslated regions ($3'$UTR). The majority of on-line pieces of software predict miRNA target sites only in the $3'$UTR, and most of the characterised miRNA binding sites occur in this region [16]. Experimental investigations revealed the occurrence of $5'$UTR and CDS miRNAs binding sites [16–18]. According to another study, human mRNAs contain many strong sites with 13 or more base pairings without bulges in the middle [19].

Previously, we had revealed miRNA binding sites in 54 mRNAs of oncogenes including *PTPN12, MSH6,* and *ZEB1* that had sites in CDSs with a high $\Delta G/\Delta G_m$ ratio [19]. These genes are involved in the development of breast cancer [20], colorectal cancer [21], lymphoma [22], and esophageal cancer [23]. The *PTPN12* gene encodes a protein from tyrosine phosphatase family. It participates in different cellular processes, including cell growth, the mitotic cycle, and oncogenic transformation [24]. *MSH6* is a DNA mismatch repair protein that participates in the recognition of mismatched nucleotides prior to their repair [25]. Mutations in this gene have been associated with hereditary nonpolyposis colon cancer and endometrial cancer [26]. The *ZEB1* gene encodes a protein which belongs to a large family of zinc finger transcriptional factors [27]. Identification of the role of miRNA in the regulation of the gene and the gene set expression is an important problem. Thus, we have examined the presence of miR-1279 target sites in paralogous genes for *PTPN12, MSH6,* and *ZEB1*.

MiRNA sites are localized within the 5′UTR, CDS, and 3′UTR of mRNA [28]. Indeed, about half of all miRNA binding sites are located in the protein-coding region [29] suggesting that studies should pay much attention to these sites. The nucleotide placement of these binding sites coding the amino acid sequence may raise important questions about binding sites localized in the CDSs.

To reduce the number of false positives, target sites may be validated by conservation of sites across species. To verify predicted sites by RNAHybrid, we have examined conservation of these sites across orthologous genes of different organisms. Significant sites in human genes have been investigated in *PTPN12, MSH6,* and *ZEB1* orthologues. Investigating these problems is necessary to establish the role of miRNA in the regulation of single gene and gene sets expressions.

In present work we analysed the conservation of miR-1279, miR-548 m, miR-548j and miR-548d-5p binding sites inorthologues and paralogues of the *PTPN12*, *MSH6* and *ZEB1* genes to verify these interactions.

2. Materials and Methods

Investigation objects were mRNAs of *PTPN12*, *MSH6*, and *ZEB1* human genes, their orthologs, and paralogs (Supplementary Tables S1, S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2013/902467) in *Anolis carolinensis* (Aca), *Ailuropoda melanoleuca* (Ame), *Bos taurus* (Bta), *Cricetulus griseus* (Cgr), *Cavia porcellus* (Cpo), *Callithrix jacchus* (Cja), *Canis lupus familiaris* (Clf), *Danio rerio* (Dre), *Equus caballus* (Eca), *Gallus gallus* (Gga), *Homo sapiens (Hsa), Loxodonta africana* (Laf), *Macaca mulatta* (Mml), *Monodelphis domestica* (Mdo), *Meleagris gallopavo* (Mga), *Mus musculus* (Mmu), *Nomascus leucogenys* (Nle), *Ornithorhynchus anatinus* (Oan), *Oryctolagus cuniculus* (Ocu), *Oreochromis niloticus* (Oni), *Otolemur garnettii* (Oga), *Pan paniscus* (Ppa), *Papio anubis* (Pan), *Pan troglodytes* (Ptr), *Pongo abelii* (Pab), *Rattus norvegicus* (Rno), *Saimiri boliviensis* (Sbo), *Sarcophilus harrisii* (Sha), *Sus scrofa* (Ssc), *Taeniopygia guttata* (Tgu), *Xenopus laevis* (Xla), and *Xenopus tropicalis* (Xtr). Nucleotide and amino acid sequences were accessed from GenBank (http://www.ncbi.nlm.nih.gov/). Nucleotide sequence of hsa-miR-1279, hsa-miR-548 m, and hsa-miR-548j were received from miRBase (http://www.mirbase.org/).

Free energy (ΔG), position of potential binding sites, and interaction schemes were calculated by RNAhybrid 2.1 software (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/). The E-RNAhybrid software (http://sites.google.com/site/malaheenee/software/) was used to compute the ratio of $\Delta G/\Delta G_m$ and *P*value. $\Delta G/\Delta G_m$ value equalled 75% and more was used as comparative criterion of miRNA and mRNA interaction force. $\Delta G/\Delta G_m$ value shows degree of complementarity of each miRNA site in target gene. ΔG value and its standard deviation were used to determine significance level of miRNA binding sites with mRNA. Interaction energies (ΔG_m) of hsa-miR-1279, hsa-miR-548 m, hsa-miR-548 j, and hsa-miR-548 d-5 p with their perfectly complementary sequence were calculated and equaled −118 kJ/moL, −139 kJ/moL, −161 kJ/moL, and −148 kJ/moL, consequently. At first, miRNA binding sites were found for human mRNAs, and then they were revealed in homologous and paralogous genes of studied animals. Diagrams of nucleotide and protein variability were produced by WebLogo software (http://weblogo.threeplusone.com/).

3. Results

3.1. Features of hsa-miR-1279 Binding Sites in mRNAs of Paralogous and Orthologous PTPN12 Genes. To investigate the role of miR-1279 in the regulation of tyrosine phosphatase family genes, we attempted to elucidate target binding sites among them. We identified the *PTPN12* gene as a reliable target of hsa-miR-1279 with binding energy of −102.1 kJ/moL and a $\Delta G/\Delta G_m$ value of 86.5%. hsa-miR-1279 consists of 17 nucleotides; 15 of these nucleotides are perfectly complementary to the binding site of human *PTPN12* mRNA. We considered miRNA binding sites with $\Delta G/\Delta G_m$ values greater than 75%. All sites with lower values were determined to be weak sites and were not considered. The hybridization energy of miR-1279 with the mRNA of tyrosine phosphatase family genes, except that of *PTPN12*, varied from −67.7 kJ/moL to −89.5 kJ/moL. Corresponding $\Delta G/\Delta G_m$ values ranged from 57.4% to 75.0%. Thus, only *PTPN12* was considered. The nucleotide sequence of the binding site in PTPN12 encoded the TKEQYE hexapeptide, which was not a conserved amino acid sequence in any other tyrosine phosphatase family. Consequently, miR-1279 may efficiently regulate only *PTPN12* gene expression.

Furthermore, we investigated the conservation of the miR-1279 binding site in orthologous genes. The human *PTPN12* gene has 23 orthologues in different animals. For determination of the conservation of the relevant site, we verified the nucleotide sequence of the binding site. The nucleotide sequences of these sites in the mRNA of the investigated genes are presented in Table 1. Sequences of the mRNA *PTPN12* binding sites in all animals were homologous, and nucleotides in first and second positions of the codons were conserved (Figure 1(a)). ACGAAGGAGCAGUAUGAA oligonucleotide of the interaction site in the CDS encoded the

Table 1: Characteristics of miR-1279 binding sites in CDSs of *PTPN12* orthologous genes.

Object	Region of CDS in *PTPN12* orthologous genes	Position in CDS, nt	$\Delta G/\Delta G_m$, %
Hsa	CAAACAAAGGAGCAAUAUGAACUUG	788	86.5
Ptr	· ·	788	86.5
Cgr	· ·	836	87.9
Ppa	· ·	836	86.5
Aca	· · · · · G · · · · · A · · G · · · · · · · · · ·	836	74.5
Ame	· · · · · U · · · · · · · · G · · · · · G · · · ·	840	83.3
Bta	· · · · · U · · · · · · · · · · · · · · G · · · ·	840	83.3
Cja	· · · · · · · · · · · A · · · · · · · · G · · · ·	836	63.1
Clf	· · · · · U · · · · · · · · G · · · · · G · · · ·	836	83.0
Dre	· · G · · C · · A · · · · · G · · · · · G · · G ·	828	71.6
Eca	· · G · · U · · · · · · · · G · · · · · G · · · ·	447	83.0
Gga	· · G ·	788	86.5
Laf	· G · · · ·	788	86.5
Mdo	· · · · · · · · · · · · · · G · · · · · · · · C ·	908	63.1
Mga	· · G · · · · · · · · · · · G · · · · · · · · · ·	836	86.5
Mmu	· · · · · · · · · · · · · · G · · · · · · · · · ·	788	86.5
Nle	· G · · · ·	677	86.5
Oan	· · · · · G · · · · · · · · · · · · · · · · · · ·	479	87.9
Ocu	· G · · · ·	479	86.5
Oga	· G · · · ·	944	86.5
Oni	· · · · · C · · · · · · · · G · · · · · G · · G ·	765	83.0
Pab	· G · · · ·	677	86.5
Pan	· G · · · ·	836	86.5
Sbo	· G · · · ·	836	86.5
Rno	· · · · · G · · · · · · · · G · · · · · · · · C ·	479	87.9
Tgu	· · · · · · · · · · · · · · G · · · · · · · · · ·	815	86.5
Xla	· · G · · · · · · · · · · · G · · C · · G U · G ·	836	71.3
Xtr	· · G · · · · · · · · · · · G · · · · · G U · G ·	788	86.5

Note: the dots are identical nucleotides or amino acids in data presented in Tables 1–6.

TKEQYE hexapeptide, which was identical in all orthologous proteins of the studied species (Table 1). Replacements in the third position of the codon in described oligonucleotide sequences did not change the amino acid content. This oligopeptide was located from positions13 to 18 in the conserved amino acid sequences of tyrosine phosphatase orthologues (Figure 1(b)). These data indicate that the main function of this site is gene regulation by miR-1279. All miR-1279 interaction sites in *PTPN12* orthologues corresponded to an open frame. Replacements of one or several nucleotides in the third position of the codon led to decreased binding energies, but did not change the location of the miR-1279 binding site. Interactions between hsa-miR-1279 and the mRNA of *PTPN12* orthologues showed significance levels of $\Delta G/\Delta G_m$ values ranging from 71.3% to 87.9% (Table 1). MiRNA binding sites were identical in 11 species from mammals to amphibians.

The binding energy of orthologous mRNA with orthologous miRNA may vary if the sequences of orthologous miRNAs in animals differ from that of hsa-miR-1279. It is possible that the interactions of hsa-miR-1279 with the mRNA of *A. carolinensis*, *C. jacchus*, *D. rerio*, *M. domestica*, and *X. laevis* tyrosine phosphatase orthologues are less effective. These data confirm that this binding site exists from the early stages of evolution and did not change for tens to hundreds of millions of years.

3.2. Features of hsa-miR-1279 Binding Sites in mRNAs of Paralogous and Orthologs MSH6 Genes. The human genome encodes several DNA mismatch repair proteins: *MSH2*, *MSH3*, *MSH4*, *MSH5*, *MSH6*, *MLH1*, and *MLH3*. We did not find conserved strong miR-1279 binding sites in the mRNAs of these genes, with the exception of the *MSH6* gene, among all considered DNA mismatch repair proteins. Nonconserved miR-1279 interaction sites in these genes had $\Delta G/\Delta G_m$ values ranging from 62.0% to 80.5%, and only the *MLH3* gene had a significant binding site ($\Delta G/\Delta G_m$ value is 73%). These data confirmed the specific binding of miR-1279 with *MSH6* mRNA among genes of the DNA mismatch repair family.

The hsa-miR-1279 binding site in human *MSH6* mRNA had perfect complementary nucleotides with a single nucleotide bulge, including 4 G-U base pairs, in spite of

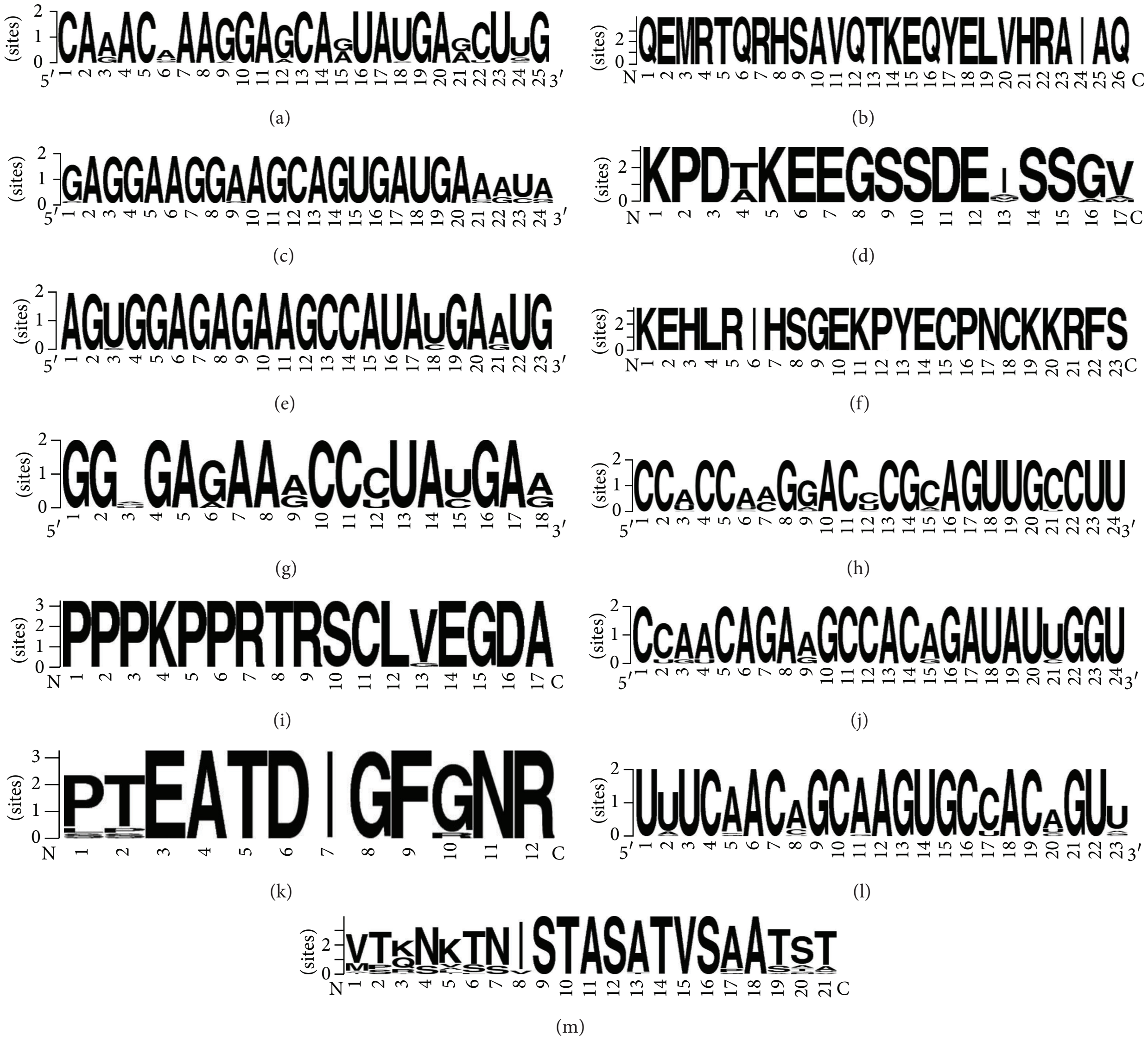


Figure 1: Fragments of nucleotide and amino acid sequences, where miRNA binding sites are located. The variability of nucleotides in miR-1279 binding sites in mRNA of *PTPN12* (a), *MSH6* (c), and *ZEB1* (e) orthologous genes and the variability of amino acids of PTPN12 (b), MSH6 (d), and ZEB1 (f) orthologous proteins; miR-1279 binding site encode TKEQYE (b), EGSSDE (d), and GEKPYE (f) oligopeptides; (g), the variability of nucleotides in miR-1279 binding sites of zinc-finger transcriptional factors family; (h), (j) and (l)—the variability of nucleotides in 548j, 548 m, and miR-548d-5p binding sites in mRNA of *PTPN12*, respectively; (i), (k), and (m), the variability of amino acids of PTPN12 orthologous proteins in regions with PRTRSC, EATDI, and STASAT oligopeptides, respectively.

its high $\Delta G/\Delta G_m$ value (89.4%) and binding energy (−105.4 kJ/moL). The GAAGGAAGCAGUGAUGA oligonucleotide sequence in the binding site of *MSH6* mRNA (Figure 1(c)) encoded the EGSSDE hexapeptide in MSH6 protein (Table 2). We observed 34 orthologues for human *MSH6* and miR-1279 binding sites with high complementarily levels in 15 orthologues. Oligonucleotides of these sites and corresponding oligopeptides are shown in Table 2. MSH6 protein from 13 mammals and human contained completely homologous EGSSDE hexapeptides and the corresponding nucleotide sequences in mRNA binding sites exhibited high homology. There was nucleotide variability in the third position of the glutamic acid codons at the beginning of the EGSSDE hexapeptide, situated from positions 13 to 30 in the highly conserved protein region of MSH6 (Figure 1(d)). These data suggest the functional importance of the miR-1279 binding site in the regulation of *MSH6* gene expression.

All binding sites in *MSH6* orthologous mRNA corresponded to an open reading frame. Replacement of one or several nucleotides in the third position of codons led to diminished hybridization energy but did not change the location of the hsa-miR-1279 binding site. Such replacements did not influence the amino acid composition properties of the protein. The $\Delta G/\Delta G_m$ value of interaction sites varied from 87.6% to 89.4% (Table 2).

3.3. Features of hsa-miR-1279 Binding Sites in mRNAs of Human ZNF Family and ZEB1 Orthologous Genes. The *ZEB1* gene is a member of the large ZNF gene family of transcription factors. We searched for miR-1279 binding sites in the CDSs of several genes from this family. *ZNF552* and *ZNF790* mRNAs had miR-1279 sites, where the $\Delta G/\Delta G_m$ values equalled 89.4% and 87.9%, respectively. Nucleotide sequences of the binding sites in these 2 genes shared high homology and encoded a GEKPYE oligopeptide. In fact, while many genes of this family had the GEKPYE oligopeptide, not all were targets for miR-1279. There were no significant binding sites for hsa-miR-1279 in *ZNF91*, *ZNF148*, *ZNF208*, *ZNF232*, *ZNF236*, *ZNF423*, *ZNF495B*, *ZNF509*, *ZNF521*, *ZNF729*, *ZNF776*, *ZNF768*, *ZNF853*, *ZNF857A*, and *ZNF865* genes within the CDS, and these genes did not encode the GEKPYE oligopeptide.

Proteins encoded by *ZEB2*, *ZNF8*, *ZNF70*, *ZNF84*, *ZNF140*, *ZNF454*, *ZNF461*, *ZNF475*, *ZNF529*, *ZNF576*, *ZNF594*, *ZNF713*, *ZNF751*, and *ZNF755* genes had this oligopeptide in structure, but the corresponding oligonucleotides were bound by hsa-miR-1279 with low hybridization energies. These oligonucleotides differed from the human *ZEB1* gene sequence by replacements in the third positions of several codons (Figure 1(g)). These genes have between 1 and 7 sites encoding the GEKPYE hexapeptide. We assume that if miR-1279 would be bound to these sequences, gene expression of these proteins would be inhibited. The expression of majority of the genes in the zinc-finger transcription factor family was not regulated by miR-1279.

The strongest binding site in genes of the zinc-finger transcription factor family was found in *ZEB1* mRNA. The binding energy of miR-1279 with *ZEB1* mRNA was 90% of the ΔG_m, and 16 out of 17 nucleotides from the hsa-miR-1279 sequence exhibited total complementarity. Thus, it is suggested that they may interact in vivo. To verify the miR-1279 binding site in the *ZEB1* gene, we investigated the interactions between hsa-miR-1279 and mRNAs of *ZEB1* orthologues from 12 animal species. The human *ZEB1* gene has 16 identified orthologues, but only 12 of them retained this hsa-miR-1279 binding site and had conserved amino acid sequences in the binding region. The nucleotide sequence of the miR-1279 binding site (GGAGAGAAGCCAUAUGA) had high homology between these 12 orthologues (Figure 1(e)). The third nucleotide of these codons sequences encoded changes in tyrosine and glutamic acid, but these changes did not influence the encoding of the GEKPYE oligopeptide. This oligopeptide was localized in the conserved region of orthologous proteins (Figure 1(f)). Characteristics of the interactions between hsa-miR-1279 and *ZEB1* orthologue mRNAs of different animals are shown in Table 3. This binding site was well conserved among mammals and birds but was found in amphibians with less accuracy.

3.4. Features of hsa-miR-548j Binding Sites in mRNAs of PTPN12 Orthologous Genes. *PTPN12* mRNA contained near perfect binding sites for several miRNAs (miR-1279, miR-548j, and miR-548 m). In the first part of this study, we validated miR-1279 binding site conservation across species. Furthermore, we considered the conservation of miR-548j interaction sites in *PTPN12* mRNA. MiR-548j has intronic origins (intron 9 in the *TPST2* gene). The nucleotide sequence of the miR-548j binding site in the CDS of *PTPN12* mRNA exhibited low variability and was present in all investigated organisms from mammals to fish (Figure 1(h)). Characteristics of these interactions are presented in Table 4. The ΔG_m value of miR-548j with a completely complementary sequence equalled −161 kJ/moL. The nucleotide sequence of the miR-548j binding site corresponded to the PRTRSC oligopeptide. This hexapeptide did not change across 16 species of mammals, reptiles, and birds but exhibited some changes in 4 species of fishes and amphibians (Figure 1(i)). Such conservation of revealed binding sites indicates the possibility of regulator role of miR-548j in *PTPN12* gene expression.

The third miRNA that binds to *PTPN12* mRNA was miR-548 m. This binding site showed a lower level of conservation than miR-1279 and miR-548j binding sites. MiR-548 m interacted with the site encoding a conserved TEATDI hexapeptide in 7 species (Figure 1 (k)). A homologous oligopeptide from other animal species exhibited mismatches in 1 or 2 amino acids but had a similar position in a conserved region of the protein. The ΔG_m value of miR-548 m with a perfectly complementary sequence equalled −139.4 kJ/moL. Characteristics of miR-548 m binding sites are presented in Table 5 and showed a high level of conservation. In fact, the binding site only had perfect conservation across 4 investigated species (chimpanzee, gibbon, horse, and elephant). The $\Delta G/\Delta G_m$ values vary from −75.1% to −80.8% (Table 5). Consequently, the effect of miR-548 m on *PTPN12* mRNA expression was less than those of miR-1279 and miR-548j.

3.5. Features of hsa-miR-548d-5p Binding Sites in mRNAs of PTPN12 Orthologous Genes. MiR-548d-5p had another binding site in *PTPN12* mRNA. The binding energy of this site was lower than in the miR-548j and miR-548 m binding sites, but the nucleotide sequence in this site exhibited conservation in orthologous genes (Table 6). The amino acid sequences up- and downstream of the hexapeptide STASAT of PTPN12 were variable (Figure 1 (m)). The nucleotide sequences of the miR-548d-5p binding sites and *PTPN12* mRNA were homologous (Figure 1(l)); however, nucleotides in the third codon position were variable, and thus, these variable nucleotides did not cause changes in the amino acids coding of hexapeptide STASAT in some cases.

3.6. Nucleotide Interactions of hsa-miR-1279 Binding Sites with mRNAs of PTPN12, MSH6, and ZEB1 . Next, we investigated the regions of the miRNA that are important for target gene regulation. Binding sites may be different according to the primary contribution of the specific region of miRNA to the hybridization energy. We observed 3 types of miRNA contributions: (1) the contribution of 5′-end dominated, (2) the contribution of the central region dominated, and (3) the contribution of 3′-end dominated (Table 7). A schematic representation of the nucleotide sequences of hsa-miR-1279

Table 2: Characteristics of miR-1279 binding sites in CDSs of *MSH6* orthologous genes and the variability of amino acids of MSH6 orthologous proteins in regions with EGSSDE oligopeptide.

Object	Region of CDS in *MSH6* orthologous genes	Region of *MSH6* orthologous proteins	Position in CDS, nt	$\Delta G/\Delta G_m$, %
Hsa	GAGGAAGGAAGCAGUGAUGAAAUA	KPDTKEEGSSDEISSGV	812	89.4
Mml	························	·················	812	89.4
Pan	························	·················	812	89.4
Nle	························	·················	602	89.4
Pab	························	·················	818	89.4
Ppa	························	·················	629	89.4
Ptr	························	·················	929	89.4
Sbo	························	·················	809	89.4
Cpo	····················G···	················A	806	89.4
Clf	························	···A·············	578	89.4
Eca	························	···A·············	758	89.4
Mdo	·········G···········GC·	············A···M	1019	87.6
Mmu	C···················CGCG	·················	812	88.3
Ocu	·····················G·U	···A········V··A·	809	89.4
Sha	·····················GC·	············A···M	998	87.6
Ssc	·······················G	············M····	815	89.4

Table 3: Characteristics of miR-1279 binding sites in CDSs of *ZEB1* orthologous genes.

Object	Region of CDS in *ZEB1* orthologous genes	Position in CDS, nt	$\Delta G/\Delta G_m$, %
Hsa	AGUGGAGAGAAGCCAUAUGAAUG	741	89.4
Nle	·······················	741	89.4
Ppa	·······················	741	89.4
Ptr	·······················	741	89.4
Sbo	·······················	741	89.4
Ame	·······················	810	89.4
Bta	·······················	792	89.4
Cja	·······················	792	89.4
Pab	·······················	792	89.4
Clf	·······················	588	89.4
Mml	·······················	750	89.4
Oga	·······················	729	89.4
Pan	·······················	770	89.4
Gga	····················G··	789	89.4
Mmu	·················C·····	730	74.1
Sha	··C····················	741	89.4
Xla	·················C··G··	798	74.1
Xtr	·················C··G··	732	74.1

binding sites in the mRNA of *PTPN12*, *MSH6*, and *ZEB1* from different animals is shown in Table 7. The miR-1279 binding site in the *PTPN12* gene of 14 animal species was 5′-dominant and evolutionary conserved. *PTPN12* mRNAs from *A. melanoleuca*, *C. lupus*, *E. caballus*, and *O. niloticus* also had perfect complementarity in 5′-end of miR-1279 but had a reduced number of conserved nucleotides than the first gene group and had several mismatches in the target region. mRNAs of *MSH6* from 11 animal species had 3′-dominant miR-1279 sites and were highly homologous in all species. The 3′-end of miR-1279 was the primary contributor in these binding sites. MiR-548j was bound to the mRNA of the investigated species through the central region or the 3′-end in *O. niloticus* and *D. rerio* (Table 7). As shown in Table 7, miR-548 m and miR-548d-5p interacted predominantly with the 3′-end and, in several cases, the central region (*A. carolinensis*, *C. jacchus*, *Gallus gallus*, *M. gallopavo*, *O. niloticus,* and *P. abelii*). All described data confirmed that the nucleotide sequences of mRNA site may bind with all miRNA regions.

Table 4: Characteristics of miR-548j binding sites in CDSs of *PTPN12* orthologous genes and the variability of amino acids of PTPN12 orthologous proteins in regions with PRTRSC oligopeptide.

Object	Region of CDS in *PTPN12* orthologous genes	Region of *PTPN12* orthologous proteins	Position in CDS, nt	$\Delta G/\Delta G_m$, %
Hsa	CCACCAAGGACCCGCAGUUGCCUU	PPPKPPRTRSCLVEGDA	652	80.3
Ame	························	·················	1010	80.3
Cja	························	·················	1010	80.3
Pan	························	·················	1010	80.3
Ppa	························	·················	1010	80.3
Ptr	························	·················	1010	80.3
Sbo	························	·················	1010	80.3
Bta	························	·················	1009	80.3
Clf	························	·················	620	80.3
Pab	························	·················	941	80.3
Eca	··G········U············	·················	938	81
Laf	························	············G····	1010	80.3
Ocu	·····U·····U············	·················	989	79.5

Table 5: Characteristics of miR-548m binding sites in CDSs of *PTPN12* orthologous genes and the variability of amino acids of PTPN12 orthologous proteins in regions with EATDI oligopeptide.

Object	Region of CDS in *PTPN12* orthologous genes	Region of *PTPN12* orthologous proteins	Position in CDS, nt	$\Delta G/\Delta G_m$, %
Hsa	CCAACAGAAGCCACAGAUAUUGGU	PTEATDIGFGNR	2263	76.3
Pan	························	············	2263	76.3
Ppa	························	············	2263	76.3
Ptr	························	············	2263	76.3
Eca	························	············	2195	76.3
Nle	························	············	1905	76.3
Clf	········G···············	············	1878	75.1
Laf	··G·····················	············	2336	76.3
Cgr	·U············G·········	L···········	2201	75
Ocu	···U················C···	·S··········	2243	80.8
Ame	········G···············	SP·······R··	2263	75.1

Table 6: Characteristics of miR-548d-5p binding sites in CDSs of *PTPN12* orthologous genes and the variability of amino acids of PTPN12 orthologous proteins in regions with STASAT oligopeptide.

Object	Region of CDS in *PTPN12* orthologous genes	Region of *PTPN12* orthologous proteins	Position in CDS, nt	$\Delta G/\Delta G_m$, %
Hsa	UUUCAACAGCAAGUGCCACAGUU	VTQNKTNISTASATVSAATST	1878	65.9
Pan	·······················	·····················	1878	65.9
Cja	·······················	·····················	1875	65.9
Nle	·······················	·····················	1872	65.9
Ptr	·······················	·····················	1872	65.9
Pab	·······················	·····················	1521	65.9
Ppa	·······················	·····················	1875	65.9
Sbo	·······················	·····················	1875	65.9
Aca	·······U··G············	····IS·V··········A·A	1881	64.5
Ame	·······················	··T··················	1881	65.9
Bta	···················U···	··K···S·············S	1011	59.7
Clf	·······················	··K·············G····	1494	65.9
Eca	···················C···	A·K···S··············	1806	55.2
Gga	·······C···········U···	MSRSAS·············T·	1872	55.5
Laf	····C········A·········	·PK················G·	1884	65.1
Mdo	················U······	·PK·V················	1950	59.4
Mga	······················G	MSRSAS·············T·	1872	60.6
Mmu	·······················	··R···S·········P·S·A	1872	59.4
Ocu	·······················	··K··················	1860	65.9
Oga	·······················	··K···S··············	1923	65.9
Rno	···················C··G	··K···S·········P·S··	1860	49.9
Sha	····G··················	IPK·V················	1920	59.4
Tgu	·G·····C···········C··C	MSKSVS·V···········N·	2146	56.9
Xla	·A·····C···········U··A	M·KSLL············SAV	1803	50.1
Xtr	·A·····C···········U··A	M·KSLS············SAA	1821	50.1

Table 7: Schemes of hsa-miR-1279 biding with mRNA of *PTPN12, MSH6,* and *ZEB1* orthologous genes, and schemes of hsa-miR-548j, hsa-miR-548m, and hsa-miR-548d-5p biding with *PTPN12* orthologous gene.

	Scheme of binding site	Objects with same binding site
mRNA *PTPN12* miR-1279	5′ N N 3′ AAAGGAGCAAUAUGA UUUCUUCGUUAUACU 3′ UC 5′	Cgr, Gga, Hsa, Laf, Mga, Mmu, Nle, Oan, Ocu, Oga, Pab, Pan, Ppa, Ptr, Rno, Sbo, Tgu, Xla, Xtr
mRNA *MSH6* miR-1279	5′ N A N 3′ GGAAGGAAGCAGUG UGA UCUUUCUUCGUUAU ACU 3′ 5′	Clf, Cpo, Eca, Hsa, Mdo, Mml, Mmu, Nle, Ocu, Pab, Pan, Ppa, Ptr, Sbo, Sha, Ssc
mRNA *ZEB1* miR-1279	5′ N C N 3′ GGAGAGAAGC AUAUGA UCUUUCUUCG UAUACU 3′ U 5′	Ame, Cja, Clf, Bta, Gga, Hsa, Mml, Nle, Oga, Pab, Pan, Ppa, Ptr, Sha, Sbo
mRNA *PTPN12* miR-548j	5′ C N C 3′ ACCAAGGA CCGCAGUUGC UGGUUUCU GGCGUUAAUG 3′ 5′	Ame, Bta, Cja, Clf, Eca, Hsa, Laf, Pab, Pan, Ppa, Ptr, Sbo
mRNA *PTPN12* miR-548m	5′ N G 3′ CAGAAGCCACAGAUAUU GUUUUUGGUGUUUAUGG 3′ AAAC 5′	Ame, Cgr, Clf, Eca, Hsa, Laf, Nle, Ocu, Pan, Ppa, Ptr
mRNA *PTPN12* miR-548d-5p	5′ A U U G 3′ GCAAG GCCACAGUU CU CGUUU UGGUGUUAA GA 3′ C U U AAA 5′	Cja, Hsa, Nle, Ocu, Oga, Pab, Pan, Ppa, Ptr, Sbo, Ame, Clf

Note: N: any nucleotide—A, G, U, or C nucleotides.

4. Discussion

Previously, it was shown that 54 human mRNAs involved in oncogenesis contained strong miRNA binding sites within their CDSs that were predicted by RNAhybrid [19]. Interactions between mRNAs and miRNAs in binding sites were selected using $\Delta G/\Delta G_m$ values, allowing us to identify miRNA interactions with near perfect complementarity. We examined the variation in sequences of binding sites and have been able to show good conservation of strong binding sites among the organisms we have investigated.

It was predicted that miR-127 was 9 bound to mRNA of *PTPN12*, *MHS6*, and *ZEB1* in sites encoding 3 different oligopeptides (Tables 2, 4, and 6). miRNA binding sites may correspond to 3 open reading frames. Interactions between *PTPN12*, *MSH6*, and *ZEB1* mRNAs and miR-1279, as well as interactions between *PTPN12* mRNA and miR-548j, miR-548 m, and miR-548d-5p correspond to different open reading frames in their mRNA targets. The ability of miRNA binding sites to encode 1 oligopeptide suggests the stable influence of miRNA on mRNA during evolution.

Similar conservation of nucleotide regions in CDSs of studied mRNAs and corresponding amino acid of proteins has been established for human *APC*, *BAD*, *EPHB2*, and *MMP2* genes (unpublished data). MiRNA binding sites in plant proteins encode regions of mRNAs in paralogous genes of the SPL family with high homology of oligonucleotides and corresponding hexapeptides (unpublished data). The nucleotide sequences adjacent to binding sites and to the corresponding amino acids have high variability (Figure 1). Highly homologous miRNA binding sites of orthologous genes and conservative oligopeptides of corresponding proteins have been established.

The binding sites of several miRNAs are located in the 3′UTRs of some orthologous genes in the Drosophila genome and have highly homologous nucleotide sequences [30]. All nucleotides of miRNAs that participate in forming binding sites with mRNA in plants are highly conserved and not only in seeds [31]. These data prove the high conservatism of many miRNAs and show the importance of all nucleotide sequences of miRNAs with respect to binding to mRNAs. The presence of many target genes for a single miRNA suggests the functional connection of these genes. According to the data collected for the zinc-finger transcription factor family, the presence of oligopeptides that associate with the miRNA binding site in mRNA is not necessarily sufficient for miRNA site prediction. Changes in the third position of codons may cause weakening of the miRNA binding site or result in lack of binding ability. We have found this effect in many genes from the zinc-finger family. Mutations in the third position of codons can lead to loss of functional importance of miRNA binding.

Target site conservation is one of the primary prediction validation methods. We have used this prediction to evaluate several factors. Investigation of 3 different gene families demonstrated that miR-1279 efficiently regulated only 1 gene from the tyrosine phosphatase family and DNA mismatch repair family. Furthermore, miR-1279 affected several mRNAs from paralogous genes of the zinc-finger transcription factor family.

Single miRNAs can also affect gene sets by conservation of the target site in orthologous genes. MiR-1279 has been shown to have a high probability of regulating *PTPN12*, *MHS6*, and *ZEB1*. Estimation of the degree of binding of miRNA with orthologous mRNAs has demonstrated that the degree of complementarity increases from fish and amphibians to mammals and is identical across primates. Similar trends have been identified in the *ZEB1* gene. Additionally, several miRNAs have been confirmed to affect a single gene. These data support that miR-1279, miR-548j, miR-548 m, and miR-548d-5p may potentially regulate *PTPN12* gene expression.

5. Conclusion

High conservation of binding sites in orthologous genes has proven the importance and antiquity of miRNA interactions with mRNAs of target genes. The expression of some paralogous genes can be regulated by a single miRNA. However, repression of the expression of entire gene families via 1 miRNA is unlikely. The expression of one gene can be regulated by several miRNAs, presupposing multiple methods of gene regulation that include dependence on the expression of intronic miRNAs. Establishment of multiple miRNA interactions with mRNAs is a complex problem, but it is necessary for definite regulation of gene expression through miRNAs. The power of miRNA linkage with mRNAs in CDSs can change at the expense of nucleotide replacement in the third position of the codons. This phenomenon has been clearly shown in miR-1279 binding sites in the mRNAs of human ZNF family genes.

Acknowledgments

This study was supported by grant from the Ministry of Education and Science, Kazakhstan Republic. The authors would like to thank V. Khailenko for the software E-RNAhybrid 2.1.

References

[1] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-4*," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[2] P. Xu, S. Y. Vernooy, M. Guo, and B. A. Hay, "The Drosophila microRNA mir-14 suppresses cell death and is required for normal fat metabolism," *Current Biology*, vol. 13, no. 9, pp. 790–795, 2003.

[3] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen, "bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila," *Cell*, vol. 113, no. 1, pp. 25–36, 2003.

[4] M. Liu and H. Chen, "The role of microRNAs in colorectal cancer," *Journal of Genetics and Genomics*, vol. 37, no. 6, pp. 347–358, 2010.

[5] J. Krol, I. Loedige, and W. Filipowicz, "The widespread regulation of microRNA biogenesis, function and decay," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 597–610, 2010.

[6] T. M. Witkos, E. Koscianska, and W. J. Krzyzosiak, "Practical aspects of microRNA target prediction," *Current Molecular Medicine*, vol. 11, no. 2, pp. 93–109, 2011.

[7] N. Rajewsky, "microRNA target predictions in animals," *Nature Genetics*, vol. 38, supplement 1, pp. S8–S13, 2006.

[8] C. Zhao, C. Huang, T. Weng, X. Xiao, H. Ma, and L. Liu, "Computational prediction of MicroRNAs targeting GABA receptors and experimental verification of miR-181, miR-216 and miR-203 targets in GABA-A receptor," *BMC Research Notes*, vol. 5, no. 91, pp. 1–8, 2012.

[9] P. Lekprasert, M. Mayhew, and U. Ohler, "Assessing the utility of thermodynamic features for microRNA target prediction under relaxed seed and no conservation requirements," *PLoS ONE*, vol. 6, no. 6, Article ID e20622, 2011.

[10] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in Drosophila," *Genome Biology*, vol. 5, no. 1, p. R1, 2003.

[11] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[12] D. Grün, Y. Wang, D. Langenberger, K. C. Gunsalus, and N. Rajewsky, "MicroRNA target predictions across seven drosophilo species and comparison to mammalian targets," *PLoS Computational Biology*, vol. 1, no. 1, Article ID e13, 2005.

[13] M. Kiriakidou, P. T. Nelson, A. Kouranov et al., "A combined computational-experimental approach predicts human microRNA targets," *Genes and Development*, vol. 18, no. 10, pp. 1165–1178, 2004.

[14] A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen, "Identification of Drosophila microRNA targets," *PLoS Biology*, vol. 1, no. 3, Article ID E60, 2003.

[15] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, "Fast and effective prediction of microRNA/target duplexes," *RNA*, vol. 10, no. 10, pp. 1507–1517, 2004.

[16] M. Schnall-Levin, O. S. Rissland, W. K. Johnston, N. Perrimon, D. P. Bartel, and B. Berger, "Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs," *Genome Research*, vol. 21, no. 9, pp. 1395–1403, 2011.

[17] L. D. Hurst, "Preliminary assessment of the impact of microrna-mediated regulation on coding sequence evolution in mammals," *Journal of Molecular Evolution*, vol. 63, no. 2, pp. 174–182, 2006.

[18] X. Zhou, X. Duan, J. Qian, and F. Li, "Abundant conserved microRNA target sites in the 5′-untranslated region and coding sequence," *Genetica*, vol. 137, no. 2, pp. 159–164, 2009.

[19] A. S. Issabekova, O. A. Berillo, M. Regnier, and A. T. Ivashchenko, "Interactions of intergenic microRNAs with mRNAs of genes involved in carcinogenesis," *Bioinformation*, vol. 8, no. 11, pp. 513–518, 2012.

[20] S. Volinia, M. Galasso, M. E. Sana et al., "Breast cancer signatures for invasiveness and prognosis defined by deep sequencing

of microRNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 8, pp. 3024–3029, 2012.

[21] G. Kaur, A. Masoud, N. Raihan, M. Radzi, W. Khamizar, and L. S. Kam, "Mismatch repair genes expression defects and association with clinicopathological characteristics in colorectal carcinoma," *Indian Journal of Medical Research*, vol. 134, no. 2, pp. 186–192, 2011.

[22] T. Ripperger, C. Beger, N. Rahner et al., "Constitutional mismatch repair deficiency and childhood leukemia/lymphoma-report on a novel biallelic ?*MSH6* mutation," *Haematologica*, vol. 95, no. 5, pp. 841–844, 2010.

[23] A. Feber, L. Xi, J. D. Luketich et al., "MicroRNA expression profiles of esophageal cancer," *Journal of Thoracic and Cardiovascular Surgery*, vol. 135, no. 2, pp. 255–260, 2008.

[24] T. Sun, N. Aceto, K. L. Meerbrey et al., "Activation of multiple proto-oncogenic tyrosine kinases in breast cancer via loss of the *PTPN12* phosphatase," *Cell*, vol. 144, no. 5, pp. 703–718, 2011.

[25] A. K. Win, J. P. Young, N. M. Lindor et al., "Colorectal and other cancer risks for carriers and noncarriers from families with a DNA mismatch repair gene mutation: a prospective cohort study," *Journal of Clinical Oncology*, vol. 30, no. 9, pp. 958–964, 2012.

[26] M. Ollikainen, W. M. Abdel-Rahman, A. Moisio et al., "Molecular analysis of familial endometrial carcinoma: a manifestation of hereditary nonpolyposis colorectal cancer or a separate syndrome?" *Journal of Clinical Oncology*, vol. 23, no. 21, pp. 4609–4616, 2005.

[27] Y. Arima, H. Hayashi, M. Sasaki et al., "Induction of *ZEB* proteins by inactivation of RB protein is key determinant of mesenchymal phenotype of breast cancer," *Journal of Biological Chemistry*, vol. 287, no. 11, pp. 7896–7906, 2012.

[28] I. Lee, S. S. Ajay, I. Y. Jong et al., "New class of microRNA targets containing simultaneous 5′-UTR and 3′-UTR interaction sites," *Genome Research*, vol. 19, no. 7, pp. 1175–1183, 2009.

[29] A. S. Issabekova, O. A. Berillo, V. A. Khailenko, S. A. Atambayeva, M. Regnier, and A. T. Ivachshenko, "Characteristics of intronic and intergenic human miRNAs and features of their interaction with mRNA," *World Academy of Science, Engineering and Technology*, no. 59, pp. 63–66, 2011.

[30] S. Miura, M. Nozawa, and M. Nei, "Evolutionary changes of the target sites of two MicroRNAs encoded in the Hox gene cluster of Drosophila and other insect species," *Genome Biology and Evolution*, vol. 3, no. 1, pp. 129–139, 2011.

[31] M. Nozawa, S. Miura, and M. Nei, "Origins and evolution of microRNA genes in Drosophila species," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 180–189, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 307145, 7 pages
http://dx.doi.org/10.1155/2013/307145

Research Article

miR156- and miR171-Binding Sites in the Protein-Coding Sequences of Several Plant Genes

Assyl Bari, Saltanat Orazova, and Anatoliy Ivashchenko

Al-Farabi Kazakh National University, 71 Al-Farabi Avenue, Building No.6, Almaty 050038, Kazakhstan

Correspondence should be addressed to Anatoliy Ivashchenko; a_ivashchenko@mail.ru

Received 26 February 2013; Revised 20 May 2013; Accepted 23 June 2013

Academic Editor: Vassily Lyubetsky

We identified the interaction sites of several miRNAs with the mRNAs from paralogs and orthologs of the *SPL* and *HAM* genes in *A. thaliana*. miRNAs from the miR156 and miR157 families in *A. thaliana* are shown to have binding sites within the mRNAs of *SPL* genes. The ath-miR156a–j binding sites located in the mRNAs of the *SPL* paralogs contain the sequence GUGCUCUCUCUCUUCUGUCA. This sequence encodes the ALSLLS motif. miR157a–d bind to mRNAs of the *SPL* family at the same site. We suggest merging the miR156 and miR157 families into one family. Several *SPL* genes in eight plants contain conserved miR156 binding sites. GUGCUCUCUCUCUUCUGUCA polynucleotide is homologous in its binding sites. The ALSLLS hexapeptide is also conserved in the SPL proteins from these plants. Binding sites for ath-miR171a–c and ath-miR170 in *HAM1*, *HAM2*, and *HAM3* paralog mRNAs are located in the CDSs. The conserved miRNA binding sequence GAUAUUGGCGCGGCUCAAUCA encodes the ILARLN hexapeptide. Nucleotides within the *HAM1*, *HAM2*, and *HAM3* miRNA binding sites are conserved in the mRNAs of 37 orthologs from 13 plants. The miR171- and miR170-binding sites within the ortholog mRNAs were conserved and encode the ILARLN motif. We suggest that the ath-miR170 and ath-miR171a–c families should be in one family.

1. Introduction

Individual microRNAs (miRNAs) and their families can be identical or very similar in closely related and phylogenetically distant plant species [1, 2]. Therefore, it is important to determine the properties of miRNA-binding sites in the protein-coding sequences (CDSs) of paralogous and orthologous genes. In plants, miRNAs regulate the expression of many genes involved in plant morphogenesis and development [3–7] and resistance to biotic and abiotic stresses [2, 8–10]. The number of identified plant miRNAs is growing, and the main challenge is to clearly identify their targets. Many miRNA-binding site prediction programs such as miRanda (http://www.microrna.org/microrna/getMirnaForm.do) [11], DIANA microT (http://diana.cslab.ece.ntua.gr/DianaTools/index.php?r=microtv4/index) [12], and PicTar (http://pictar.mdc-berlin.de/) [13] search for the sites in the 3′-untranslated region (3′UTR) of the mRNA. However, in plant and animal cells, miRNA-binding sites have been identified in the 5′-untranslated region (5′UTR) and CDS [14–18]. Computational methods can predict many miRNA-binding sites in mRNAs. However, a significant proportion of false-positive miRNA-binding sites are identified. Therefore, it is necessary to develop methods of improving the reliability of site prediction. One way to improve the reliability of binding site prediction is to check if the sites are present in orthologous genes. The aim of our research is to identify the interaction sites of several miRNAs within the CDSs of paralogous and orthologous mRNAs and establish the features of these interactions. *SPL* and *HAM* genes code for transcription factors and play a key role in the regulation of plant reproductive development [19, 20]. The expression of these genes is controlled by miRNAs. In this paper, we present the characteristics of the binding sites for the miR156, miR157, miR170, and miR171 families in several paralogous and orthologous *SPL* and *HAM* genes in *A. thaliana* and other plant species.

2. Materials and Methods

The gene sequences from *Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Glycine max, Medicago truncatula, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Ricinus communis, Selaginella moellendorffii, Sorghum bicolor, Vitis vinifera,* and *Zea mays* were obtained from GenBank (http://www.ncbi.nlm.nih.gov/). The miRNAs sequences were retrieved from miRBase (http://www.mirbase.org/). The free energy (ΔG) of hybridization between miRNAs and mRNAs, the position of potential binding sites, and the interaction schemes were calculated using the RNAHybrid 2.1 program (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) [21]. The E-RNAhybrid program (http://sites.google.com/site/malaheenee/software/) was used to compute the $\Delta G/\Delta G_m$ value and P value. The $\Delta G/\Delta G_m$ value was used as a comparative criterion for the miRNA and mRNA interaction force. A $\Delta G/\Delta G_m$ value of more than 75% indicates a significant degree of complementarity between the miRNA and its target. This percentage corresponds to $P < 0.005$. The ΔG value and its standard deviation were used to determine the validity of predicted miRNA-binding sites in the mRNA. The maximal interaction energy (ΔG_m) for miR156, miR157, miR170, miR171, and their families was equal to the binding energy of perfectly complementary sequences. Graphs of the nucleotide and amino acid sequence variability were created by using the WebLogo program (http://weblogo.berkeley.edu/) [22]. To improve the reliability of predicted miRNA-binding sites in the mRNAs of genes in *A. thaliana*, we confirmed their presence in the mRNAs of orthologous genes in other plants.

3. Results

3.1. Binding of the miR156 and miR157 Families with the mRNAs of SPL Paralogs. We found that, among 328 miRNAs in *A. thaliana,* the miR156 and miR157 families are shown to have strong binding sites within the squamosa promoter binding protein-like (*SPL*) gene family of transcription factors. The miR156 family consists of 10 miRNAs (miR156a–j), and miR156a–f have identical nucleotide sequences (miRBase). miR156 family members are predicted to be associated with the mRNAs of genes encoding the DNA-binding proteins SPL1–SPL16 with varying degrees of prediction reliability. Among the 16 genes in this family, CDSs of eight paralogs have been targeted by miR156a: *AT1G27360 (SPL11), AT1G27370 (SPL10), AT1G69170 (SPL6), AT2G42200 (SPL9), AT3G57920 (SPL15), AT5G43270 (SPL2), AT5G50570 (SPL13)*, and *AT5G50670 (SPL13)*. The genes *AT5G50570* and *AT5G50670* have significant conservation and are located at a distance of 33 kilobases (kb) from each other. Therefore, we only studied the properties of the miRNA-binding sites within the *AT5G50570* gene. The miR156a-binding sites within the *SPL* mRNAs are identical and consist of the conserved nucleotides GUGCUCUCUCUCUUCUGUCA. The open reading frame encoding the conserved ALSLLS motif begins with the GCU triplet of the miRNA-binding site sequence.

Table 1 shows the nucleotide sequences of the miR156a-binding sites in the mRNAs of eight *SPL* genes and the amino acid sequences of the corresponding paralogous SPL proteins containing the ALSLLS oligopeptide. The first two nucleotides (GU) of the oligonucleotide are involved in miRNA-binding; the second and third positions are part of the nonconserved codon (NGU) of paralogous genes and, therefore, the corresponding amino acids are variable (Figure 1). This variability may indicate the importance of the GU dinucleotide in enhancing the binding of miR156a with the mRNAs.

The $\Delta G/\Delta G_m$ value for the miR156a-binding sites ranged from 90.2% to 91.4%, which indicates a strong interaction between this miRNA and the mRNAs of the *SPL* gene family (Table 2). The sequences of miR156g, miR156h, miR156i, and miR156j differ from the miR156a–f sequence by one or two nucleotides; however, miR156a–j bind to the same site within each of the *SPL* paralog mRNAs, which is specific for the miR156 family. The $\Delta G/\Delta G_m$ value varied from 88.4% for miR156g to 100% for miR156j (Table 2). Thus, the binding is strong in all cases.

The miR157a–c sequences differ from the miR156a sequence by one nucleotide at the 5′-end. The miR157d sequence differs by one nucleotide from the miR156h sequence and by two nucleotides from miR156a sequence. miR157a–d bind to the mRNAs of the *SPL* family at the same site as miR156a–j (Table 2). The $\Delta G/\Delta G_m$ value of the miR157a–d-binding sites ranged from 89.5% to 92.5% (Table 2). Therefore, we suggest that the miRNAs of the miR156 and miR157 families belong to the same family. For example, *Oryza sativa* and *Zea mays* have only the miR156 family and not the miR157 family (miRBase).

The *SPL3* and *SPL5* mRNA-binding sites for miR156a are located in the 3′UTR, and their nucleotide sequences have significant conservation with those in the CDS. The $\Delta G/\Delta G_m$ value for these binding sites was 93.1% and 84.7%, respectively.

Table 3 represents the interaction schemes for the miR156 and miR157 families with the mRNAs from *AT1G27360*, *AT1G27370*, *AT1G69170*, *AT2G42200*, *AT3G57920*, *AT5G43270*, and *AT5G50570* paralogs. The position of the binding sites for various miRNAs differs in paralogous genes. Thus, the binding of the miRNA with the mRNA occurs over the entire nucleotide sequence, not only within the "seed" region.

3.2. Binding of the miR156 Family with the mRNAs of SPL Orthologs. SPL genes from *A. lyrata, O. sativa, Populus trichocarpa, Physcomitrella patens, Ricinus communis, Sorghum bicolor, Vitis vinifera,* and *Z. mays* are targeted by the miR156 family. For all of the studied genes, the GUGCUCUCUCUCUUCUGUCA polynucleotide is completely conserved in the ath-miR156a-binding sites within the mRNA (Figure 2(a)). Consequently, the ALSLLS hexapeptide is also conserved in the SPL proteins of these plant species (Figure 2(b)). High variability of the nucleotide sequences adjacent to the binding site and therefore variability of the amino acids before and after the ALSLLS motif were observed (Figures 2(a) and 2(b)).

Table 1: Nucleotide variability of miR156a binding sites in mRNA of *SPL* paralogous genes and amino acid variability of SPL paralogous proteins in regions with the ALSLLS oligopeptide in *A. thaliana*.

Gene	Region of mRNA	Region of protein
AT1G27360	CACC$\boxed{\text{GUGCUCUCUCUCUUCUGUCA}}$ACC	SQDIHR$\boxed{\text{ALSLLS}}$TSSDP
AT1G27370	U..A......................U	.H.FYS.......T..S
AT1G69170	ACU.....................G.U	ASRST.......AQ.QQ
AT2G42200	A..U.....................AU	IG.SNC......NPHQ.
AT3G57920	AG.U.....................A.	.T.SSC......N.YPI
AT5G43270	G.UG.....................AU	...LDG......N.TTW
AT5G50570	G.UU....................U..	IH.SDC......S..SS
AT5G50670	G.UU....................U..	IH.SDC......S..SS

The conservative sequence is set in box.

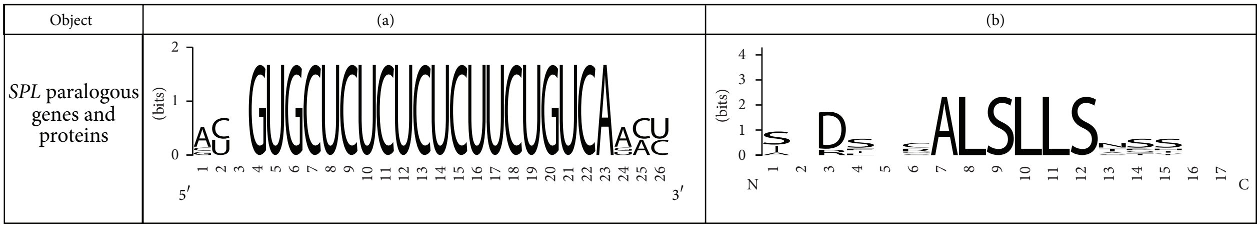


Figure 1: The variability of nucleotides (a) in the binding sites of miR156a with mRNA of *SPL* paralogs and the variability of amino acids (b) in SPL proteins containing the ALSLLS hexapeptide in *A. thaliana*.

The mRNAs of the *SPL3* orthologous genes in *A. lyrata* and *P. trichocarpa* have miR156a-binding sites in the 3′UTR. The nucleotide sequences of these sites differ slightly, and the $\Delta G/\Delta G_m$ value was equal to 93.1% and 87.3%, respectively. The *SPL5* mRNA from *A. thaliana* and *V. vinifera* also has miR156a-binding sites in the 3′UTR. The $\Delta G/\Delta G_m$ value for these sites was equal to 84.7% and 91.4%, respectively. The *SPL5* mRNAs from *P. patens* and *Z. mays* have miR156a-interaction sites within the CDS. The $\Delta G/\Delta G_m$ value for these binding sites was equal to 90.9% and 90.7%, respectively. We suggest that the location of the binding sites in the CDS and 3′UTR of *SPL3* and *SPL5* orthologs may change because of their close position to the border between the CDS and 3′UTR.

3.3. Binding of miR171a–c and miR170 with the mRNAs of HAM Paralogs.

We have determined the binding sites for ath-miR171a within the mRNAs of *AT4G00150 (HAM3)*, *AT2G45160 (HAM1)*, and *AT3G60630 (HAM2)*. These genes belong to the GRAS transcription factor family in *A. thaliana* [20]. The binding sites for ath-miR171a in the CDS contain the GAUAUUGGCGCGGCUCAAUCA polynucleotide, which encodes the ILARLN hexapeptide in the corresponding proteins (Table 4).

The genes *HAM1*, *HAM2*, and *HAM3* are targets for ath-miR171b, ath-miR171c, and ath-miR170. The characteristics of the ath-miR171a–c and ath-miR170 interaction sites with the mRNAs of *HAM1, HAM2, and HAM3* paralogs are listed in Table 5. All of these miRNA-binding sites are located in the CDS and have the same position for each gene.

ath-miR171a binds perfectly with the mRNAs of *HAM* paralogs, and the $\Delta G/\Delta G_m$ value was equal to 100%. ath-miR171b and c connect without the triplet at the 3′end; therefore, the $\Delta G/\Delta G_m$ value was equal to 86.9%, which indicates a strong interaction with these RNAs. Although ath-miR170 belongs to another family, it strongly associates with each paralog of the *HAM* genes. The $\Delta G/\Delta G_m$ value was 98.8%. Binding sites for ath-miR171a–c and ath-miR170 are in the CDS and their positions are the same in each paralog mRNA (Table 5). The conserved GAUAUUGGCGCG-GCUCAAUCA oligopeptide encodes the conserved ILARLN hexapeptide in *HAM1*, *HAM2*, and *HAM3* paralogs (Table 4). A comparison of the predicted binding sites for ath-miR171a–c and ath-miR170 in *HAM1*, *HAM2*, and *HAM3* suggests that these miRNAs belong to the same family. We grouped these miRNAs into the ath-miR171 family because ath-miR171a has full complementarity with all of the sites in the mRNAs of *HAM1*, *HAM2*, and *HAM3* paralogs.

The nucleotide sequences of miR171a–c and miR170 form structures of complementary nucleotides with the binding sites within the mRNAs of *HAM* paralogs (Table 6). The position of the binding sites for various miRNAs differs in paralogous genes. This indicates a strong miRNA interaction with the mRNA. We noted that, in the center of these structures, there are eight GC pairs that make the main contribution to the interaction energy for the mRNA and miRNA pair. This is inconsistent with the concept in which the main contribution to the binding is made by "seeds" located at the 5′-end of the miRNA [23–25].

Table 2: Characteristics of miR156a–j and miR157a–d binding sites in CDSs of *SPL* paralogous genes in *A. thaliana*.

Gene	Position in CDS, nt	$\Delta G/\Delta G_m$, %						
		miR156a–f	miR156g	miR156h	miR156i	miR156j	miR157a–c	miR157d
AT1G27360	1211	91.4	88.6	93.9	99.3	100	91.7	92.5
AT1G27370	2365	91.1	88.4	93.6	99.3	100	91.4	92.5
AT1G69170	1295	91.4	88.6	93.9	99.3	100	91.0	92.5
AT2G42200	936	90.7	87.9	93.1	99.3	100	91.7	91.7
AT3G57920	844	90.7	87.9	93.1	99.3	100	91.7	91.7
AT5G43270	1186	90.7	87.9	93.1	99.3	100	91.7	91.7
AT5G50570	1100	90.2	87.9	92.6	98.8	100	89.5	91.2

Table 3: Schemes of miR156a–j and miR157a–d binding sites in the CDSs of *SPL* paralogous genes in *A. thaliana*.

<table>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCA 3′</td></tr>
<tr><td></td><td>|||||| |||||||||||||</td></tr>
<tr><td>miR156a–f</td><td>3′ CACGAGUGAGAGAAGACAGU 5′</td></tr>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCA 3′</td></tr>
<tr><td></td><td>|||||| |||||||||||||</td></tr>
<tr><td>miR156g</td><td>3′ CACGAGUGAGAGAAGACAGU 5′</td></tr>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCA 3′</td></tr>
<tr><td></td><td>||||||||| ||||||||||</td></tr>
<tr><td>miR156h</td><td>3′ CACGAGAGAAAGAAGACAGU 5′</td></tr>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCA 3′</td></tr>
<tr><td></td><td>||||||||||||||||||||</td></tr>
<tr><td>miR156i</td><td>3′ GACGAGAGAGAGAAGACAGU 5′</td></tr>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCA 3′</td></tr>
<tr><td></td><td>||||||||||||| |||||||</td></tr>
<tr><td>miR156j</td><td>3′ CACGAGAGAGAGAAGACAGU 5′</td></tr>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCAA 3′</td></tr>
<tr><td></td><td>||||||||| |||||||||||</td></tr>
<tr><td>miR157a–c</td><td>3′ CACGAGAGAUAGAAGACAGUU 5′</td></tr>
<tr><td><i>SPL</i> mRNA</td><td>5′ GUGCUCUCUCUCUUCUGUCA 3′</td></tr>
<tr><td></td><td>||||||||| ||||||||||</td></tr>
<tr><td>miR157d</td><td>3′ CACGAGAGAUAGAAGACAGU 5′</td></tr>
</table>

3.4. Binding of miR171a–c and miR170 with the mRNAs of HAM Orthologs. For each of the *HAM1*, *HAM2*, and *HAM3* paralogs, orthologs were found in 13 species *(A. lyrata, A. thaliana, Brachypodium distachyon, Glycine max, Medicago truncatula, O. sativa, P. patens, P. trichocarpa, R. communis, S. moellendorffii, S. bicolor, V. vinifera*, and *Z. mays*) and miR171a-binding sites were identified in these genes. The nucleotide sequences of the binding sites are highly conserved in the mRNAs of 37 orthologous genes (See Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2013/307145). All of the nucleotides in the miR171a-binding site (GAUAUUGGCGCGGCUCAAUCA) were conserved and encode the same ILARLN motif (Supplemental Table S2). Nucleotides adjacent to the ath-miR171a-binding sites in the mRNA and amino acids near the ILARLN motif in the HAM1, HAM2, and HAM3 orthologous proteins were variable. Therefore, conservation of the miR171a-binding sites is more important for the regulation of proteins than that of the adjacent nucleotides in the mRNA and the corresponding amino acids in the proteins. The amino acids located upstream and downstream of the ILARLN motif are variable.

4. Discussion

In the study of miRNAs, there are many challenges. It is important to identify the targets for a particular miRNA and to determine the degree of miRNA binding to its target. The results from previous studies are contradictory. In particular, it has been suggested that the miRNA binds preferentially to the 3′UTR of the mRNA and that the binding is determined by the "seed" at the 5′-end of the miRNA [23–25]. However, miRNAs were shown to bind to the 5′UTR and CDS [26–28]. Recently, these results were further supported by a number of publications [14–17, 29]. It was shown that approximately 70% and 80% of miRNA-binding sites are located in the 5′UTR and CDS of mRNAs in animal and plant genes, respectively. Our results show that miRNA-binding sites can be located in the CDS of mRNAs and that the sites are highly conserved in the evolution of higher plants. There may be a reason for the localization of miRNA-binding sites in the CDS. These sites will be more conserved than those in the variable 5′UTR and 3′UTR mRNA regions. Localization of the binding sites in the CDS contributes to earlier miRNA-binding than localization in 3′UTR during the posttranscriptional regulation of gene expression. Our results reveal that such links could have been established a long time ago and are highly stable. The results were obtained with a high probability prediction of the miRNA-binding sites in the mRNAs of *SPL* and *HAM* genes. The occurrence of these interactions and their conservation in many plants show the fundamental role of gene expression regulation by miRNAs. The degree of miRNA binding to the mRNA is an indicator of the regulatory role of the miRNA in gene expression.

The absence of binding sites for a miRNA in mRNAs of some paralogs suggests that there is a way to protect part of

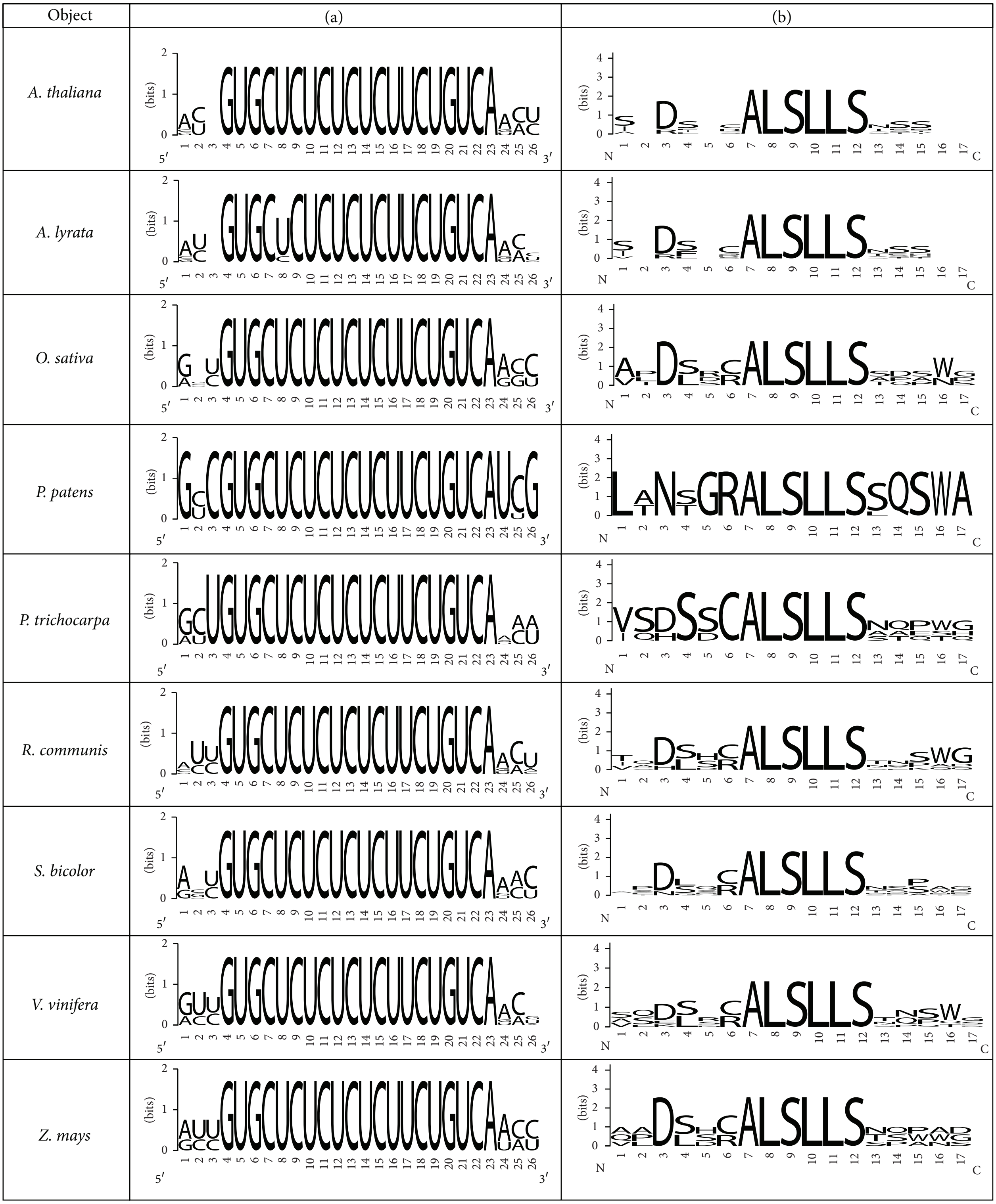


Figure 2: The variability of nucleotides (a) in the binding sites of miR156a with mRNAs of *SPL* orthologs and the variability of amino acids (b) in SPL proteins that contain the ALSLLS hexapeptide.

the gene family expression from the effect of that miRNA. For example, about half of the *SPL* family genes are targets for the miR156 family. In animal cells, only the *PTPN12* gene out of the 16 *PTPN* paralogs strongly binds to has-miR-1279. Moreover, this relationship was highly conserved during evolution (unpublished data). Among the eight genes of the *MSH* family, only the mRNA of the *MSH6* gene has the binding site for has-miR-1279, and this binding site is preserved in

Table 4: Nucleotide variability of miR171a binding sites in mRNA of *HAM* paralogous genes and amino acid variability of HAM paralogous proteins in regions with the ILARLN oligopeptide in *A. thaliana*.

Genes	Region of mRNA	Region of protein
AT4G00150	UCAGGG$\boxed{\text{GAUAUUGGCGCGGCUCAAUCA}}$ACAG	TCLAQG$\boxed{\text{ILARLN}}$QQLSSPV
AT3G60630	G..A.......................CA.C	PV..........HN.NNNN
AT2G45160	G..A.......................C..U	TV..........HH.NTSS

The conservative sequence is set in box.

Table 5: Characteristics of miR171a–c and ath-miR170 binding sites in mRNAs of *HAM* paralogous genes in *A. thaliana*.

Gene	Position in CDS, nt	$\Delta G/\Delta G_m$, %		
		miR171a	miR171b,c	miR170
AT2G45160	884	100	86.9	98.8
AT3G60630	803	100	86.9	98.8
AT4G00150	674	100	86.9	99.5

Table 6: Schemes of miR171a–c and miR170 binding sites in CDSs of *HAM* paralogous genes in *A. thaliana*.

<table>
<tr><td><i>HAM</i> mRNA</td><td>5′</td><td>GAUAUUGGCGCGGCUCAAUCA</td><td>3′</td></tr>
<tr><td></td><td></td><td>|||||||||||||||||||||</td><td></td></tr>
<tr><td>miR171a</td><td>3′</td><td>CUAUAACCGCGCCGAGUUAGU</td><td>5′</td></tr>
<tr><td><i>HAM</i> mRNA</td><td>5′</td><td>AGGGAUAUUGGCGCGGCUCAA</td><td>3′</td></tr>
<tr><td></td><td></td><td> ||||||||||||||||||</td><td></td></tr>
<tr><td>miR171b,c</td><td>3′</td><td>CGACUAUAACCGUGCCGAGUU</td><td>5′</td></tr>
<tr><td><i>HAM</i> mRNA</td><td>5′</td><td>GAUAUUGGCGCGGCUCAAUCA</td><td>3′</td></tr>
<tr><td></td><td></td><td>|||||||||||||||||||||</td><td></td></tr>
<tr><td>miR170</td><td>3′</td><td>CUAUAACUGUGCCGAGUUAGU</td><td>5′</td></tr>
</table>

the orthologs of the *MSH6* gene. Such selective binding of an individual miRNA with mRNAs of other families of protein-coding genes is possible (unpublished data).

In this paper, we show that miRNA-binding sites with identical nucleotide sequences may be located in the CDS and 3′UTR of various genes, for example, *SPL3* and *SPL5* orthologs. However, a binding site in the 3′UTR is less conserved than that in the CDS. A similar result was obtained earlier with the *ZNF* gene family in animal cells [16].

Analysis of interaction schemes of the miRNA with the mRNA in Tables 3 and 4 shows that there is no clear preference for a particular part of the miRNA in binding to the mRNA. These data show a high degree of conservation of miRNA nucleotide sequences during evolution [30, 31].

According to the miRBase database, ath-miR156a–j and ath-miR157a–d belong to different families. However, all of these miRNA families have a common binding site in the mRNAs of the *SPL* gene family. It is likely that ath-miR156a–j and ath-miR157a–d are members of the same family. Similar arguments support the conclusion that ath-miR170 and ath-miR171a–c should be combined into the same family. The previously mentioned results should be considered for the distribution of miRNA in different families.

Acknowledgment

This work was supported by The Ministry of Education and Science of Kazakhstan (Grant no. 0143/GF).

References

[1] J. T. Cuperus, N. Fahlgren, and J. C. Carrington, “Evolution and functional diversification of *MIRNA* genes,” *Plant Cell*, vol. 23, no. 2, pp. 431–442, 2011.

[2] Z. Tang, L. Zhang, C. Xu et al., “Uncovering small RNA-mediated responses to cold stress in a wheat thermosensitive genic male-sterile line by deep sequencing,” *Plant Physiology*, vol. 159, no. 2, pp. 721–738, 2012.

[3] G. Wu and R. S. Poethig, “Temporal regulation of shoot development in *Arabidopsis thaliana* by miRr156 and its target SPL3,” *Development*, vol. 133, no. 18, pp. 3539–3547, 2006.

[4] A. Yamaguchi, M.-F. Wu, L. Yang, G. Wu, R. S. Poethig, and D. Wagner, “The MicroRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of *LEAFY, FRUITFULL, and APETALA1*,” *Developmental Cell*, vol. 17, no. 2, pp. 268–278, 2009.

[5] S. Schwarz, A. V. Grande, N. Bujdoso, H. Saedler, and P. Huijser, “The microRNA regulated SBP-box genes SPL9 and SPL15 control shoot maturation in Arabidopsis,” *Plant Molecular Biology*, vol. 67, no. 1-2, pp. 183–195, 2008.

[6] S. H. Cho, C. Coruh, and M. J. Axtella, “miR156 and miR390 regulate tasiRNA accumulation and developmental timing in *Physcomitrella patens*,” *Plant Cell*, vol. 24, no. 12, pp. 4837–4849, 2012.

[7] U. Chorostecki, V. A. Crosa, A. F. Lodeyro et al., “Identification of new microRNA-regulated genes by conserved targeting in plant species,” *Nucleic Acids Research*, vol. 40, no. 18, pp. 8893–8904, 2012.

[8] Y. Meng, C. Shao, H. Wang, and M. Chen, “The regulatory activities of plant microRNAs: a more dynamic perspective,” *Plant Physiology*, vol. 157, no. 4, pp. 1583–1595, 2011.

[9] T. Hewezi, T. R. Maier, D. Nettleton, and T. J. Baum, “The arabidopsis microrna396-GRF1/GRF3 regulatory module acts as a developmental regulator in the reprogramming of root cells during cyst nematode infection,” *Plant Physiology*, vol. 159, no. 1, pp. 321–335, 2012.

[10] J. J. Kim, J. H. Lee, W. Kim, H. S. Jung, P. Huijser, and J. H. Ahn, “The microrNA156-SQUAMOSA promoter binding protein-like3 module regulates ambient temperature-responsive flowering via flowering locus in Arabidopsis,” *Plant Physiology*, vol. 159, no. 1, pp. 461–478, 2012.

[11] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, “Human microRNA targets,” *Plos Biology*, vol. 2, no. 11, article e363, 2004.

[12] M. Kiriakidou, P. T. Nelson, A. Kouranov et al., “A combined computational-experimental approach predicts human microRNA targets,” *Genes and Development*, vol. 18, no. 10, pp. 1165–1178, 2004.

[13] A. Krek, D. Grün, M. N. Poy et al., “Combinatorial microRNA target predictions,” *Nature Genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[14] F. Grey, R. Tirabassi, H. Meyers et al., “A viral microRNA downregulates multiple cell cycle genes through mRNA 5’UTRs,” *Plos pathogens*, vol. 6, no. 6, article e1000967, 2010.

[15] F. Moretti, R. Thermann, and M. W. Hentze, “Mechanism of translational regulation by miR-2 from sites in the 5′ untranslated region or the open reading frame,” *RNA*, vol. 16, no. 12, pp. 2493–2502, 2010.

[16] M. Schnall-Levin, O. S. Rissland, W. K. Johnston, N. Perrimon, D. P. Bartel, and B. Berger, “Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs,” *Genome Research*, vol. 21, no. 9, pp. 1395–1403, 2011.

[17] Y. Wang, A. Itaya, X. Zhong et al., “Function and evolution of a microRNA that regulates a caspi^{2+} -ATPase and triggers the formation of phased small interfering rnas in tomato reproductive Growth,” *Plant Cell*, vol. 23, no. 9, pp. 3185–3203, 2011.

[18] L. da Sacco and A. Masotti, “Recent insights and novel bioinformatics tools to understand the role of microRNAs binding to 5′ untranslated region,” *International Journal of Molecular Sciences*, vol. 14, no. 1, pp. 480–495, 2013.

[19] G. Wu, M. Y. Park, S. R. Conway, J.-W. Wang, D. Weigel, and R. S. Poethig, “The sequential action of miR156 and miR172 regulates developmental timing in arabidopsis,” *Cell*, vol. 138, no. 4, pp. 750–759, 2009.

[20] E. M. Engstrom, C. M. Andersen, J. Gumulak-Smith et al., “Arabidopsis homologs of the petunia HAIRY MERISTEM gene are required for maintenance of shoot and root indeterminacy,” *Plant Physiology*, vol. 155, no. 2, pp. 735–750, 2011.

[21] J. Krüger and M. Rehmsmeier, “RNAhybrid: MicroRNA target prediction easy, fast and flexible,” *Nucleic Acids Research*, vol. 34, pp. W451–W454, 2006.

[22] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, “WebLogo: a sequence logo generator,” *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[23] L. P. Lim, N. C. Lau, P. Garrett-Engele et al., “Microarray analysis shows that some microRNAs downregulate large numbers oftarget mRNAs,” *Nature*, vol. 433, no. 7027, pp. 769–773, 2005.

[24] K. K.-H. Farh, A. Grimson, C. Jan et al., “Biochemistry: the widespread impact of mammalian microRNAs on mRNA repression and evolution,” *Science*, vol. 310, no. 5755, pp. 1817–1821, 2005.

[25] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets,” *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[26] U. A. Ørom, F. C. Nielsen, and A. H. Lund, “MicroRNA-10a binds the 5′UTR of ribosomal protein mRNAs and enhances their translation,” *Molecular Cell*, vol. 30, no. 4, pp. 460–471, 2008.

[27] I. Lee, S. S. Ajay, I. Y. Jong et al., “New class of microRNA targets containing simultaneous 5/-UTR and 3/-UTR interaction sites,” *Genome Research*, vol. 19, no. 7, pp. 1175–1183, 2009.

[28] X. Zhou, X. Duan, J. Qian, and F. Li, “Abundant conserved microRNA target sites in the 5/-untranslated region and coding sequence,” *Genetica*, vol. 137, no. 2, pp. 159–164, 2009.

[29] A. S. Issabekova, O. A. Berillo, M. Regnier, and A. T. Ivashchenko, “Interactions of intergenic microRNAs with mRNAs of genes involved in carcinogenesis,” *Bioinformation*, vol. 8, no. 11, pp. 513–518, 2012.

[30] S. Miura, M. Nozawa, and M. Nei, “Evolutionary changes of the target sites of two MicroRNAs encoded in the Hox gene cluster of Drosophila and other insect species,” *Genome Biology and Evolution*, vol. 3, no. 1, pp. 129–139, 2011.

[31] M. Nozawa, S. Miura, and M. Nei, “Origins and evolution of MicroRNA genes in plant species,” *Genome Biology and Evolution*, vol. 4, no. 3, pp. 230–239, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 621604, 14 pages
http://dx.doi.org/10.1155/2013/621604

Research Article

Evaluating Phylogenetic Informativeness as a Predictor of Phylogenetic Signal for Metazoan, Fungal, and Mammalian Phylogenomic Data Sets

Francesc López-Giráldez,[1] Andrew H. Moeller,[1] and Jeffrey P. Townsend[1,2,3]

[1] *Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520, USA*
[2] *Department of Biostatistics, Yale University, 135 College Street, New Haven, CT 06520, USA*
[3] *Program in Computational Biology and Bioinformatics, 300 George Street, Yale University, New Haven, CT 06520, USA*

Correspondence should be addressed to Jeffrey P. Townsend; jeffrey.townsend@yale.edu

Received 1 March 2013; Accepted 3 June 2013

Academic Editor: William Piel

Phylogenetic research is often stymied by selection of a marker that leads to poor phylogenetic resolution despite considerable cost and effort. Profiles of phylogenetic informativeness provide a quantitative measure for prioritizing gene sampling to resolve branching order in a particular epoch. To evaluate the utility of these profiles, we analyzed phylogenomic data sets from metazoans, fungi, and mammals, thus encompassing diverse time scales and taxonomic groups. We also evaluated the utility of profiles created based on simulated data sets. We found that genes selected via their informativeness dramatically outperformed haphazard sampling of markers. Furthermore, our analyses demonstrate that the original phylogenetic informativeness method can be extended to trees with more than four taxa. Thus, although the method currently predicts phylogenetic signal without specifically accounting for the misleading effects of stochastic noise, it is robust to the effects of homoplasy. The phylogenetic informativeness rankings obtained will allow other researchers to select advantageous genes for future studies within these clades, maximizing return on effort and investment. Genes identified might also yield efficient experimental designs for phylogenetic inference for many sister clades and outgroup taxa that are closely related to the diverse groups of organisms analyzed.

1. Introduction

The genomes of nearly 400 eukaryotes and nearly 3000 prokaryotes are now or are in the process of being sequenced. Most of these organisms have thousands of genes, yet only a few of those have been commonly used as markers for phylogenetic studies [1]. In cases where choice has been exercised, genes have been selected for sequencing based on rough impressions of the genes' utilities in previous studies of taxa that are to varying degrees divergent from the taxa of interest. Recent sequencing of multiple genomes within major branches of the tree of life provides a much greater selection of markers and excites hope that more accurate procedures for experimental design may be adopted. However, despite the phenomenal growth in sequence information available, the optimal way to employ genome-wide data sets to inform more clade-specific molecular phylogenetic studies remains elusive [2, 3] due to a lack of methods that quantitatively assess the power of genes to resolve particular nodes in a phylogeny.

Although a few rules of thumb for selecting genes for phylogenetic studies have been advocated (e.g., percent sequence divergence; 4, 5; or proportion of parsimony-informative sites; 6), their successful use is highly context dependent. Conventional wisdom dictates selection of a gene that evolves at an appropriate pace for the phylogenetic question of interest, but this axiom often fails to illuminate the correct decision. Fairly complex distributions of rates across characters can yield information regarding some periods of the history encompassed by the phylogeny but not others [4–6].

In response to the crucial role of gene selection in experimental design [4, 7–11], Townsend [5] proposed a metric that predicts utility across historical time for known genes. Based on the estimated full distribution of rates across characters, the Townsend [5] informativeness yields a graphical appraisal of a gene's signal for any historical epoch. To estimate informativeness, prior data on the molecular evolutionary rates for each site of a locus is required. This prior information may be derived from three potential sources: (1) preliminary data on the candidate loci from a well-studied subset of the taxa of interest; (2) data on the candidate loci from a well-studied sister clade; or (3) comparative genomic data from sequenced genomes within and/or outside the clade of interest. Thus, Townsend [5] informativeness metrics can be obtained without reference to sequence data from the taxa of interest.

To evaluate phylogenetic informativeness as a procedure for selecting loci to sequence for phylogenetic studies that incorporate broad taxon sampling, we utilized empirical data sets for which the process of evolution may only be approximated, as well as simulated data for which we could specify the process and the true tree. In each case, we tested the performance of the Townsend [5] phylogenetic informativeness with trees with more than four taxa and its robustness to the effects of homoplasy. We analyzed data sets encompassing different time scales and taxonomic groups: (i) mammal data sets [12] consisting of 50 genes each (~33,440 aa and ~100,649 bp) sequenced in 20 species; (ii) a fungal data set [13] consisting of 46 genes (~13,082 aa) sequenced in 28 species; (iii) an animal/fungal data set [14, 15] consisting of 50 genes (~12,089 aa) sequenced in 25 species. In parallel, we simulated 50 amino acid and 50 DNA sequence alignments of 300 sites each. Genes for empirical and simulated data sets were ranked by phylogenetic informativeness and analyzed for their ability to recapitulate known node identity and robustness using measures of branch support associated with maximum likelihood and maximum parsimony optimality criteria. Genes from the empirical data sets examined here that were identified as performing well could be especially useful for phylogenetic inference in organisms related to the clades analyzed. Our results represent the first phylogenomic test of the phylogenetic informativeness method presented by Townsend [5], supporting it as a metric of potential phylogenetic signal in both nucleotide and amino acid data sets.

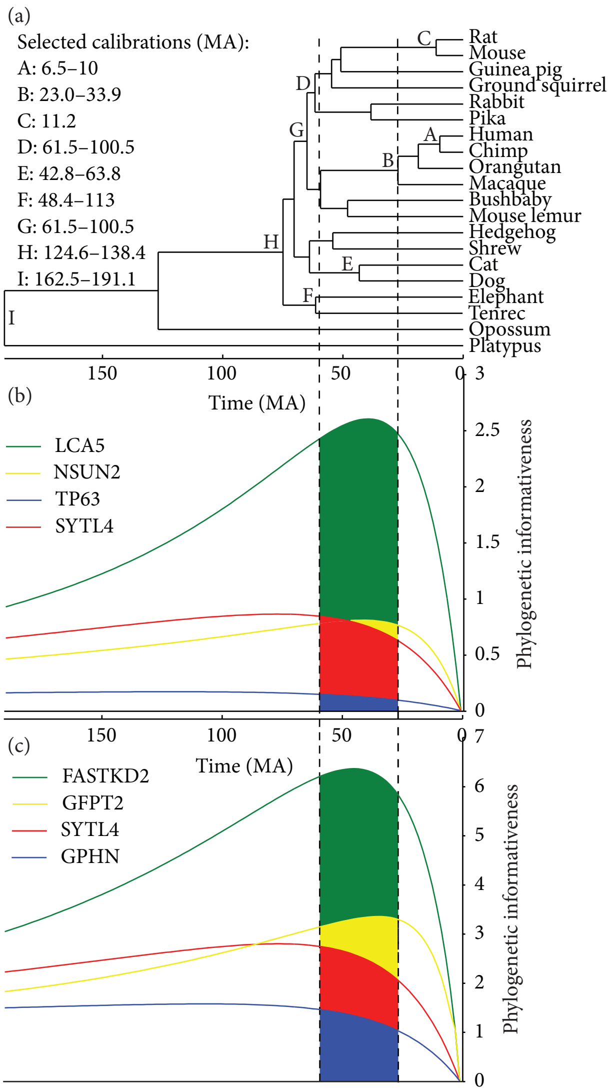


Figure 1: Phylogenetic informativeness profiles for the OrthoMaM data set. (a) Chronogram and the calibration points used to calculate site rates. Phylogenetic informativeness profiles over a 190 Myr period for (b) four amino acid sequence alignments and (c) four DNA sequence alignments, on the same time scale as in panel (a). Integration of the area below the profiles can provide a ranking of the predicted utility of genes for that epoch (here, the epoch encompassing the branch leading to primates). Integration results will be the largest for the genes that have the highest probability of exhibiting mutations during the given epoch that will not be obscured in subsequent branches. To quantitatively establish genes that will be most informative for the entire phylogeny, integrals over the whole time scale were calculated.

2. Materials and Methods

2.1. Sequence Data Composition. We obtained four amino acid and two DNA sequence data sets for analysis. In all cases, species were selected whose phylogeny has been well established. Data sets encompassed different time scales and taxonomic levels, and were extracted from four different sources (Table 1). From the OrthoMaM database [12] of orthologous genomic markers for placental mammals, we obtained a data set of amino acid and a data set of nucleotide sequences. We selected 50 single-copy orthologous genes that were present in 20 species (Figure 1(a)) and that had lengths of ~2000 bp and ~666 aa (see Supplementary Tables 1 and 2 in Supplementary Material available online at http://dx.doi.org/10.1155/2013/621604). The large number of genes in OrthoMaM facilitated a gene size constraint to minimize the influence of sequence length in the phylogenetic inference. Forty genes were in common between the amino acid and DNA data sets. Northern tree shrew, cow, horse, little brown bat, and nine-banded armadillo species were also present in OrthoMaM database but were excluded from the analyses to ensure a comparison of outcomes

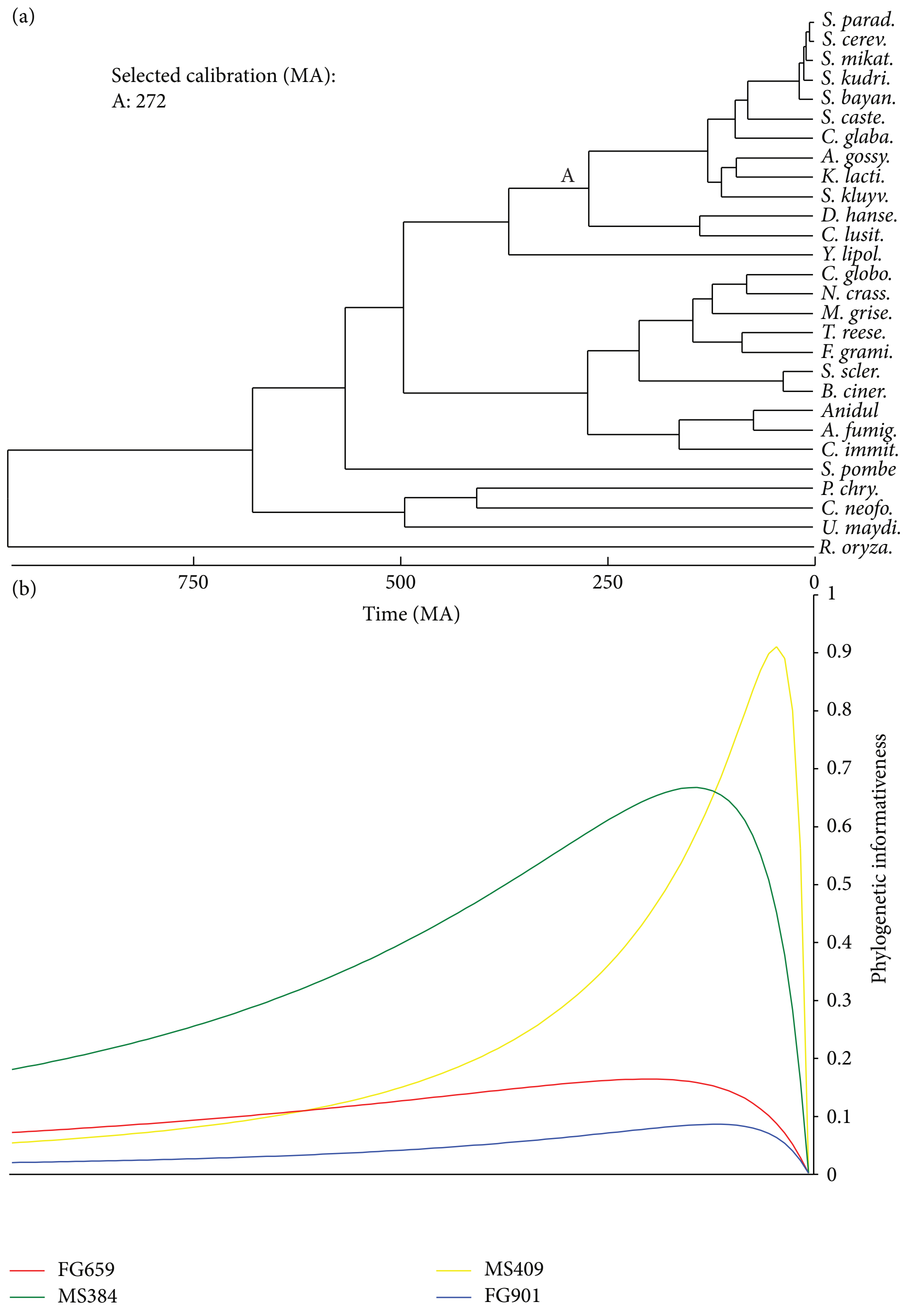


Figure 2: Phylogenetic informativeness profiles for the FUNYBASE data set. (a) Chronogram and the calibration points used to calculate site rates. (b) Phylogenetic informativeness profiles over a 973 Myr period for four amino acid sequence alignments, on the same time scale as in panel (a).

against a well-resolved and uncontroversial phylogenetic tree. From the FUNYBASE database [13] of fungal orthologous sequences, we obtained amino acid sequences for 46 genes (Supplementary Table 3) in 28 fungal species (Figure 2(a)). *Stagonospora nodorum* and *Aspergillus oryzae* were also present in FUNYBASE but they were excluded due to their weakly supported phylogenetic placement [16]. The third source was Taylor and Berbee [15], modified from Rokas et al. [14]—abbreviated as the TBR data set—from which we obtained an alignment of 50 amino acid sequence regions (Supplementary Table 4) from 8 animal, 15 fungal, and 2 plant species (Figure 3(a)).

The two sets of simulated alignments, amino acid and nucleotide, were generated with Seq-Gen v1.3.2 software [17]. Simulated alignments allowed us to fix the alignment length, the evolutionary model, and its parameters. To ensure that the simulations encompassed a realistic instance, the parameter values for mean site rate, proportion of invariant sites, and sequence length approximated the values found in the FUNYBASE data set. Also, both amino acid and DNA

Table 1: Data sets used in the study.

Data set	Source	Gene number	Mean length ± SD
OrthoMaM (AA)	Ranwez et al. (2007) [12]	50	668.8 ± 13.3
OrthoMaM (DNA)	Ranwez et al. (2007) [12]	50	2013.0 ± 43.5
FUNYBASE (AA)	Marthey et al. (2008) [13]	46	284.4 ± 130.0
TBR (AA)	Taylor and Berbee (2006) [15]	50	241.3 ± 106.3
Simulations (AA)	Seq-Gen	50	300 ± 0
Simulations (DNA)	Seq-Gen	50	300 ± 0

sequences were simulated on the FUNYBASE chronogram (see below). The gamma rate heterogeneity values were varied to explore a wide range. JTT and K2P (with $\kappa = 2$) models were used for amino acid and nucleotide alignments, respectively. For both amino acid and DNA, 50 different data sets were simulated, each based on one of 10 different mean rates ranging from 0.0001 to 0.001 substitutions per site per Myr and ranging across five gamma rate heterogeneity values, including no rate heterogeneity and $\alpha = 0.5, 1, 2$, and 3. In all cases, the gamma distribution was discretized into 10 rate categories. For every alignment, 20% of sites were set to be invariant. For each of these 50 alignments, we generated 10 replicates. The Seq-Gen program assigns each site to either the invariant category or one of the gamma categories stochastically within each simulation. As a result, the number of invariant sites and the number of sites in each category vary from simulation to simulation. However, this fact would not guarantee that each rate category has the same number of sites. To ensure that each replicated alignment had the same exact rate distribution but not the same amino acid or DNA sequence, replicates were created such that each replicate contained 60 invariant sites and 24 sites for each of the 10 rate categories if rate heterogeneity was specified.

The sequences downloaded from OrthoMaM and FUNYBASE were aligned using MUSCLE v3.6 [18] with default settings. Gblocks v0.91b [19] was used to remove ambiguously aligned positions from the alignments. In Gblocks, the minimum number of sequences for a flank position was set to 16. Only sites with more than half of sequences with gaps were treated as a gap position and eliminated from the final alignment. Default settings were applied for the rest of options.

2.2. Divergence Times and Chronogram. To compute the rates of evolution of amino acid and nucleotide sites for all nonsimulated data sets, we specified an ultrametric evolutionary tree. The concatenated amino acid sequences were used in each case to estimate the phylogeny with the parallel version of MrBayes v3.1.2 [20, 21]. The length of the concatenated sequences totaled 16,802, 13,082, and 12,089 aa for OrthoMaM, FUNYBASE, and TBR alignments, respectively. We allowed mixed models with invariant sites and gamma-shaped rate variation with four rate categories. All parameters were unlinked; thus, the models and parameters were estimated during the analysis separately for each locus. Ten independent runs were conducted using 4 MCMC chains and random starting trees of 500,000 generations each, sampling trees every 100 generations. We discarded the first 100,000 generations as burn-in after visualization in the program Tracer v1.4 [22], long after the log likelihood reached apparent stationarity.

For convenience, we used a time-calibrated phylogeny (chronogram). While absolute dates of internal nodes were not relevant to any inferences herein, their relative depths were aligned with the ultrametric profiles for predictive purposes. We obtained the chronogram for each data set (see Figures 1(a)–3(a)) by passing the phylogenetic tree with the highest likelihood to r8s software v1.71 [23]. This software allows incorporation of multiple calibration points, fixing or constraining minimal or maximal ages to the nodes. For the OrthoMaM chronogram (Figure 1(a)), we used diverse calibration points from [24]. For the FUNYBASE chronogram (Figure 2(a)), the tree was calibrated by fixing the split of *D. hansenii* and *C. albicans* from the other yeasts at 272 Myr [25]. For the TBR chronogram (Figure 3(a)), three calibration points were used as in the intermediate solution in [15]. Divergence times were estimated by the penalized likelihood method with a truncated Newton algorithm in r8s, setting the smoothing parameter to 0.06 for OrthoMaM and 0.01 for FUNYBASE trees. The optimization of the smoothing parameter was obtained using the cross-validation feature in r8s following the instructions of the program manual (available at http://loco.biosci.arizona.edu/r8s/). As indicated in [15], there was an absence of predictable lineage-specific rate correlations in the TBR tree. Thus, these data were processed following the recommendation of the r8s program documentation, with the Langley-Fitch method that assumes a global substitution rate instead of the penalized likelihood method.

2.3. Evolutionary Rates and Phylogenetic Informativeness. Using the alignment data and its corresponding chronogram, molecular evolutionary rates were estimated for each gene at each alignment position. We used Rate4site [26] and DNArates (Olsen, unpublished) programs to obtain the substitution rates at amino acid and nucleotide sites, respectively. These programs were chosen because they provide ML approach to estimate the rates for each site independently and a simple model to avoid overparameterization. In the Rate4site program, rates were inferred assuming a JTT model for the topology and branch lengths of the input phylogenetic tree without any optimization. In the DNArates program, rates were inferred assuming K2P (with $\kappa = 2$) model.

For each gene, the phylogenetic informativeness profile ρ as a function of time, T, was calculated as

$$\rho\left(T; \lambda_1, \ldots, \lambda_2\right) = \sum_{i=1}^{n} 16\lambda_i^2 T e^{-4T\lambda_i}, \tag{1}$$

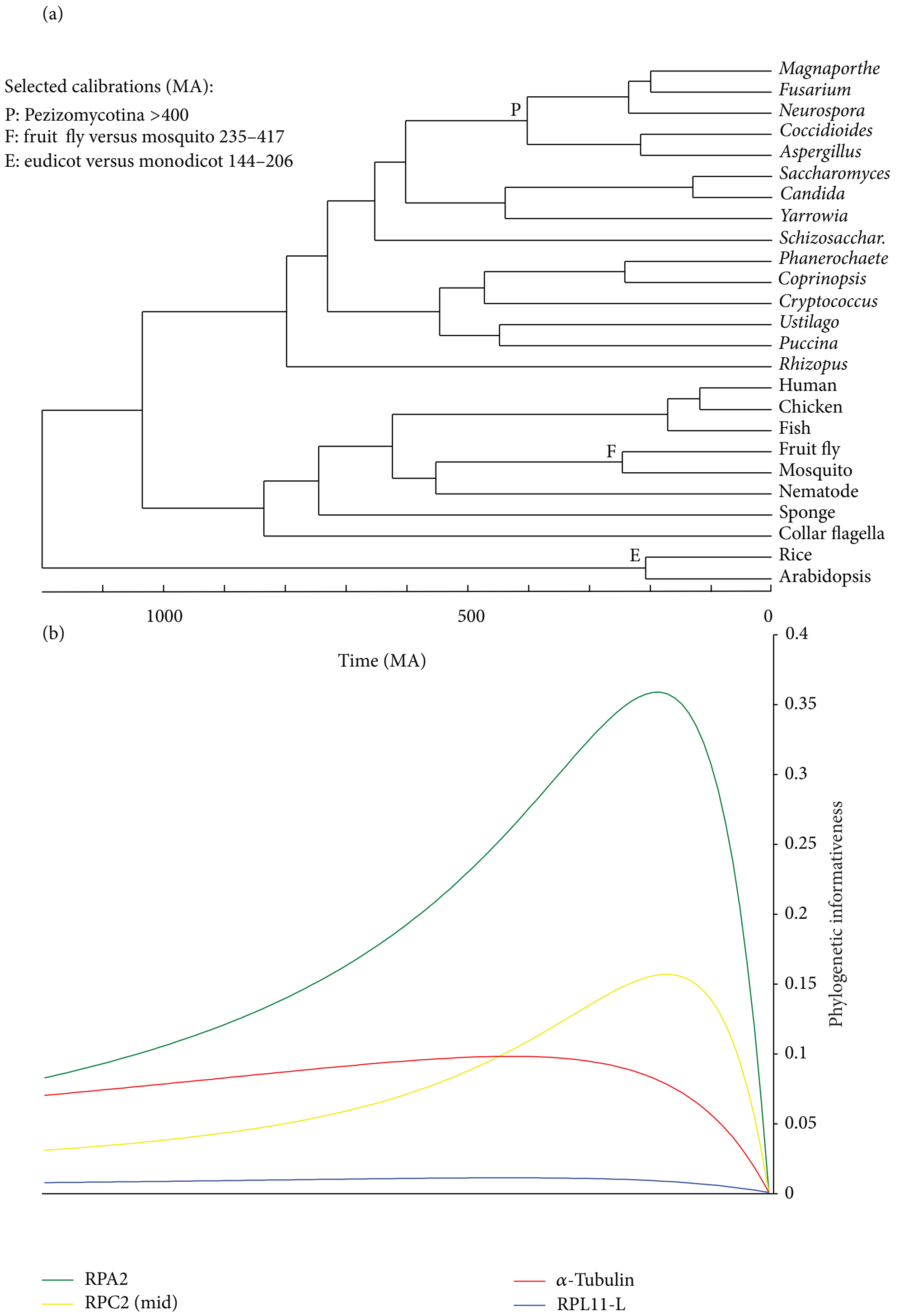


Figure 3: Phylogenetic informativeness profiles for the TBR data set. (a) Chronogram and the calibration points used to calculate site rates. (b) Phylogenetic informativeness profiles over a 1193 Myr period for four amino acid sequence alignments, on the same time scale as in panel (a).

substituting the estimated rates λ_i of evolution of each site [5]. This formula provides the probability that character i would provide an unambiguous synapomorphy lying within an asymptotically short internode between two pairs of sister taxa whose common ancestor is at time T. To convey the informativeness of a particular data set, the equation was plotted at a continuum of depths, from time 0 to the root of the phylogenetic trees (Figures 1–4). The differential phylogenetic informativeness (DPI) of each gene was evaluated quantitatively by integrating over the phylogenetic

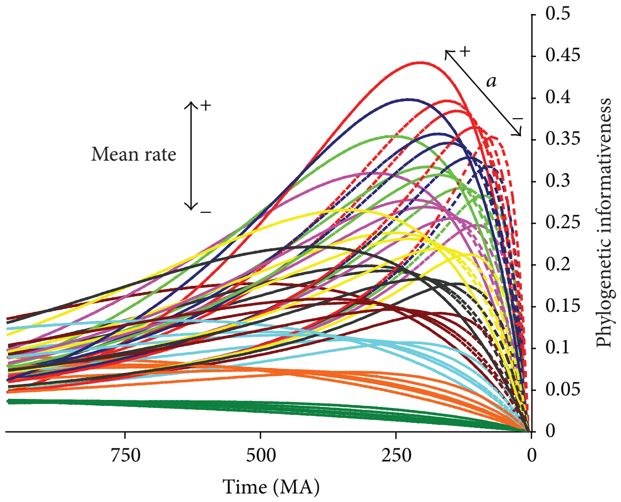


Figure 4: Phylogenetic informativeness profiles for simulated amino acid and DNA alignments, on the same time scale as in Figure 1 (973 Myr period). Each of the 10 different colors represents a different mean rate, from 0.0001 (slowest, bottom) to 0.001 (fastest, top) substitutions per site per million years. Dashed lines are profiles from alignments simulated with gamma rate heterogeneity (α = 0.5, 1, 2, and 3).

informativeness profile from the origin (h_1) to the terminus (h) of the epochs of interest, $\int_{h1}^{h2} \rho(T;\lambda)T$. Assigning h_1 and h_2 so as to encompass all branching points of a phylogeny provided a summary of the relative informativeness of each gene to resolve all nodes in the phylogeny. Using DPI, we ranked the genes for each data set.

Both the calculations of the molecular evolutionary rate and of the phylogenetic informativeness profiles were performed using the PhyDesign web application [27].

2.4. Phylogenetic Analysis. To evaluate the performance of each locus, we analyzed the accuracy and robustness with which each locus recovered the well-established topology. In each case, the phylogeny resulting from our analysis of the concatenated amino acid alignments exactly matched established topology as described in the literature [12, 16, 28, 29]. To assess the fidelity with which individual genes recovered the reference tree on a holistic scale, we calculated the topological distance between each individual gene tree and the concatenated tree using the Symmetric Difference [30] computed by the TreeDist program included in the Phylip v3.68 package [31]. To measure the robustness with which each gene recovered the correct topology on a node-by-node basis, we applied four metrics of branch support based on two different optimization criteria, maximum likelihood (ML) and maximum parsimony (MP). The ML support metrics were nonparametric bootstrap (ML-BP) and the approximate Likelihood-Ratio Test statistic (aLRT statistic; [32]). The MP support metrics were nonparametric bootstrap (MP-BP) and Decay Index (DI) [33].

For all individual aligned orthologous markers, we determined the amino acid and nucleotide substitution models that best fit the data using the command-line mode of ProtTest v1.4 [34] and ModelGenerator v0.85 [35], respectively. In both programs, the models for each gene were selected to minimize the BIC criterion.

We inferred ML gene trees with PHYML v2.4.4 [36] using bootstrap proportions (BP) based on 100 bootstrap replicates, with the model and parameters as described above. For empirical data sets, we discretized the gamma site-rate distribution into four rate categories. For simulated sequences, we discretized the gamma site-rate distribution into 10 categories, both for inference and for simulation. Ten instead of four rate categories were used in the simulated data to obtain more diverse sequences and substitution rates during the simulation process. We used PAUP* v4.0b10 [37] for MP-BP analyses based on 1000 BP replicates. A heuristic search with TBR branch-swapping on 20 random sequence addition replicate starting trees was employed.

Although BP values are probably the most frequently used type of support values, they scale nonlinearly with the number of synapomorphies, conveying little information when they are low and reaching an asymptote of 100 rapidly when they are high. In contrast, aLRT and DI are not constrained by an upper limit. The aLRT is based on the conventional LRT under the null hypothesis that the inferred branch has length 0 [32]. Our analysis, applied the aLRT statistic value—that is, two times the difference between the maximum log-likelihood values of the best and the second best alternative arrangements around the branch of interest—with the modified version of PHYML v2.4.5 [32]. The last support measure, DI, also known as Bremer support, is the number of parsimony steps from the best tree to the next best tree without the branch of interest. DIs were calculated with the help of AutoDecay v5.04 for PERL using reverse constraints in PAUP*.

2.5. Statistical Analysis. We used three different statistical approaches to evaluate the performance of phylogenetic informativeness. First, we correlated the DPI gene ranking with the tree distance from the gene tree to the well-established tree topology. We expected that the genes ranked highest would show a low tree distance—that is, recovering a topology closest to the reference tree. Second, we measured how well phylogenetic informativeness predicts branch support on a node-by-node basis. To do so, genes were compared in pairs based on their predicted performance (DPI), comparing the predicted best with the predicted worst, the predicted second best with the predicted second worst, and so on. Then, we correlated their predicted proportionate performance:

$$\text{PPP} = \frac{\text{DPI}_1}{\text{DPI}_1 + \text{DPI}_2}, \tag{2}$$

where $\text{DPI}_i > \text{DPI}_j$ and $0.5 \le \text{PPP} \le 1$. With n_i denoting the number of nodes with higher support using gene i, and n_j denoting the number of nodes with higher support using gene j, the empirical proportionate performance of each gene pair in terms of nodes better supported was calculated as

$$\mathrm{EPP} = \frac{n1}{n1 + n2}, \quad (3)$$

where $0 \leq \mathrm{EPP} \leq 1$. To measure the strength of linear relationship between predicted and empirical performances, we calculated Pearson's correlation coefficient (r).

Third, we examined the cumulative global support for the well-established phylogeny as each gene was added according to several sampling schemes. To do so, we first calculated the proportionate likelihood-ratio support (PLRS) and the proportionate decay index support (PIDS) provided by each gene for each node in the well-established phylogeny. For each node, we divided the aLRT statistic and DI by the number of genes. Then, we calculated the average PLRS and PIDS across nodes for each gene. This average value can be interpreted as the relative contribution or global support of each gene for the well-established phylogeny. We plotted the cumulative path of the global support for each data set according to several sampling schemes. In an ideal experiment, this cumulative support would dramatically rise with the top-ranked prioritized loci and increase little as less informative markers were used.

Sampling the genes from the highest to the lowest proportionate support would represent the ideal situation when deciding about sampling genes for sequencing. Logically, the other way around represents the worst-case scenario. We also plotted the hypothetical average path between these two extremes. Finally, we compared these paths with the plot when prioritizing sampling with DPI values.

All alignment operations, data parsing, and communication of data to and from software were performed with Perl programming including Bioperl modules [38]. We also manipulated trees using Phyutility v2.2 [39]. All software used for the analyses mentioned corresponded to the Linux version. Only results from the ML analyses (i.e., ML-BP and aLRT) are shown. Relevant differences between ML and MP results are discussed in the text.

3. Results

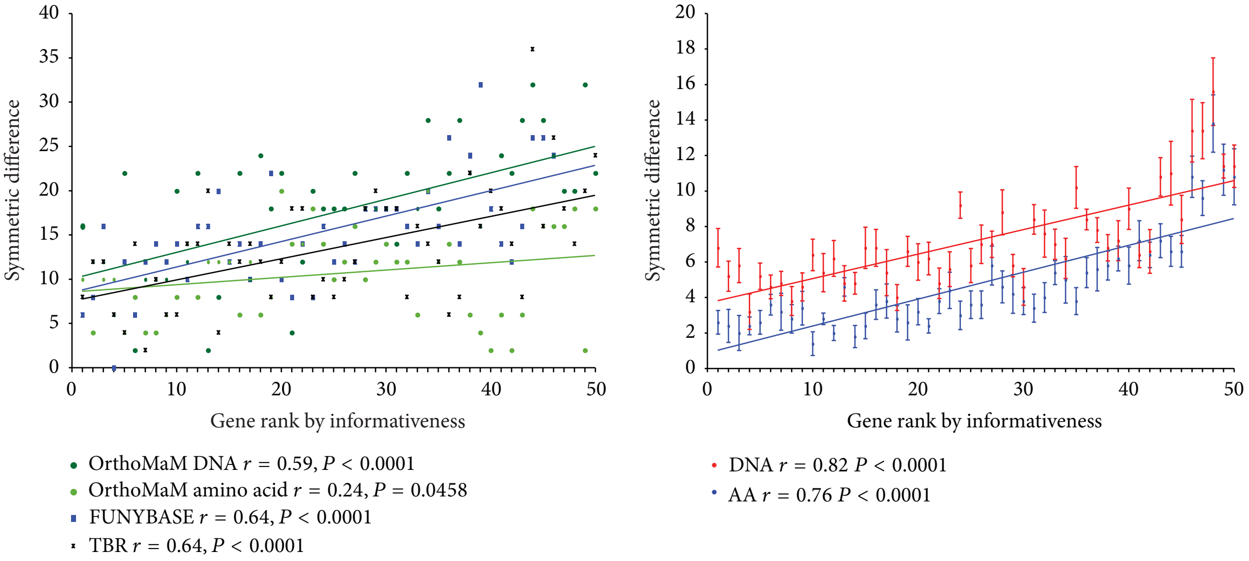
3.1. Phylogenetic Informativeness Profiles. Graphical profiles of the phylogenetic informativeness for four loci scaled to match with the ultrametric trees (Figures 1–3) illustrated the great diversity of levels of informativeness among genes in all data sets. Plotted genes were chosen to provide extremal exemplars with different performances: the best and worst genes across the whole time scale and two other genes which showed most variation in recent compared with ancient times and vice versa. The phylogenetic informativeness profiles for the rest of the genes lie approximately within the range of the extremal profiles. The OrthoMaM data set illustrates this variation in informativeness well. Although the genes TP63 and LCA5 shared approximately the same number of sites (680 and 682 aa, resp.), LCA5 exhibited greater informativeness over the whole tree (Figure 1(b)).

Compared to SYTL4 (2046 bp; Figure 1(c)), GFPT2 exhibited higher informativeness in recent times but lower informativeness for more ancient times and yet was composed of about the same number of sites (~2000 bp). The effects of variation of rates across sites on phylogenetic informativeness profiles were also observed in the simulated sequences (Figure 4). At the same global rate, the higher the gamma-shape parameter (alpha), the closer the profiles to a singular rate distribution. When alpha was low, corresponding to higher rate heterogeneity, the profiles peaked closer to recent times due to the presence of a set of faster evolving sites.

Direct comparison between amino acid and DNA informativeness profiles for 40 genes for which amino acid and DNA were both extracted from OrthoMaM data sets demonstrated correlated patterns of informativeness, with two significant differences. First, amino acid profiles showed more variation from low to high informativeness potential. Second, DNA alignments had higher profiles than amino acid sequences (e.g., see SYTL4 Figures 1(b) and 1(c)), mainly due to their threefold greater number of sites. However, comparing per site profiles (data not shown), the differences in informativeness disappeared or in some cases even became inverted. Lower and flatter amino acid profiles were still present, probably due to silent substitutions. Silent substitutions, which mostly occur in the third position of a codon and have no effect upon the amino acid sequence, can cause a higher rate of evolution of nucleotide sites without affecting the rate of evolution of amino acid sites. Thus, silent substitutions in genes with flat amino acid profiles caused by the lack of sequence variation can produce higher DNA profiles.

3.2. Phylogenetic Informativeness as Predictor of Node Identity and Branch Support. All DPI rankings significantly positively correlated with the ability to recover the correct topology (Figure 5). The OrthoMaM DNA data set exhibited the weakest ($r = .24$) and simulated amino acid and DNA data sets exhibited the strongest ($r = .76$ and $r = .82$, resp.) correlations. Other indices of informativeness such as gene length, number of variable sites, and number of parsimony informative sites were also tested for correlation with the symmetric differences (data not shown). All these other indices did not exhibit significant correlations except for the FUNYBASE and for the amino acid and DNA simulated data sets (all three correlations for these three data sets were significant, $P < .05$). The analyses using MP yielded strikingly similar results (data not shown).

Phylogenetic informativeness as predictor of ML-BP yielded significant correlations for all data sets (Figure 6). The TBR data set exhibited the weakest ($r = .34$) and the simulated amino acid alignments the strongest ($r = .94$) correlation. Generally, stronger correlations with accuracy and robustness (Figures 5 and 6) were revealed in the simulated data sets than in the empirical data. Other summary statistics for genes such as gene length, number of variable sites, and number of parsimony informative sites also correlated with performance. Locus length was significantly correlated only with ML-BP in the FUNYBASE data set ($n = 23$, $r = .65$, $P < .001$). In addition, in all data sets with the exception of the DNA OrthoMaM data set, ML-BP exhibited



(a)

(b)

Figure 5: Phylogenetic accuracy. Symmetric-difference tree distance as a function of the gene ranking based on phylogenetic informativeness for (a) OrthoMaM amino acid (dark green circles), OrthoMaM DNA (light green circles), FUNYBASE (blue squares), and TBR (black crosses) data sets and (b) simulated amino acid (blue) and DNA (red) data. The symmetric difference was calculated comparing each gene tree estimated by ML to the well-established tree as obtained from the concatenated amino acid sequences. The corresponding Pearson's correlation coefficient (r) and P value for each data set are indicated. For the simulated data, the mean and its standard error were plotted using the replicates.

significant correlations ($P < .05$) with the number of variable and parsimony informative sites. For the DNA OrthoMaM data set, the only measure that was significantly correlated with node identity and branch support was the phylogenetic informativeness. To ensure that this result did not arise as a peculiar consequence of our length-based selection of genes, we evaluated these correlations with an additional, nonoverlapping subset of OrthoMaM DNA sequences with a lower mean number of sites (~1000 bp). The phylogenetic informativeness was again the only measure significantly correlated to both node identity ($n = 50$, $r = .44$, $P < .001$) and branch support ($n = 25$, $r = .59$, $P < .001$). MP analyses yielded similar results. For MP, the DNA simulated alignments exhibited the weakest ($n = 25$, $r = .44$, $P = .013$) and the FUNYBASE the strongest ($n = 25$, $r = .79$, $P < .0001$) correlations.

DPI rankings provided close to the optimal experimental design path (i.e., sampling the genes from the highest to the lowest average PLRS, which would represent the ideal situation when deciding about sampling genes for sequencing) for all amino acid data sets (Figure 7), outperforming haphazard sampling, especially for FUNYBASE and simulated amino acid data. DPI rankings for the TBR data set showed more deviation from the ideal sampling. Prioritizing with DPI rankings for both OrthoMaM and simulated DNA data (Figure 8) also yielded results close to optimal experimental design path, outperforming the haphazard path. The best and worst experimental design paths for DNA data sets showed less difference from each other than did their respective amino acid counterparts. This result is consistent with the highest variability in amino acid informativeness potential mentioned earlier when comparing overlapping genes for amino acid and DNA OrthoMaM data sets. A similar trend was observed in both simulated and empirical data. Empirical data showed higher variability in the performance of individual genes. We also repeated cumulative plots ranking genes based on the other summary statistics mentioned previously: gene length, number of variable sites, and number of parsimony informative sites. For all data sets with the exception of DNA OrthoMaM, these rankings performed better than haphazard sampling. For the DNA OrthoMaM data set, the phylogenetic informativeness ranking performed noticeably better than these other measures, which in some parts of the plot crossed or were below the average path. The analyses using MP yielded similar results.

4. Discussion

We systematically examined the Townsend [5] phylogenetic informativeness as a metric for assessing phylogenetic signal. Our results demonstrate that prioritizing rankings obtained with phylogenetic informativeness was significantly correlated with the ability of a gene to recover the right topology as well as with higher branch support measures. In addition, we found that the informativeness metrics significantly outperformed haphazard experimental design and predicted a close-to-optimal prioritization of gene sequencing. Although Townsend [5] phylogenetic informativeness was based on analysis of the canonical four-taxon problem and although it does not specifically account for the misleading effects of

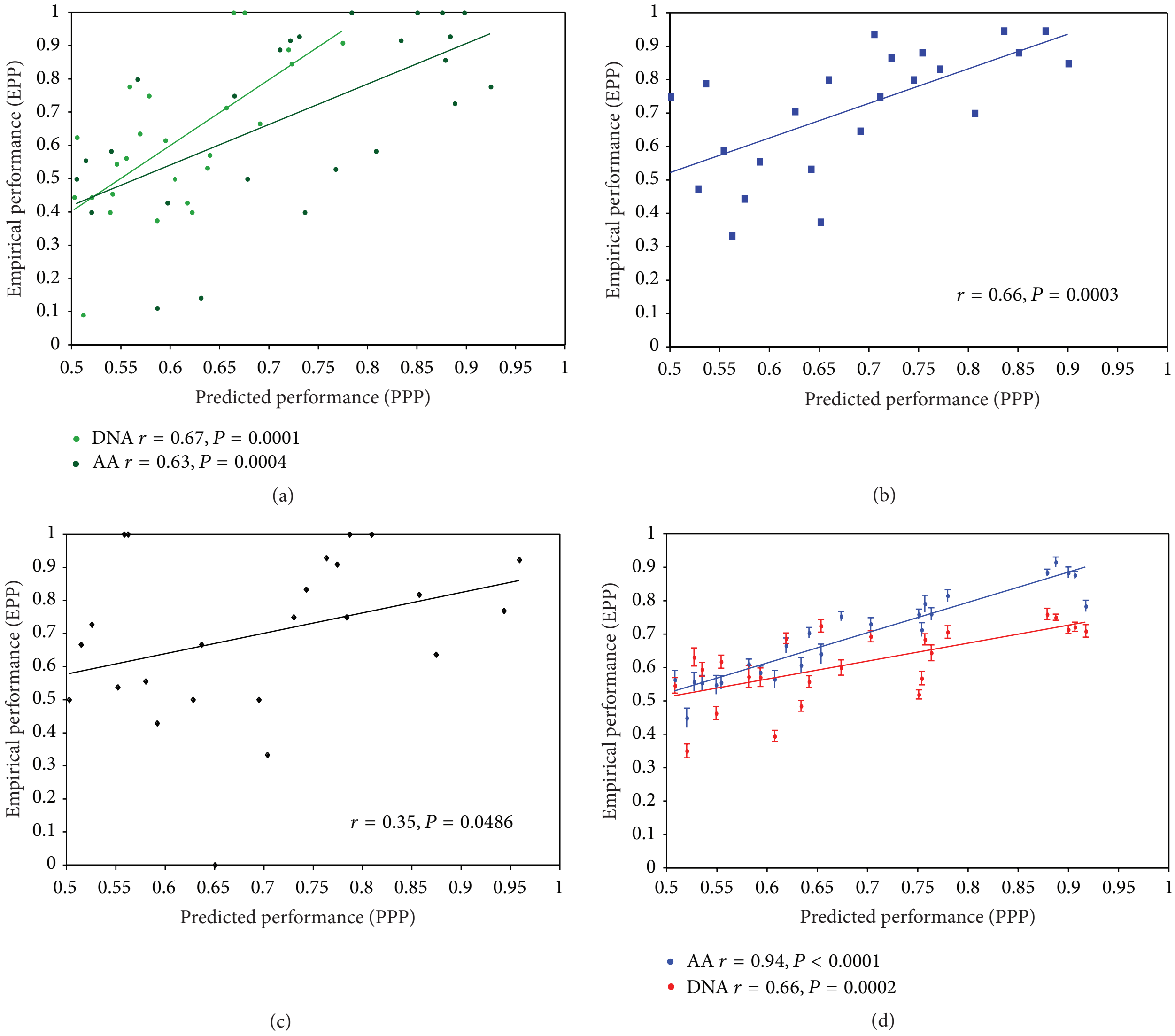


Figure 6: Phylogenetic informativeness as a predictor of maximum likelihood bootstraps (ML-BPs). Variation in ML-BP support (EPP) as a function of the variation in phylogenetic informativeness (PPP; see Material and Methods for more details) for (a) OrthoMaM amino acid (dark green), OrthoMaM DNA (light green), (b) FUNYBASE (blue), (c) TBR (black), and (d) simulated amino acid (blue) and DNA (red) data sets. The corresponding Pearson's correlation coefficient (r) and P value for each data set are indicated. For the simulated data, the mean and its standard error were plotted using the replicates.

homoplasy, our analysis suggests that the metric is robust despite these limitations. We examined its predictions in phylogenomic data sets spanning diverse time scales and taxonomic groups for both amino acid and DNA sequences and supplemented our empirical analyses with controlled simulations. Furthermore, we validated the results for both parsimony and maximum likelihood optimality criteria. We conclude that phylogenetic informativeness profiles provide advantageous guidance for phylogenetic projects in the selection and prioritization of loci to sequence for maximal return on effort and investment.

Despite its crucial role, pursuit of analytical methods for experimental design in phylogenetics has been sparse. Until recently, the only prominent procedure developed to deal with the question of experimental design in the context of topological uncertainty has been the empirical saturation plot [40]. In this plot, a lack of more or less increasing linear sequence divergence with time would indicate saturation in the set of characters analyzed. However, the plots are hard to fit unambiguously to data and do not lend themselves to immediate quantifications of informativeness for specific epochs. Other graphical methods to visualize phylogenetic signal have been advanced, such as likelihood mapping [41] and plotting Treeness triangles [42]. Although well suited for post hoc analyses, a major issue with these graphical approaches is that they are not easy to interpret or very practical for large-scale surveys. More importantly, neither puts forward an applied methodology for ranking genes for

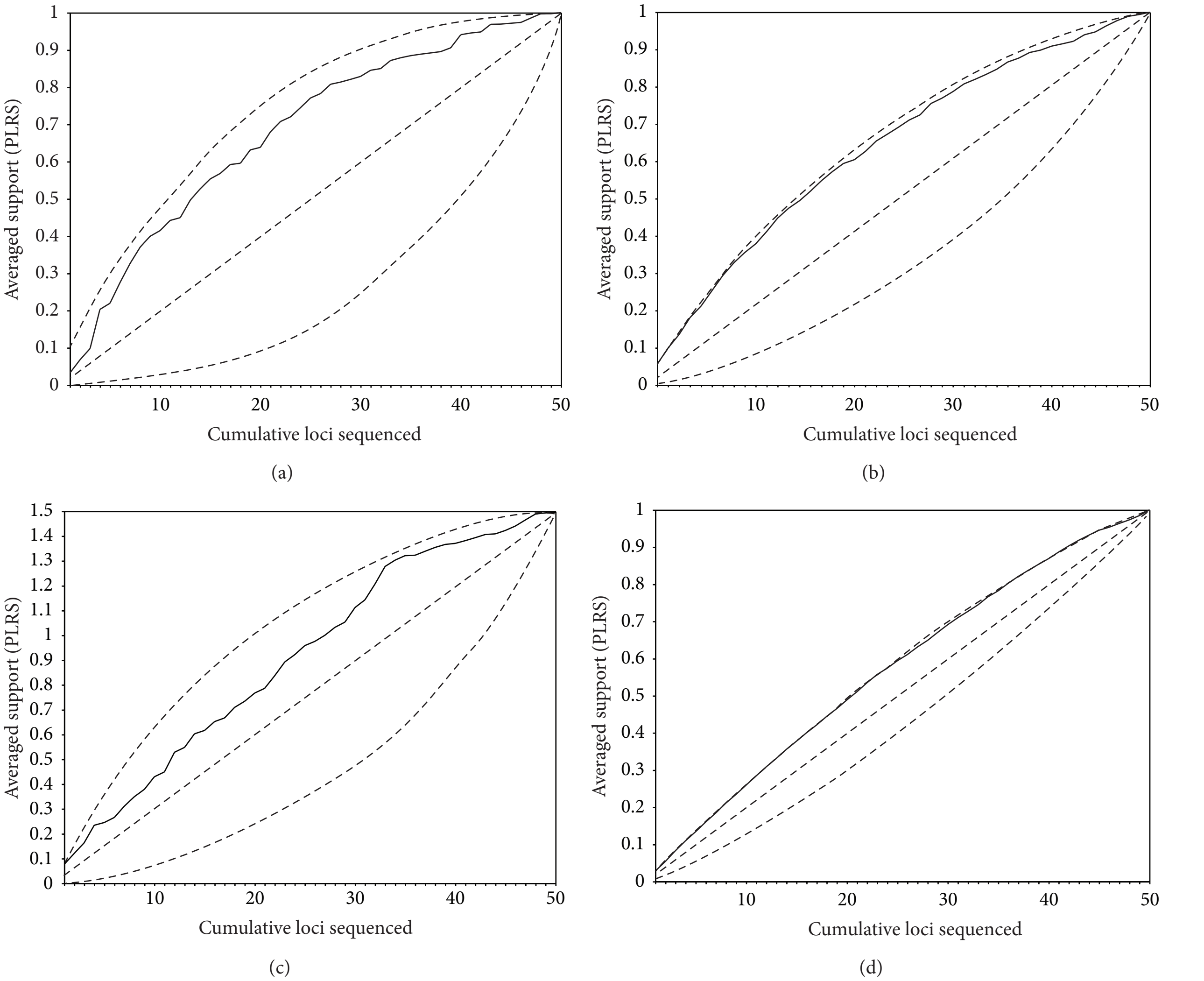


Figure 7: Cumulative proportionate likelihood-ratio supports (PLRS) averaged across nodes for (a) OrthoMaM amino acid, (b) FUNYBASE, (c) TBR, and (d) simulated amino acid data sets. Genes are ranked by differential phylogenetic informativeness encompassing all branches in the tree. The upper dashed line represents cumulative PLRS when loci are prioritized, posthoc, from highest to lowest PLRS values. The lower dashed line represents cumulative PLRS when loci are prioritized, posthoc, from lowest to highest PLRS values. The intermediate dashed line is the hypothetical average one would achieve sampling at random from loci available.

phylogenetic utility. Recently, responding to this necessity, two nongraphical strategies were proposed: (i) [43] suggests ranking genes by comparing the cophenetic correlation coefficients among individual protein distances matrices and (ii) [16] advocates ranking phylogenetic performance of genes using a topological metric, comparing individual gene topology against a reliable reference tree. These two approaches also provide insights into conflicting phylogenetic signal among genes, a practice followed more or less formally by phylogeneticists [44]. Both methods were tested in a similar set of fungal genomes, however, yielding different gene rankings [16]. Apparently, using topological distances is a superior strategy [16]. Although they have utility for ranking genes, such topological distance measures require a reference topology for the taxa of interest and extensive individual gene phylogenetic analyses. Since they yield an absolute rank rather than a function that modulates over historical time, they do not provide a domain of utility that may be extended to taxa within or outside the original analysis and thus cannot be the focus to determine genes that will be most useful for investigating phylogenetic questions at a given taxonomic level.

In order to estimate phylogenetic informativeness, one requires the site rate distribution for each locus. To obtain the rates, two prior pieces of information are needed: (1) an alignment of loci of interest pruned to contain a set of taxa for which the tree topology is fairly well known and (2) an ultrametric tree for those taxa. The ultrametric tree can be either a chronogram—an ultrametric tree with branch lengths proportional to time—or it can be in unspecified molecular evolutionary units. This prior information may be derived from three potential sources: (1) preliminary data on

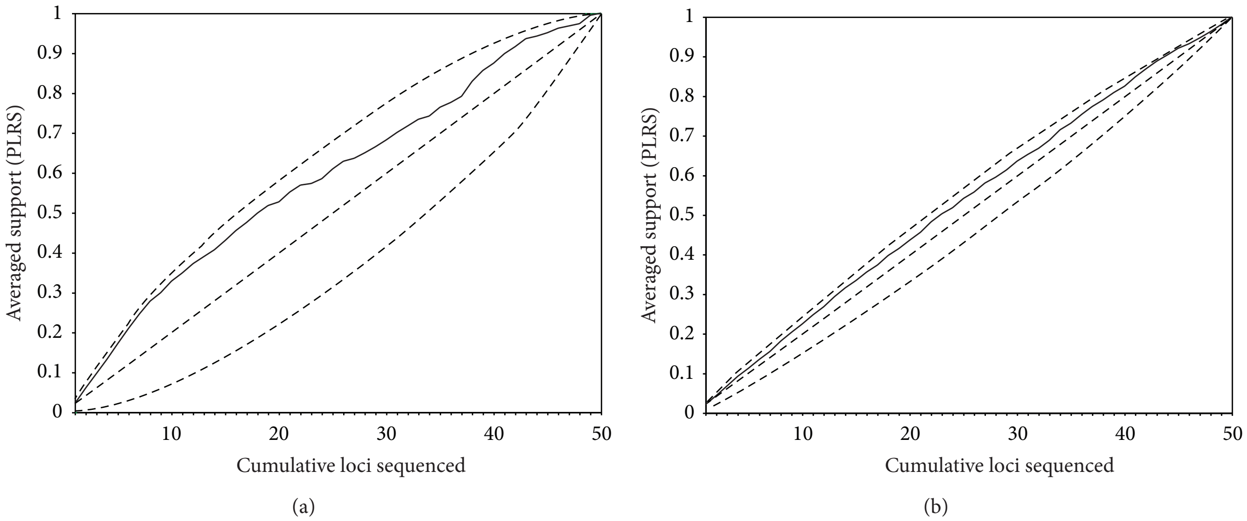


Figure 8: Cumulative proportionate likelihood-ratio supports (PLRS) averaged across nodes for (a) OrthoMaM DNA and (b) simulated DNA data sets. Genes are ranked by differential phylogenetic informativeness encompassing all branches in the tree. The upper dashed line represents cumulative PLRS when loci are prioritized, posthoc, from highest to lowest PLRS values. The lower dashed line represents cumulative PLRS when loci are prioritized, posthoc, from lowest to highest PLRS values. The intermediate dashed line is the hypothetical average one would achieve sampling at random from loci available.

the candidate loci from a well-studied subset of the taxa of interest; (2) data on the candidate loci from a well-studied sister clade; or (3) comparative genomic data from sequenced genomes within and/or outside the clade of interest. Informativeness profiles can be generated with the online application PhyDesign [27] and used to rank loci based on their ability to resolve nodes of interest, allowing assessment of the relative signal of genes within large data sets. However, erroneous rate estimations, such as those caused by an incorrect input topology, will affect the accuracy of informativeness profiles. Thus, iterative refinement/recalculation of PI while increasing taxonomic sampling is recommended for researchers seeking to identify the best candidate loci for phylogenetic reconstruction.

To apply all the information at hand and to assess results in a familiar way in this study, we used the phylogenetic informativeness of genes over the full epoch integrating from time 0 to the root. While generality and ease of presentation were gained by this procedure, the strength of the correlations observed herein was likely reduced as a cost of that generality. In fact, the rank order of genes by informativeness varies over history due to the pattern of variation of rates among sites. Comparison of the profiles of informativeness for the different data sets against their chronograms (Figures 1–3) illustrated the different gene potential for signal across their evolution. As a simple example, in Figure 2, MS409 (MetRS, mitochondrial methionyl-tRNA synthetase) from FUNYBASE shows a great potential for questions of recent fungi evolution and much lower potential signal for all of the rest of fungal history. In addition, genes showing the same mean rate can have different phylogenetic profiles [5], even though it is a common practice to talk about slow-evolving genes and rapid-evolving genes to define their temporal applicability. An incontrovertible example of this is found in the simulated sequences profiles (Figure 4), in that each variant with a different gamma-shape parameter showed a different informativeness profile. Numerous examples of the importance of incorporating rate variation among sites for the correct phylogenetic inference are found in the literature [45–48].

Phylogenetic informativeness ably predicted gene performance in all data sets, encompassing diverse evolutionary contexts. In contrast, it is common practice to explore the adequacy of a methodology for a single set of empirical data or under a single criterion for phylogenetic inference [5, 16, 43]. However, due to the complexity of the biological process that generates phylogenetic data, extrapolating strong conclusions from individual data sets can be inadvisable.

Initially, we had expected that phylogenetic noise might hinder gene prediction of performance for data sets with more ancient nodes. Some theoretical work associated with particular data sets has indicated that homoplasy obscures the phylogenetic signal for periods older than 600 MA, and eliminates the signal as 1000 MA is approached [42, 49]. However, we did not observe a pattern indicating that phylogenetic informativeness predicts better for one data set or another based on the age of the events encompassed by their phylogenies.

The contrast between the hypothetical ideal and worst prioritizing rankings for DNA data sets (Figure 8) was less than the contrast between their respective amino acid counterparts (Figure 7). A greater similarity in phylogenetic performance among DNA sequences than among amino acid sequences might be responsible for this pattern. The greater similarity of informativeness for DNA sequences can be attributed to the consistent presence of a homogeneous class of fast-evolving silent sites in DNA sequences. In contrast, amino acid sequences have no consistent, a priori identifiable fast-evolving site class like degenerate third codon position sites in DNA but instead may range from extreme constraint on all sites to lack of constraint on many sites. Thus, polymorphism at silent coding sites can lead to high-informativeness DNA profiles for genes with flat amino acid profiles and, for the same reason, make DNA profiles more similar to each other compared with amino acid profiles. The number of characters and the signal retention for each character will dictate the different phylogenetic performance of these two data types [50]. For most inference purposes, the net phylogenetic informativeness is the prediction of interest, as it should correlate with empirical results, such as the degree of support of a node. However, per site phylogenetic informativeness can be calculated to quantify the cost versus benefit of sequencing and to compare relative phylogenetic potential without the confounding effect of sequence length. For example, a top-ranked gene may show good net phylogenetic informativeness profiles, but there may be one or more shorter markers (requiring less sequencing effort) that may exhibit better per site profiles. A combination of shorter genes requiring the equivalent sequencing effort of a longer marker might lead to the best results.

Two explanations may underlie the observation that amino acid data sets tended to show stronger correlations between informativeness rankings and tree distances or BPs. The expanded character-state space accessible for amino acids compared with DNA sequences can diminish the potential misleading effects of homoplasy in phylogenetic inference. Because signal is accounted for in the Townsend [5] informativeness but the potential for misleading homoplasy is not, greater homoplasy in DNA sequences might have led to worse predictions for DNA markers than for amino acid markers. Although functional constraints in proteins can limit character-state space available for a given amino acid site, simulations indicate that small increases in the character-state space increase accuracy of phylogenetic inference [50]. A second reason for the better performance of informativeness predictions for amino acid sequences could be better alignment. The long evolutionary distances between sequences used can make alignment of homologous residues of DNA sequence much more challenging than alignment of corresponding amino acid sequences.

Simulation studies have been successfully applied to address questions of character and taxon sampling strategies [51–53] and also for comparing methods of branch support [54, 55]. Differences between results achieved with empirical data sets and results achieved with simulated data sets are likely to derive from the strict adherence to the model of evolution in simulations compared to frequent deviation from the model typical in empirical data. Simulations oversimplify the substitution process. To perform simulations, we incorporated a specific evolutionary model. Thus, the regularities of the model dictated regularity in the results. Stochasticity in the nature of the substitution process in empirical data precludes better predictions of gene performance in empirical data sets than in simulated data sets. Even when phylogenetic informativeness predicts that there are a considerable number of sites evolving at optimal rates, changes will not necessarily map at all or in sufficient numbers to the branches of interest. Mutation, selection, and genetic drift processes determine the number and position of differences observed. Stochasticity of the substitution process will affect any attempt at gene performance prediction.

We found that gene length, number of parsimony informative sites, and/or number of variable sites were also significantly correlated with the tree distances measures for the FUNYBASE and the simulated data sets. For all data sets except OrthoMaM, the number of variable sites and the number of parsimony informativeness sites were also significantly correlated with BPs. For genes in FUNYBASE, significant correlations of gene length and number of variable sites with gene performance have been observed previously [16]. Accordingly, [55] found that these three indices were significantly correlated with bootstrap values for some branches. However, none of these parameters could systematically be used as a predictor of single gene performance [55]. Moreover, the number of variable sites and particularly parsimony informative sites represent posthoc indices giving estimates of the amount of signal present in the alignments that cannot be justifiably projected to a different set of taxa or to a novel depth in a phylogeny, limiting their utility. Interestingly, the only parameter that predicted the gene performance in OrthoMaM DNA data set was the phylogenetic informativeness metric. This result reinforces the idea that incorporating character rate evolution for gene performance predictions is a key factor and that phylogenetic informativeness metrics can be successfully applied to systematically facilitate more cost-effective phylogenetic research. Extending the method of Townsend [5] to account for additional nuances of molecular evolution, such as the accumulation of homoplasy [56], will further bolster its applications to phylogenetic experimental design.

5. Conclusions

Phylogenetic analyses with broad taxonomic sampling, such as Tree of Life projects, can, with just a few of the right genes, accumulate sufficient data to build a reliable phylogeny (e.g., less than 10 genes) [16]. Ideally, the set of genes required would be minimized. Thus, as a consequence of the extensive selection of potential loci that may be pursued based on genome projects, a decision on what genes need to be sequences has to be made. Despite this need, pursuit of analytical methods for experimental design in phylogenetics has been sparse. We explored the impact of using the Townsend [5] phylogenetic informativeness and found it to be an advantageous procedure for using genome-scaled sequence data to identify loci with high utility for phylogenetic inference. This choice of genes is critical, not only for their global performance across depths of the tree but also for their performance in resolving particular timescales. By estimating the full distribution of rates at each site, phylogenetic informativeness profiles showed how signal content varies among genes and across time. Thus, one may select genes that will perform best for the epoch of interest, whether for recent divergence times or for more ancient divergence times. Prioritization predictions made by the Townsend [5] phylogenetic informativeness correlate with the accuracy and robustness with which a gene sequence recovers the correct topology. These predictions are valid for amino acid and DNA markers of diverse groups of organisms spanning broad time scales, especially when the time scale is not significantly deeper than the peak of informativeness. This quantitative and objective informativeness metric can play a critical role in augmenting the efficiency and accuracy of many phylogenetic studies at multiple time scales.

Acknowledgment

The most computationally time intensive analyses were performed in the Yale University Biomedical HPC Center clusters.

References

[1] M. P. Cummings and A. Meyer, "Magic bullets and golden rules: data sampling in molecular phylogenetics," *Zoology*, vol. 108, no. 4, pp. 329–336, 2005.

[2] J. B. Dacks and W. F. Doolittle, "Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help," *Cell*, vol. 107, no. 4, pp. 419–425, 2001.

[3] F. Delsuc, H. Brinkmann, and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 361–375, 2005.

[4] A. Graybeal, "Is it better to add taxa or characters to a difficult phylogenetic problem?" *Systematic Biology*, vol. 47, no. 1, pp. 9–17, 1998.

[5] J. P. Townsend, "Profiling phylogenetic informativeness," *Systematic Biology*, vol. 56, no. 2, pp. 222–231, 2007.

[6] A. H. Moeller and J. P. Townsend, "Phylogenetic informativeness profiling of 12 genes for 28 vertebrate taxa without divergence dates," *Molecular Phylogenetics and Evolution*, vol. 60, no. 2, pp. 271–272, 2011.

[7] C. A. M. Russo, N. Takezaki, and M. Nei, "Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny," *Molecular Biology and Evolution*, vol. 13, no. 3, pp. 525–536, 1996.

[8] R. Zardoya and A. Meyer, "Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates," *Molecular Biology and Evolution*, vol. 13, no. 7, pp. 933–942, 1996.

[9] N. Goldman, "Phylogenetic information and experimental design in molecular systematics," *Proceedings of the Royal Society B*, vol. 265, no. 1407, pp. 1779–1786, 1998.

[10] M. Miya and M. Nishida, "Use of mitogenomic information in Teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion," *Molecular Phylogenetics and Evolution*, vol. 17, no. 3, pp. 437–455, 2000.

[11] R. L. Mueller, "Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis," *Systematic Biology*, vol. 55, no. 2, pp. 289–300, 2006.

[12] V. Ranwez, F. Delsuc, S. Ranwez, K. Belkhir, M.-K. Tilak, and E. J. P. Douzery, "OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics," *BMC Evolutionary Biology*, vol. 7, no. 1, article 241, 2007.

[13] S. Marthey, G. Aguileta, F. Rodolphe et al., "FUNYBASE: a FUNgal phylogenomic dataBASE," *BMC Bioinformatics*, vol. 9, article 456, 2008.

[14] A. Rokas, D. Krüger, and S. B. Carroll, "Evolution: animal evolution and the molecular signature of radiations compressed in time," *Science*, vol. 310, no. 5756, pp. 1933–1938, 2005.

[15] J. W. Taylor and M. L. Berbee, "Dating divergences in the Fungal Tree of Life: review and new analyses," *Mycologia*, vol. 98, no. 6, pp. 838–849, 2006.

[16] G. Aguileta, S. Marthey, H. Chiapello et al., "Assessing the performance of single-copy genes for recovering robust phylogenies," *Systematic Biology*, vol. 57, no. 4, pp. 613–627, 2008.

[17] A. Rambaut and N. C. Grassly, "Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Computer Applications in the Biosciences*, vol. 13, no. 3, pp. 235–238, 1997.

[18] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[19] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[20] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[21] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist, "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference," *Bioinformatics*, vol. 20, no. 3, pp. 407–415, 2004.

[22] A. Rambaut and A. J. Drummond, Tracer v1. 4, 2007, http://beast.bio.ed.ac.uk/Tracer.

[23] M. J. Sanderson, "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock," *Bioinformatics*, vol. 19, no. 2, pp. 301–302, 2003.

[24] M. J. Benton and P. C. J. Donoghue, "Paleontological evidence to date the tree of life," *Molecular Biology and Evolution*, vol. 24, no. 1, pp. 26–53, 2007.

[25] I. Miranda, R. Silva, and M. A. S. Santos, "Evolution of the genetic code in yeasts," *Yeast*, vol. 23, no. 3, pp. 203–213, 2006.

[26] I. Mayrose, A. Mitchell, and T. Pupko, "Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account," *Journal of Molecular Evolution*, vol. 60, no. 3, pp. 345–353, 2005.

[27] F. López-Giráldez and J. P. Townsend, "Phydesign, a webapp for phylogenetic informativeness profiles," *BMC Evolutionary Biology*, vol. 11, article 152, 2012.

[28] W. J. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien, "Molecular phylogenetics and the origins of placental mammals," *Nature*, vol. 409, no. 6820, pp. 614–618, 2001.

[29] B. Schierwater, M. Eitel, W. Jakob et al., "Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis," *PLoS Biology*, vol. 7, no. 1, article e20, 2009.

[30] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[31] J. Felsenstein, *Phylip, Phylogeny Inference Package*, version 3.6. Distributed by the Author, Department of Genome Sciences, University of Washington, Seattle, Wash, USA, 2005.

[32] M. Anisimova and O. Gascuel, "Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative," *Systematic Biology*, vol. 55, no. 4, pp. 539–552, 2006.

[33] K. Bremer, "Branch support and tree stability," *Cladistics*, vol. 10, no. 3, pp. 295–304, 1994.

[34] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[35] T. M. Keane, C. J. Creevey, M. M. Pentony, T. J. Naughton, and J. O. McInerney, "Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified," *BMC Evolutionary Biology*, vol. 6, article 29, 2006.

[36] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[37] D. L. Swofford, PAUP*, Phylogenetic Analysis Using Parsimony, *∗and other methods.. Version 4. Sinauer Associates, Sunderland, Mass, USA, 2003.

[38] J. E. Stajich, D. Block, K. Boulez et al., "The Bioperl toolkit: perl modules for the life sciences," *Genome Research*, vol. 12, no. 10, pp. 1611–1618, 2002.

[39] S. A. Smith and C. W. Dunn, "Phyutility: a phyloinformatics tool for trees, alignments and molecular data," *Bioinformatics*, vol. 24, no. 5, pp. 715–716, 2008.

[40] A. Graybeal, "Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates," *Systematic Biology*, vol. 43, no. 2, pp. 174–193, 1994.

[41] K. Strimmer and A. Von Haeseler, "Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 13, pp. 6815–6819, 1997.

[42] W. T. White, S. F. Hills, R. Gaddam, B. R. Holland, and D. Penny, "Treeness triangles: visualizing the loss of phylogenetic signal," *Molecular Biology and Evolution*, vol. 24, no. 9, pp. 2029–2039, 2007.

[43] E. E. Kuramae, V. Robert, C. Echavarri-Erasun, and T. Boekhout, "Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom," *BMC Evolutionary Biology*, vol. 7, article 134, 2007.

[44] J. W. Leigh, E. Susko, M. Baumgartner, and A. J. Roger, "Testing congruence in phylogenomic analysis," *Systematic Biology*, vol. 57, no. 1, pp. 104–115, 2008.

[45] M. Hasegawa, A. Di Rienzo, T. D. Kocher, and A. C. Wilson, "Toward a more accurate time scale for the human mitochondrial DNA tree," *Journal of Molecular Evolution*, vol. 37, no. 4, pp. 347–354, 1993.

[46] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees," *Molecular Biology and Evolution*, vol. 10, no. 3, pp. 512–526, 1993.

[47] Z. Yang, "Among-site rate variation and its impact on phylogenetic analyses," *Trends in Ecology and Evolution*, vol. 11, no. 9, pp. 367–372, 1996.

[48] M. P. Simmons, L.-B. Zhang, C. T. Webb, A. Reeves, and J. A. Miller, "The relative performance of Bayesian and parsimony approaches when sampling characters evolving under homogeneous and heterogeneous sets of parameters," *Cladistics*, vol. 22, no. 2, pp. 171–185, 2006.

[49] D. Penny, B. J. McComish, M. A. Charleston, and M. D. Hendy, "Mathematical elegance with biochemical realism: the covarion model of molecular evolution," *Journal of Molecular Evolution*, vol. 53, no. 6, pp. 711–723, 2001.

[50] M. P. Simmons, A. Reeves, and J. I. Davis, "Character-state space versus rate of evolution in phylogenetic inference," *Cladistics*, vol. 20, no. 2, pp. 191–204, 2004.

[51] Z. Yang, "On the best evolutionary rate for phylogenetic analysis," *Systematic Biology*, vol. 47, no. 1, pp. 125–133, 1998.

[52] J. J. Wiens, "Can incomplete taxa rescue phylogenetic analyses from long-branch attraction?" *Systematic Biology*, vol. 54, no. 5, pp. 731–742, 2005.

[53] Y. Suzuki, G. V. Glazko, and M. Nei, "Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 25, pp. 16138–16143, 2002.

[54] M. P. Cummings, S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka, "Comparing bootstrap and posterior probability values in the four-taxon case," *Systematic Biology*, vol. 52, no. 4, pp. 477–487, 2003.

[55] A. Rokas, B. I. Williams, N. King, and S. B. Carroll, "Genome-scale approaches to resolving incongruence in molecular phylogenies," *Nature*, vol. 425, no. 6960, pp. 798–804, 2003.

[56] J. P. Townsend, Z. Su, and Y. Tekle, "Phylogenetic signal and noise, predicting the power of a dataset to resolve phylogeny," *Systematic Biology*, vol. 5, pp. 835–849, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 397278, 7 pages
http://dx.doi.org/10.1155/2013/397278

Research Article

Origin and Status of Homologous Proteins of Biomineralization (Biosilicification) in the Taxonomy of Phylogenetic Domains

Igor E. Pamirsky[1] and Kirill S. Golokhvast[2]

[1] *Analytical Centre of the Mineralogical and Geochemical Researches, Institute of Geology and Nature Management FEB RAS, 1 Relochny Lane, Blagoveshchensk, Russia*
[2] *Department of Oil and Gas Deal, Laboratory of Nanotoxicology, Far Eastern Federal University, 37 Pushkinskaya Street, Vladivostok, Russia*

Correspondence should be addressed to Igor E. Pamirsky; parimski@mail.ru and Kirill S. Golokhvast; droopy@mail.ru

Received 1 March 2013; Revised 16 April 2013; Accepted 3 June 2013

Academic Editor: Vassily Lyubetsky

The taxonomic affiliation (in the systematisation of viruses, and biological domains) of known peptides and proteins of biomineralization (silicateins, silaffins, silacidins and silicase) and their primary structure homologues were analyzed (methods *in silico*; using Uniprot database). The total number of known peptides and proteins of biosilicification was counted. The data of the quantitative distribution of the detected homologues found in nature are presented. The similarity of the primary structures of silaffins, silacidins, silicateins, silicase, and their homologues was 21–94%, 45–98%, 39–50%, and 28–40%, respectively. These homologues are found in many organisms, from the Protista to the higher plants and animals, including humans, as well as in bacteria and extracellular agents, and they perform a variety of biological functions, such as biologically controlled mineralisation. The provisional classification of these biomineralization proteins is presented. The interrelation of the origin of the first organic polymers and biomineralization is discussed.

1. Introduction

Minerals formed with the participation of various organisms are diverse [1–3]. Such mineral formations are called biominerals, and the process of their formation is known as biomineralization. Biologically mediated and biologically controlled biomineralizations are distinct [4]. As a rule, the latter products involve biominerals of endogenous origin. At the same time, calling the formation of some endogenous biominerals, such as urinary stones, a controlled process is difficult. Considering the latest international research in this field [5–15], controlled biomineralization, in a broad sense, should be understood as not only a process of the formation of mineral particles but also their subsequent transformations, that is, metabolism. The metabolism of physiogenic biominerals, the biochemical processes of which are determined genetically and proceed with the direct participation of a number of molecules of protein origin, is of special interest.

Peptides and proteins involved biomineralization can be termed “proteins of biomineralization” (POB) and are divided into the following groups: (1) peptides and polypeptides, which form an organic matrix on which minerals are formed, (2) enzymes that catalyse the formation of inorganic structures, (3) enzymes that catalyse the hydrolysis of inorganic structures, and (4) proteins transporting the structural components of biominerals. The high relevance of a study on POBs is caused by the depth of fundamentality, especially with respect to the interconnection of biology, geology, and medicine and the importance of the practical value, such as the synthesis of materials with specific functions for various technologies. However, many questions about the regularities and features of the biomineralization mechanisms (in particular, from the point of view of the POB) remain open, and the search for structural and functional homologues of known POBs in different organisms appears to represent one solution for these issues. In this regard, beginning the study of the most ancient primitive organisms,

Table 1: Quantitative distribution of the homologues with respect to origin.

Domains of biological taxonomy	Titles of POB			
	Silaffins	Silacidins	Silicateins	Silicase
Viruses	1%	0.8%	—	—
Archaea	0.6%	—	—	—
Bacteria	38.4%	—	—	—
Eukaryote	60%	99.2%	100%	100%
Total number of homologues	741	82	686	249

Note: peptides and proteins belonging to the group of proteins under study shown in the list of homologues were not considered.

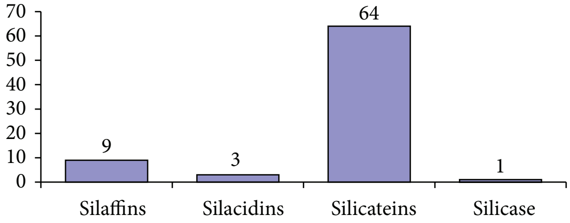


Figure 1: The number of known proteins of biomineralization in diatoms and sponges (from UniProt and the literature by the end of 2012).

which were first on the earth, to enable the construction of a biomineral using endogenous protein molecules appears logical. The magnetotactic bacteria forming magnetite belong to these organisms as well as diatoms and sponges, forming silicon dioxide, and other organisms.

Today, representatives of biosilification proteins cover all groups of the above POB classifications. Prominent representatives of these proteins are silaffins and silacidins in diatoms and silicateins and silicase in river and sea sponges. Silaffins and silacidins (phosphoproteins) are small peptides that catalyse the formation of silica nanospheres from silicic acid and control their size, playing a central role in the formation of the cell walls of diatoms. Silicateins catalyse the generation of amorphous silica from silicic acid esters, participating in the formation of the silicon skeleton. Silicase is the only known enzyme that performs the depolymerisation of silica. Silicon transporters (SIT) are proteins involved in the transmembrane transport of silicic acid. This paper is devoted to the study of enumerated peptides and proteins (silicon transporter (SIT) data not shown).

2. Materials and Methods

The amino acid sequences of all the studied proteins, except silicase and silacidins, were taken from the computational biology database server, UniProt (release 2012_10, http://www.uniprot.org). The data on silicase were taken from the paper by Schröder et al. [14], and the data on silacidins were taken from the paper by Richthammer et al. [15]. A comparative study of the homology of the amino acid sequences of peptides and proteins was performed using the same server (at the time of data retrieval from the database, there was information on approximately 30 million sequences). For comparison of the primary structures of biomineralization proteins and their families and groups, the multiple sequence alignment mode "Align" was used, implemented through "ClustalW 2.0.12" (with the mode settings not configured). The search for protein homologues was performed by pairwise alignment with the tool "BLAST" (BLASTP 2.2.26, Sep-21-2011, implemented through NCBI), as provided by the server, with the following parameters: database - UniProtKB, Threshold - 0, 1-0,0001, Matrix - Auto, Filtering - None (for proteins)/Filter low complexity regions (for peptides), Gapped - yes, Hits - 250. A similar approach was used in other studies [16–18].

Homologues of the primary structure were examined for three typical representatives of each group of proteins and peptides (except silicase). For the silaffin predecessor, short silaffins were chosen from *Cylindrotheca fusiformis* (ID Q9SE35, 265 Am), and silaffins were chosen from *Thalassiosira pseudonana* (ID Q5Y2C0, 231 Am; ID Q5Y2C1, 485 Am). For silacidins, natsilacidin A and silacidins B and C were selected. For silicateins, silicatein-α (ID B1GSK9, 334 Am) was selected from *Geodia cydonium*, and silicateins A1 (ID B5B2Z1, 329 Am) and A2 (ID B5LT52, 329 Am) were selected from *Latrunculia oparinae*.

3. Results and Discussion

3.1. The Distribution of the Homologues with respect to Origin and Their Taxonomic Affiliation. The number of studied biomineralization peptides and proteins, with the information detailed in the literature and submitted to the Uniprot database (with the polypeptide chains of some proteins represented partially in the database because their amino acid sequences have not been fully deciphered), is shown in Figure 1. Interestingly, since the discovery of silicase, none of its counterparts have been found.

The quantitative distribution of homologues with respect to origin and their taxonomic affiliations in the systematisation of viruses and biological domains are presented in Tables 1 and 2. Due to the peculiarity of the method used, the obtained data do not imply an affiliation of other (unknown, undiscovered, or not included in the number of results due to the limited number of issued results) and only show homologues to the taxa mentioned in the Tables.

Table 2: Taxonomy homologues proteins of biomineralization.

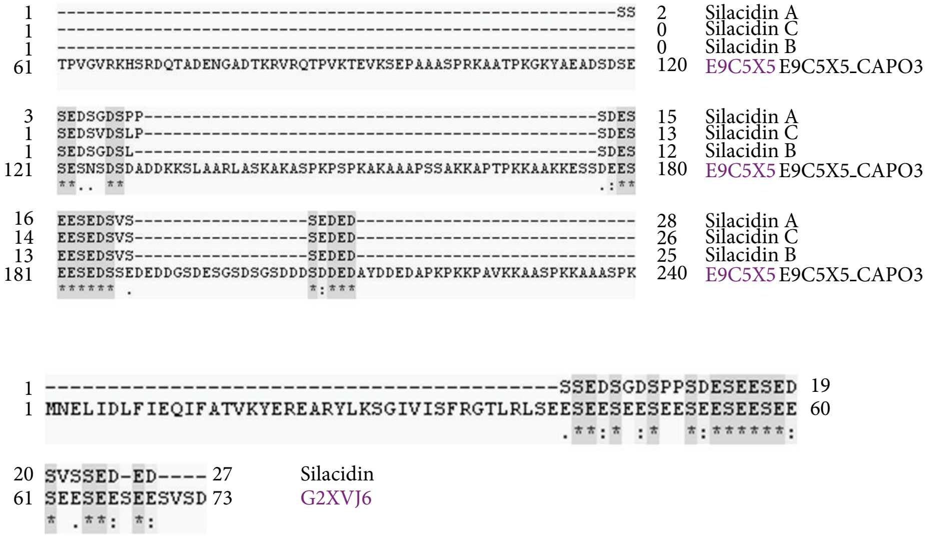
Taxonomy of viruses and organisms	Age (billions of years)	POB			
		1	2	3	4
Mimiviridae, Herpesviridae, Baculoviridae and so forth (vira**)	Fossils are not found [19]	+	+		
Bacilli, Gammaproteobacteria and so forth (bacteria)	3.8–3.5* [19]	+			
Dictyosteliomycota	—	+	+	+	
Lobosea	0.75 [20]			+	
Heterolobosea	—	+			
Kinetoplastida	0.099–0.093 [21]		+		
Conoidasida	—	+			
Aconoidasida	—	+			
Oligohymenophorea	—	+			
Demospongiae	0.542–0.516 [21]			+	+
Tricoplacia	—	+	+	+	
Hydrozoa	0.635–0.542 [21]	+		+	
Anthozoa	0.635–0.542 [21]			+	
Holothuroidea	0.513–0.505 [21]			+	
Chromadorea	—	+	+	+	+
Bdelloidea	—			+	
Branchiopoda	0.520–0.516 [21]	+	+	+	
Monogenea	—			+	
Polychaeta	0.520–0.516 [21]			+	
Gastropoda, Bivalvia	0.542–0.252 [21]	+		+	
Arachnida	0.418–0.416 [21]			+	
Insecta	0.412–0.391 [21]	+		+	+
Malacostraca	0.530–0.513 [21]			+	
Maxillopoda	0.268–0.265 [21]			+	+
Appendicularia	—		+	+	+
Leptocardii	0.541–0.485 [21]			+	
Actinopterygii	0.478–0.468 [21]	+		+	+
Amphibia	0.488–0.443 [21]	+		+	+
Aves	0.252–0.201 [21]	+		+	+
Mammalia	0.235–0.221 [21]	+	+	+	+
Homo sapiens***	0.005 [21]	+		+	+
Eurotiomycetes, Homobasidiomycetes and so forth (fungi)	0.048–0.46 [21]	+	+		
Phaeophyceae, Mamiellophyceae and so forth	0.485–0.150 [21]	+	+		
Liliopsida	0.122–0.112 [21]	+		+	
Magnoliopsida	0.388 –0.383 [21]	+			

Note: 1: silaffins; 2: silacidins; 3: silicateins; 4: silicase; +: homologue present; **(classification ICTV); ***(human is presented as a species as an exception). *3.5 authentic finding in siliceous rocks, 3.8 problematic.

3.2. Matrix Proteins (Silaffins and Silacidins)

3.2.1. Silaffins. The silaffins (four individual polypeptides and a predecessor of short silaffins) presented in the database belong to two species of diatoms. Homology was found only between polypeptides from *T. pseudonana*, which were Q5Y2C1 and Q5Y2C2 (485 and 501 Am) and Q5Y2C0 and B8BRK6 (include to 231 Am), and was 99% for each pair (names of the proteins are not shown, but the identification numbers are in the database). The short silaffins 1B, 1A2, and 1A1 (peptide lengths of 29, 18, and 29 Am from *C. fusiformis*) are identical to each other up to 32–60%, but their common polypeptide predecessor (265 Am) from *T. pseudonana* is not homologous.

The proteins identical to the studied silaffins up to 21–94% (matrix blosum 62; E value from 0 to 5.0×10^{-6}) are found mainly in cellular organisms (eukaryotes and bacteria) and in some viruses (Tables 1 and 2). The biological and molecular functions of the majority of found homologues are not associated with biomineralization or are unknown. Biomineralization proteins are detected only among silaffin-1 homologues. There are approximately 80 dentin sialophosphoprotein (DSPP): noncollagenous matrix dentin proteins in mammals regulating the mineralisation and the size and



```
  1 ----------------------------------------------------------SS   2  Silacidin A
  1 ------------------------------------------------------------   0  Silacidin C
  1 ------------------------------------------------------------   0  Silacidin B
 61 TPVGVRKHSRDQTADENGADTKRVRQTPVKTEVKSEPAAASPRKAATPKGKYAEADSDSE 120  E9C5X5 E9C5X5_CAPO3

  3 SEDSGDSPP-----------------------------------------------SDES  15  Silacidin A
  1 SEDSVDSLP-----------------------------------------------SDES  13  Silacidin C
  1 SEDSGDSL------------------------------------------------SDES  12  Silacidin B
121 SESNSDSDADDKKSLAARLASKAKASPKPSPKAKAAAPSSAKKAPTPKKAAKKESSDEES 180  E9C5X5 E9C5X5_CAPO3
    **.. **                                                 .:**

 16 EESEDSVS------------------SEDED-----------------------------  28  Silacidin A
 14 EESEDSVS------------------SEDED-----------------------------  26  Silacidin C
 13 EESEDSVS------------------SEDED-----------------------------  25  Silacidin B
181 EESEDSSEDEDDGSDESGSDSGSDDDSDDEDAYDDEDAPKPKKPAVKKAASPKKAAASPK 240  E9C5X5 E9C5X5_CAPO3
    ****** .                  *:***


  1 -----------------------------------------SSEDSGDSPPSDESEESED 19
  1 MNELIDLFIEQIFATVKYEREARYLKSGIVISFRGTLRLSEESEESEESEESEESEESEE 60
                                             .**:* :*  *:******:

 20 SVSSED-ED---- 27    Silacidin
 61 SEESEESEESVSD 73    G2XVJ6
    * .**: *:
```

Figure 2: Example of an alignment of the amino acid sequences of silacidins A (silacidin A, silacidin), B (silacidin V), C (silacidin S), DNA topoisomerase I *Capsaspora owczarzaki* (E9C5X5), and a undocumented protein from the fungus *Botrytis cinerea* (G2XVJ6).

shape of apatite crystals. The presence of phosphorylated amino acids in these proteins is an additional similarity to silaffins. The majority of serines in silaffins are known to be phosphorylated and are a source of anions in the formation of silica [11]. Lustrin A from the mollusc *Haliotis rufescens* is another homologue of silaffin-1, reinforcing the pearl layer of shells and pearls. Previously, Shen et al. [5] noted the similarity of the structures of lustrin, proteins (frustulins) forming the silicon skeleton of diatoms, extracellular matrix proteins composing the mineralised matrix of bone and dental tissues in mammals, and proteins composing avian egg shells. Some proteins (B8LDT2, B8LDT6, and B3ITC3) of the cellular wall of the diatom *T. pseudonana* and the bivalve *Crassostrea nippona*, with its calcite shell, are likely shown in the list of homologues but are uncharacterised and thus may participate in biomineralization.

Cementing or bonding proteins are worth paying attention to along with the other silaffin-1 homologues, such as the silk proteins of the silkworm *Bombyx mori*, the lacewing *Mallada signata*, the emby *Aposthonia gurneyi*, and *Haploembia solieri*, which are used for building cocoons and spider passages, and the cementing protein 3B of the worm *Phragmatopoma californica* required for the binding of sand and sea shell remains in the construction of habitation. These proteins are also characterised by a generous amount of serine and repeating regions, but they are not related to biomineralization. However, a detailed comparative study of their expressed adhesion properties to different surfaces (including minerals) may help improve the understanding of the mechanism of matrix biomineralization protein action.

Mucins should be mentioned (silaffin Q5Y2C2 homologues), as presented by several dozen representatives (not involved in biomineralization). Some mucin-like proteins are known to participate in the process of the mineralisation of mollusc shells [22] as well as of the bone, teeth, and cartilage of vertebrates [23]; that is, their biological functions are similar to those of silaffins. However, there were no mucin-like mineralising proteins in the list of the 250 homologues. A separate comparison of a typical example of such proteins (shellfish protein Q9BKM3; selected randomly) with Q5Y2C2 showed a low degree of sequence similarity of approximately 17%, which explains the results of the search for homologues in the database.

3.2.2. Silacidins. Homology between the A, B, and C silacidins was approximately 86%. A homology search revealed 54, 17, and 14 results for the A, B, and C silacidins, respectively. The identity to the detected proteins is 45–98% (with matrix pam 30). A significant difference in the length (2,7–64 times) between the silacidin polypeptide chains and these proteins explains the low similarity at the level of entire sequences (E value of 2.0×10^{-3} to 6.7×10^{-2}). At the same time, there is high homology with the individual chains sections of most found proteins (an example is shown in Figure 2). The functions of the majority of these homologues are currently unknown, and the remainder of the proteins mainly contain zinc ions and nucleic acids. In the mode of the high statistical threshold of significance (value $\geq$ 10) in the list of homologues, biomineralization proteins were reflected, such as the dentin matrix protein from the proboscis dog *Rhynchocyon petersi* and lemur *Lemur catta* and osteopontin-bone proteins from mouse *Mus musculus* and sea carp *Sparus aurata*.

Asterisks indicate identical amino acids, and "·" and ":" indicate chemically similar amino acids. The numbers indicate the ranges of the amino acid residues corresponding to the line. The topoisomerase sequence is incomplete.

3.3. Silica Polymerising (Silicateins) and Depolymerising (Silicase) Enzymes

3.3.1. Silicateins. With the method used, the degree of the identity of the primary structure of silicateins was approximately 40–99%. There was no one protein with a serine catalytic centre among all known silicatein homologues (homology of 39–50% with a matrix blosum of 62 and E value of 3.0×10^{-89} to 1.0×10^{-88}). Cathepsins L, S, and K (with the cysteine type catalytic centre) play the primary role in a wide range of unicellular and multicellular eukaryotes (all classes are listed in Table 2). If cathepsins L and S are not related to biomineralization processes, cathepsin K (tissue-specific enzymes of osteoclasts that break down bone matrix proteins) is only indirectly related to biomineralization (the direct contact with the formation of enzyme-substrate complexes is unknown). The molecular and biological functions of the other homologues are unknown.

3.3.2. Silicase. The silicase homology search of the sponge *S. domuncula* showed that all the proteins that are identical to it at 28–40% (matrix blosum 62, E value of 8.0×10^{-43} to 1.0×10^{-17}) are related to carbonate dehydratase. These homologues are predominantly found in organisms from the "Cambrian explosion" (see Tables 1 and 2), and silicase itself is most likely the "youngest" enzyme of all POBs. The biological roles of the homologues are known to varying degrees. For example, some are involved in bone resorption and osteoclast differentiation, but considering the mechanism of action, the analogy with silicase is impossible to draw.

Despite the relatively high level of similarity (and in most cases, conservation) in the primary organisation, the immediate analogues among the found silicase silicatein homologues are not shown. In the case of silicateins (serine protease), these omissions are explained by the difference in the structure of the catalytic centre from cysteine cathepsins. Other serine proteases may be able to operate with silicic acids. Silicase, as a representative of carbonic anhydrase, is of great interest to us. We also did not find information about the direct functional analogues of sponge silicase in the literature.

The silicase enzymes are supposed to be present in the silicate bacteria responsible for the destruction of Si–O bonds in the crystal lattice of clay minerals and the Si–C bonds in organosilicon compounds, but these enzymes are not isolated in a pure form [24, 25]. Some cellular organisms may produce silicase-like enzymes under certain circumstances. Such enzymes may well contribute to the assimilation (exchange) of silicon, entering the organism with water and nutriments. It is impossible not to take into account the existence of the lithophagy phenomenon in mammals [26]. Various silicon clays are widely used in modern medicine, including orally. Adhering to certain logic, it makes sense to use representative examples, such as the human enzyme chitotriosidase from the family of chitinases that implements the hydrolysis of chitin, which is a characteristic of the covering tissue of fungi, insects, and crustaceans. Although this substrate of chitinase is not a structural element of the human body, specific cells produce these enzymes under certain situations (in the case of some mucopolysaccharidoses) [27].

4. The Origin of Biomineralization Proteins

Silaffin homologues are found amongst representatives of all the kingdoms of organisms and virus taxa (see Table 2). A similar pattern is observed for silacidin homologues. Thus, it seems logical that matrix proteins (not even involved in biomineralization) combined with other plastic organic molecules form the basis of subcellular structures and cells in general. The POB homologues and proteins of biomineralization themselves detected in the representatives of specified taxa may have common ancestors, but such a hypothesis requires a detailed phylogenetic analysis. However, from the point of view of evolution, bacteria and viruses are of the greatest interest to us. Geological findings indicate that bacteria are the most ancient organisms on the earth. At the same time, the one-time (random) emergence of complex living systems such as bacteria is unlikely, but modern science is yet unable to answer clearly what transitional forms (stages) preceded the emergence of single-celled organisms. If subcellular agents (viruses) preceded a cellular form of life, they had to be carriers of the first matrix proteins. However, virus fossils still have not been found. In any case, initially for the construction of organic biological systems in the same time period, a set of organic molecules had to be specific, including those able to fulfil structural and metabolic functions. Where did these organic molecules for the construction of such systems, especially those proteins for which their synthesis was stipulated genetically, come from? According to some hypotheses [3, 28–33], minerals, acting as templates, catalysts, and/or metabolites, promoted the synthesis and interaction of organic molecules and the emergence of life. In an article devoted to mineral evolution, Hazen et al. [3] highlighted the probability that matrix (structural) proteins should be among the first organic polymers and that the emergence of life is associated with the achievement of a minimum level of mineral evolution. Kostetsky [31, 32] provided a fairly detailed and universal scenario of the simultaneous abiotic synthesis of nucleic acids and proteins (collagen, histones, and others) on apatite matrix (also on carbonate apatite, calcite, aragonite, cristobalite, and mica) and then through the formation of organic-crystalline complex, the emergence of protocells, which were derived from minerals, and the subsequent reproduction of the matrix mechanism, genetic code, DNA structure, and other crystal-chemical features. It follows that an inorganic matrix and its synthesis on organic molecules, including polymers, must have occurred at approximately the same time, which is hard to believe in practice. Nevertheless, the previous prebiotic synthesis of molecules does not exclude the emergence of unrelated protocells but does include homologous polypeptides and polynucleotides. This picture fits into Zavarzin's opinion [34], who hypothesised the existence of a "universal ancestor" to be logically contradictory, emphasising the obligatoriness of the diversity and functional complementarity of the original group of microorganisms. Presumably, the main metabolism types were formed no later than 3.5 billion years ago (cyanobacteria, as discussed), and replacing inaccessible metals with those available in the enzyme structure was likely one of the main methods of early metabolic evolution [28].

Bacteria copied all the possible abiotic reactions associated with clay minerals, with the main difference being their speed [35]. Thus, biomineralization is essentially the reverse process, with biological molecules and supramolecular structures acting as intermediaries of mineralisation [36].

The repeated occurrence of biomineralization in the process of evolution as part of metabolism, depending on the environmental conditions, may mean that organisms already had the necessary "tools" (proteins). Thus, proteins able to "work" with biominerals (especially matrix proteins) may have been found before, and generic ancestors of these organisms may have already had a minimal set of such proteins. In this regard, the mass mineralisation in the Cambrian and further diversification of mineral skeletons in the Ordovician (accompanied by significant changes in the environment) are logical outcomes when many new taxa were "ready" to metabolise the compounds of calcium, phosphate, and silicon.

5. Conclusions

The proteins with primary structures that are moderately and highly homologous to silaffins, silacidins, silicase, and silicateins occur in many different organisms, from Protozoans to the higher plants and animals, including humans, as well as in bacteria and extracellular agents.

The biological and molecular functions of these homologues vary (e.g., protein binding, binding of metal ions, transferase activity, and proteolysis), but most of them are not directly related to the formation of biomineral particles. Only a few homologues are direct analogues of silaffins and silacidins or are able to participate only indirectly in biomineralization (silicase silicatein homologues).

The data on silaffins and silacidins allow the evolutionary relationships of different biomineralization types to be considered more closely. The formation mechanisms of silica, phosphate, and calcium carbonate particles on such protein matrices are likely fundamentally similar.

Research in this area enhances the understanding of the mechanisms of the formation of physiogenic and pathogenic biominerals as well as the origins of life and the coevolution of the living and nonliving.

Acknowledgment

The project was supported by the Scientific Fund of Far Eastern Federal University.

References

[1] H. C. W. Skinner and A. H. Jahren, "Biomineralization," in *Treatise on Geochemistry*, pp. 117–184, Elsevier, Amsterdam, The Netherlands, 2003.

[2] N. P. Yushkin et al., *The Origin of the Biosphere and the Co-Evolution of Mineral and Biological Worlds*, Institute of geology, Komi Science Center, Ural Branch RAS, Syktyvkar, Russia, 2007.

[3] R. M. Hazen, D. Papineau, W. Bleeker et al., "Mineral evolution," *American Mineralogist*, vol. 93, no. 11-12, pp. 1693–1720, 2008.

[4] X. Wang, S. Hu, L. Gan, M. Wiens, and W. E. G. Müller, "Sponges (Porifera) as living metazoan witnesses from the Neoproterozoic: biomineralization and the concept of their evolutionary success," *Terra Nova*, vol. 22, no. 1, pp. 1–11, 2010.

[5] X. Shen, A. M. Belcher, P. K. Hansma, G. D. Stucky, and D. E. Morse, "Molecular cloning and characterization of Lustrin A, a matrix protein from shell and pearl nacre of Haliotis rufescens," *Journal of Biological Chemistry*, vol. 272, no. 51, pp. 32472–32481, 1997.

[6] K. Shimizu, J. Cha, G. D. Stucky, and D. E. Morse, "Silicatein α: cathepsin L-like protein in sponge biosilica," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 11, pp. 6234–6238, 1998.

[7] J. N. Cha, K. Shimizu, Y. Zhou et al., "Silicatein filaments and subunits from a marine sponge direct the polymerization of silica and silicones *in vitro*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 2, pp. 361–365, 1999.

[8] N. Kröger, R. Deutzmann, and M. Sumper, "Polycationic peptides from diatom biosilica that direct silica nanosphere formation," *Science*, vol. 286, no. 5442, pp. 1129–1132, 1999.

[9] N. Kröger, R. Deutzmann, C. Bergsdorf, and M. Sumper, "Species-specific polyamines from diatoms control silica morphology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 26, pp. 14133–14138, 2000.

[10] N. Kröger, R. Deutzmann, and M. Sumper, "Silica-precipitating peptides from diatoms: the chemical structure of silaffin-1A from Cylindrotheca fusiformis," *Journal of Biological Chemistry*, vol. 276, no. 28, pp. 26066–26070, 2001.

[11] N. Kröger, S. Lorenz, E. Brunner, and M. Sumper, "Self-assembly of highly phosphorylated silaffins and their function in biosilica morphogenesis," *Science*, vol. 298, no. 5593, pp. 584–586, 2002.

[12] N. Kröger, "Prescribing diatom morphology: toward genetic engineering of biological nanomaterials," *Current Opinion in Chemical Biology*, vol. 11, no. 6, pp. 662–669, 2007.

[13] W. E. G. Müller, J. Li, H. C. Schröder, L. Qiao, and X. Wang, "The unique skeleton of siliceous sponges (Porifera; Hexactinellida and Demospongiae) that evolved first from the Urmetazoa during the Proterozoic: a review," *Biogeosciences Discussions*, vol. 4, no. 1, pp. 385–416, 2007.

[14] H. C. Schröder, A. Krasko, D. Brandt et al., "Silicateins, silicase and spicule-associated proteins: synthesis of demosponge silica skeleton and nanobiotechnological applications," in *Porifera Research: Biodiversity, Innovation and Sustainability*, pp. 581–592, Museum Nacional, Rio de Janeiro, Brazil, 2007.

[15] P. Richthammer, M. Börmel, E. Brunner, and K.-H. van Pée, "Biomineralization in diatoms: the role of silacidins," *ChemBioChem*, vol. 12, no. 9, pp. 1362–1366, 2011.

[16] K. S. Golokhvast, I. E. Pamirsky, and A. M. Panichev, "Homology of bacteria proteins, diatoms and sponges participating in biomineralization, and human proteins, and other animals," *Pacific Science Review*, vol. 13, no. 1, pp. 39–46, 2011.

[17] I. E. Pamirsky, E. A. Borodin, and M. A. Shtarberg, *Regulation of Proteolysis of Plant and Animal Inhibitors*, LAP Lambert Academy Publishing GmbH, Saarbrücken, Germany, 2012.

[18] I. E. Pamirsky and K. S. Golokhvast, "Bioinformatic study of homology and the evolution of proteins involved in biomineralization," *Informatics and Control Systems*, vol. 1, no. 27, pp. 80–86, 2011 (Russian).

[19] G. A. Danukalova, *Paleontology in the Tables. Guidance*, GERS Publishing, Tver, Russia, 2009.

[20] S. M. Porter, "The proterozoic fossil record of heterotrophic eukaryotes," in *Neoproterozoic Geolobiology and Paleobiology*, S. Xiao and A. J. Kaufman, Eds., pp. 1–21, Springer, Dordrecht, The Netherlands, 2006.

[21] 2013, http://paleodb.org/.

[22] F. Marin, P. Corstjens, B. de Gaulejac, E. de Vrind-De Jong, and P. Westbroek, "Mucins and molluscan calcification: molecular characterization of mucoperlin, a novel mucin-like protein from the nacreous shell layer of the fan mussel Pinna nobilis (Bivalvia, Pteriomorphia)," *Journal of Biological Chemistry*, vol. 275, no. 27, pp. 20667–20675, 2000.

[23] R. J. Midura and V. C. Hascall, "Bone sialoprotein—a mucin in disguise?" *Glycobiology*, vol. 6, no. 7, pp. 677–681, 1996.

[24] M. P. Kolesnikov, "Form of silicon in plants," *Biological Chemistry Review*, vol. 41, pp. 301–332, 2001 (Russian).

[25] N. E. Samsonova, "The role of silicon in the formation of the phosphate regime of sod-podzolic soils," *Agrochem*, vol. 8, pp. 11–18, 2005 (Russian).

[26] A. M. Panichev, K. S. Golokhvast, A. N. Gulkov, and I. Y. Chekryzhov, "Geophagy and geology of mineral licks (kudurs): a review of russian publications," *Environmental Geochemistry and Health*, vol. 35, no. 1, pp. 133–152, 2013.

[27] R. U. Vysotskyay and N. N. Nemova, *Lysosomes and Lysosomal Enzymes of Fish*, Nauka, Moscow, Russia, 2008.

[28] M. A. Fedonkin, "The narrowing of the basis of life and geochemical evkariotizatsiya biosphere: the causal relationship," *Journal of Paleontology*, vol. 6, pp. 33–40, 2003 (Russian).

[29] S. N. Golubev, "Mineral crystals inside the organisms and their role in origin of life," *Zhurnal Obshchei Biologii*, vol. 48, pp. 784–806, 1987.

[30] S. N. Golubev, "Alive crystals," *Nature*, vol. 3, pp. 13–21, 1989.

[31] E. Y. Kostetsky, "The possibility of the formation of protocells and their structural components on the basis of the apatite matrix and cocrystallizing minerals," *Journal of Biological Physics*, vol. 31, no. 3-4, pp. 607–638, 2005.

[32] E. Y. Kostetsky, "How did the life begin?" *Achieve Life Sciences*, vol. 2, pp. 38–67, 2010 (Russian).

[33] A. H. Knoll, "Biomineralization and evolutionary history," *Reviews in Mineralogy and Geochemistry*, vol. 54, no. 1, pp. 329–356, 2003.

[34] G. A. Zavarzin, "The formation of biogeochemical cycles," *Journal of Paleontology*, vol. 6, pp. 16–24, 2003 (Russian).

[35] E. B. Naimark, V. A. Erouschev-Shack, N. P. Chizhikova, and E. I. Kompantseva, "Interaction of clay minerals with microorganisms: a review of experimental data," *Zhurnal Obshchei Biologii*, vol. 70, no. 2, pp. 155–167, 2009.

[36] I. S. Barskov, "Biomineralization and evolution. Coevolution of the mineral and biological worlds," in *Biosphere Origin and Evolution*, N. Dobretsov, N. Kolchanov, A. Rozanov, and G. Zavarzin, Eds., pp. 211–218, Springer, Berlin, Germany, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 760681, 11 pages
http://dx.doi.org/10.1155/2013/760681

Research Article

Molecular Identification and Ultrastructural and Phylogenetic Studies of Cyanobacteria from Association with the White Sea Hydroid *Dynamena pumila* (L., 1758)

O. A. Koksharova,[1,2] T. R. Kravzova,[3] I. V. Lazebnaya,[4] O. A. Gorelova,[3] O. I. Baulina,[3] O. E. Lazebny,[5] T. A. Fedorenko,[3] and E. S. Lobakova[3]

[1] *Lomonosov Moscow State University, Belozersky Institute of Physico-Chemical Biology, Leninskie Gory 1-40, Moscow 119992, Russia*
[2] *Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia*
[3] *Lomonosov Moscow State University, Faculty of Biology, Leninskie Gory 1-12, Moscow 119991, Russia*
[4] *N.I.Vavilov Institute of General Genetics, Russian Academy of Science, Gubkin Street 3, Moscow 119333, Russia*
[5] *N.K.Kol'tsov Institute of Developmental Biology, Russian Academy of Science, Vavilova Street 26, Moscow 119334, Russia*

Correspondence should be addressed to O. A. Koksharova; oa-koksharova@rambler.ru

Received 28 February 2013; Accepted 22 April 2013

Academic Editor: Vassily Lyubetsky

Three new cyanobacterial strains, that have been previously purified from the hydroid *Dynamena pumila* (L., 1758), isolated from the White Sea, were studied using scanning and transmission electron microscopy methods and were characterized by using almost complete sequence of the 16S rRNA gene, internal transcribed spacer 16S-23S rRNA, and part of the gene for 23S rRNA. The full nucleotide sequences of the rRNA gene clusters were deposited to GenBank (HM064496.1, GU265558.1, JQ259187.1). Comparison of rRNA gene cluster sequences of *Synechococcus* cyanobacterium 1Dp66E-1, *Oscillatoriales* cyanobacterium 2Dp86E, and *Nostoc* sp. 10Dp66E with all sequences present at the GenBank shows that these cyanobacterial strains do not have 100% identity with any organisms investigated previously. Furthermore, for the first time heterotrophic bacterium, associated with *Nostoc* sp. 10Dp66E, was identified as a member of the new phylum Gemmatimonadetes, genus of *Gemmatimonas* (GenBank accession number is JX437625.1). Phylogenetic analysis showed that cyanobacterium *Synechococcus* sp. 1Dp66E-1 forms the unique branch and belongs to a cluster of *Synechococcus*, including freshwater and sea strains. *Oscillatoriales* cyanobacterium 2Dp86E belongs to a cluster of *Leptolyngbya* strains. Isolate *Nostoc* sp. 10Dp66E forms unique branch and belongs to a cluster of the genus *Nostoc*, with the closest relative of *Nostoc commune* isolates.

1. Introduction

Cyanobacterial symbioses and associations with eukaryotes are widely distributed in aquatic and terrestrial environments. Various cyanobacterial strains form associations with sponges, hydroids, corals, dinoflagellates, radiolarians, and tintinnids [1–8]. Mostly, the studies of cyanobacterial associations are limited to subtropical and tropical marine water systems. Symbioses and associations between cyanobacteria and marine animals of high latitudes are poorly studied. Recently we isolated the several cyanobacteria from the associations with White Sea hydroid *Dynamena pumila* (L., 1758) [7]. Purification of the symbiotic cyanobacteria from the associations with animals is rarely successful. Four cyanobacteria were isolated and purified into culture from native samples of hydroid *D. pumila* and were identified based on morphological characteristics as representatives of the I, II, and IV subsections of phylum BX Cyanobacteria [7].

The goal of this study was the identification of the new cyanobacteria from association with the hydroid *Dynamena pumila* and their characterization on the basis of ultrastructural and molecular features.

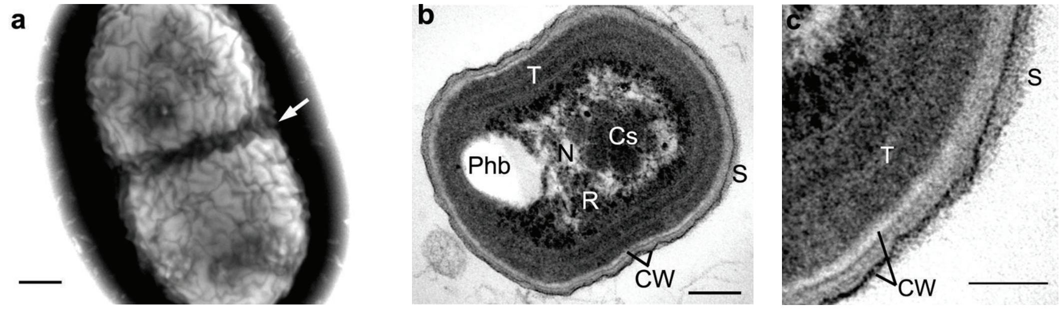


Figure 1: Ultrastructure of cyanobacterium 1Dp66E-1: **a**: TEM negative staining image; **b**: TEM ultrathin section; **c**: enlarged detail of Figure 1(b). Cs: carboxysome; CW: cell wall; N: nucleoid; Phb: poly-β-hydroxybutyrate; R: ribosomes; S: S-layer; T: thylakoid. Arrows indicate site of constriction. Scale bars: 0.2 μm (**a**, **b**) and 0.1 μm (**c**).

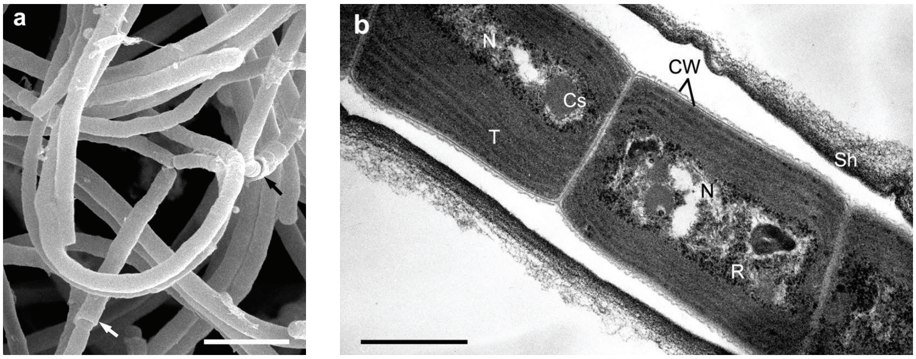


Figure 2: Ultrastructure of cyanobacterium 2Dp86E: **a**: SEM image; **b**: TEM ultrathin section. Cs: carboxysome; CW: cell wall; N: nucleoid; R: ribosomes; T: thylakoids; Sh: sheath. Arrows indicate sleeve-like thickening of sheath. Scale bars: 5 μm (**a**) and 0.5 μm (**b**).

2. Materials and Methods

2.1. Organisms and Growth Conditions. Cyanobacterial isolates 1Dp66E-1, 2Dp86E, and 10Dp66E have been previously purified from the samples of the hydroid *Dynamena pumila* (L., 1758) collected in June and August of 2006 in the Rugozero Bay of the Kandalaksha Gulf of the White Sea (66° 34′ N, 33° 08′ E) [7]. Cyanobacterial isolates 1Dp66E-1, 2Dp86E, and 10Dp66E have been grown on solid (1.5% agar) and liquid medium BG-11 [9] at 25°C and an illumination of 50 μmol photons m^{-2} s^{-1}.

2.2. Light Microscopy. The isolated cyanobacteria have been studied under Leitz Laborlux D microscope (Ernst Leitz Wetzlar GmbH, Germany).

2.3. Transmission Electron Microscopy (TEM). For negative staining, drop of cyanobacteria suspension was placed onto grids and treated with 1% (w/v) sodium phosphotungstate solution and dried. For the preparation of ultrathin sections, cyanobacteria were fixed in 2% (w/v) glutaraldehyde in 0.1 M sodium cacodylate buffer (pH 7,2) for 0.5 h and then postfixed in 1% (w/v) osmium tetroxide in the same buffer for 4 h, dehydrated by incubation in dilutions of ethanol, including absolute ethanol saturated with uranyl acetate, and embedded in araldite. Thin sections were prepared on an LKB-8800 (Sweden) ultratome and stained with lead citrate [10]. Ultrathin sections were examined with transmission electron microscopes JEM-100B and JEM-1011 (JEOL, Japan).

2.4. Scanning Electron Microscopy (SEM). Cyanobacterial samples were fixed as described above and dehydrated through an ethanol series, with an overnight exposure in absolute acetone followed by critical-point drying in a Dryer HCP-2 (Hitachi, Japan), coated with Au-Pd alloy in an IB-3 Ion Coater (Eiko, Japan), and examined with a JSM-6380LA scanning electron microscope (JEOL, Japan).

2.5. DNA Isolation and PCR Amplification. The DNA samples from cyanobacterial isolates 1Dp66E-1, 2Dp86E, and 10Dp66E were isolated according to the method of Koksharova [11]. The synthetic oligonucleotides (Table 1, "Synthol," Moscow, Russia) have been used as primers in the PCR reactions [12] in case of cyanobacterial isolates 1Dp66E-1, 2Dp86E, and 10Dp66E.

PCR was carried out on a Tercik DNA amplifier (DNA Technology, Russia) by using DreamTaq PCR Master Mix (Fermentas, EU), under the following conditions: 1 cycle at 94°C for 10 min, 25 cycles at 94°C for 45 sec, 54°C for 45 sec, 68°C for 2 min, 1 cycle at 68°C for 7 min, and a final soak step at 4°C. PCR products were resolved in 1.5% agarose gel containing ethidium bromide at 5 microgram mL^{-1}. All experiments were repeated at least three times.

Table 1: Primers.

Strain and a name of the primer	5′-3′ nucleotide sequence of the primer	Reference and a reaction where a primer was used
10Dp66E, *16S27F*	AGAGTTTGATCCTGGCTCAG	PCR
1Dp66E-1, 2Dp86E *16S378F*	GGGGAATTTTCCGCAATGGG	[23], PCR
1Dp66E-1, 2Dp86E, 10Dp66E *23S30R*	CTTCGCCTCTGTGTGCCTAGGT	[24], PCR
1Dp66E-1, 2Dp86E *16S534F*	GCCCACAGCTCAACTGTGG	This work, PCR
1Dp66E-1, 2Dp86E *23S1665R*	CGCTCTAACCACCTGAGC	This work, PCR
2Dp86E *CYA781 Rfil*	GACTACTGGGGTATCTAATCCCATT	This work, sequencing
1Dp66E-1 *CYA781Rnonfil*	GACTACAGGGGTATCTAATCCCTTT	This work, sequencing
2Dp86E *FIL6*F	CATGTCGCGAATCTTTCAG	This work, sequencing
2Dp86E *FIL7F*	CGTCCTACAATGCTACAGAC	This work, sequencing
2Dp86E *FIL8F*	CGAGTGGTCACTCTAGG	This work, sequencing
2Dp86E *FIL9F5*	CGACTGACTGGACTAATGG	This work, sequencing
2Dp86E *FIL10R*	AACCGCTGACATCCTGCT	This work, sequencing
2Dp86E *FIL11R*	GTGAGCCCGTTGTAGCTT	This work, sequencing
1Dp66E-1 *UNI 3F*	ATGGACGAAAGCCAGGGGAGC	This work, sequencing
1Dp66E-1 *UNI 4R*	GACCTGCGATTACTAGCGATG	This work, sequencing
1Dp66E-1 *UNI 5F*	GGCGTTCAACGTGTCCGATCC	This work, sequencing
1Dp66E-1 *UNI 6R*	GAGTCCTCAGCTGAACATGTCC	This work, sequencing
1Dp66E-1 *UNI 7F*	GTGAGCCCGTTGTAGCTT	This work, sequencing
1Dp66E-1 *UNI 1F*	CTCGGTGTCGTAGCTAACGC	This work, sequencing
1Dp66E-1 *UNI 1R*	CATCCTGCTTGCAAAGCAG	This work, sequencing
1Dp66E-1 *UNI 8R*	GAGGTTTACAGCCCAGAG	This work, sequencing
1Dp66E-1 *UNI 9R*	GCATCGAATTAAACCAC	This work, sequencing
1Dp66E-1 *UNI 10F*	GACAGGTGGTGCATGGC	This work, sequencing
10Dp66E *10Dp F*	GCTGACCTGCAATTACTAGC	This work, sequencing
10Dp66E *10Dp R*	GCGCTTTCGCCACCGGTGTTC	This work, sequencing
1Dp66E-1, 2Dp86E *16S15-12-08F*	GCCCACAGCTCAACTGTGG	This work, sequencing

Table 1: Continued.

Strain and a name of the primer	5′-3′ nucleotide sequence of the primer	Reference and a reaction where a primer was used
1Dp66E-1, 2Dp86E *23S15-1208R*	GGCCATCCTGGACTCGAAC	This work, sequencing
1Dp66E-1, 2Dp86E *23S16-1208R*	CGCTCTAACCACCTGAGC	This work, sequencing
1Dp66E-1, 2Dp86E 23S31R	CTCAACCATAGTCTAGAAAC	This work, sequencing
1Dp66E-1, 2Dp86E *23S32R*	GTCCTGAACGACCTAGAG	This work, sequencing

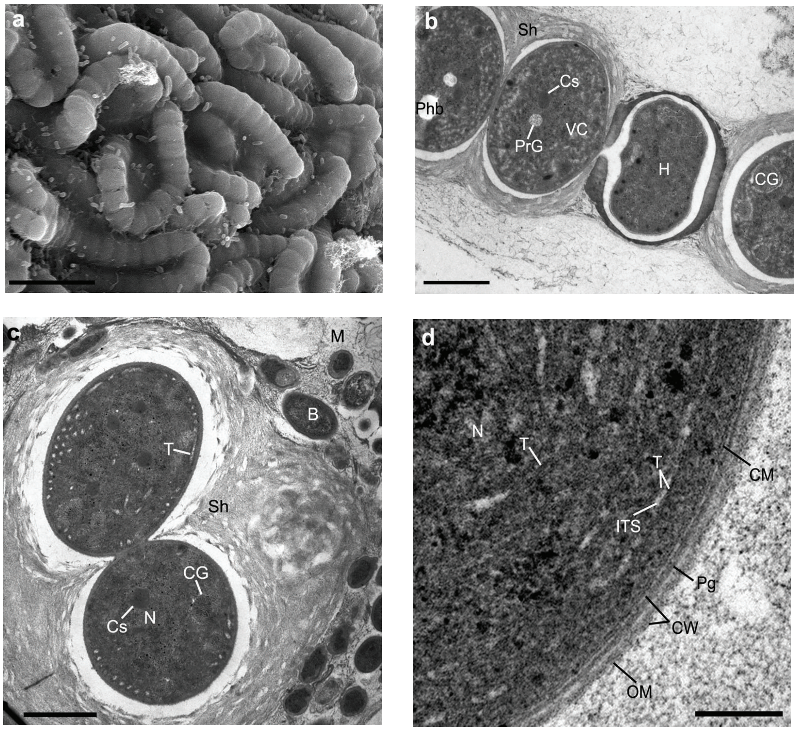


Figure 3: Ultrastructure of cyanobacterium 10Dp66E: **a**: SEM image; **b**: TEM ultrathin section of vegetative trichome; **c**: TEM ultrathin section of cell cluster; **d**: TEM ultrathin section of the cell part in the cell wall and thylakoids areas. B: bacteria; CG: cyanophycin granule; CM: cytoplasmic membrane; Cs: carboxysome; CW: cell wall; H: heterocyst; ITS: intrathylakoid space; M: intercellular matrix; N: nucleoid; OM: outer membrane; Pg: peptidoglycan; Phb: poly-β-hydroxybutyrate; PrG: protein granules; Sh: sheath; T: thylakoids; VC: vegetative cell. Scale bars: 10 μm (**a**), 1 μm (**b**, **c**) and 0.2 μm (**d**).

2.6. Cloning and Sequencing of PCR Products. DNA fragments obtained during PCR were cloned with CloneJet PCR Cloning Kit no. K1231 (Fermentas, EU). Transformation of competent XL-1 cells of *Escherichia coli* and plasmid purification were performed accordingly [13]. DNA sequencing was performed with ABI PRISM BigDye Terminator v. 3.1 at the Applied Biosystems 3730 DNA Analyzer (Center for Collective Use “Genome”). Sequences were edited and assembled with Bioedit (Invitrogen, Carlsbad, CA). Sequences were analyzed with BLAST software (http://www.ncbi.nlm.nih.gov/BLAST/) in order to identify their closest relatives. The full nucleotide sequences of the rRNA gene cluster of all three cyanobacteria were accomplished and deposited to GenBank under accession numbers HM064496.1, GU265558.1, and JQ259187.1.

2.7. Phylogenetic Analysis. Search of the nucleotide sequences in the database GenBank, homologous to the sequenced genes of studied species of cyanobacteria, was performed using BLAST (http://www.ncbi.nlm.nih.gov/entrez/viewer) with the option—the least degree of similarity (minimum identity). The sequences of selected species were aligned using the algorithm ClustalW v.1.6 (MEGA 5.1) [14]. Phylogenetic reconstructions were performed using clustering neighbor-joining method (MEGA 5.1) with the preselection of an adequate model of nucleotide substitutions. Statistical

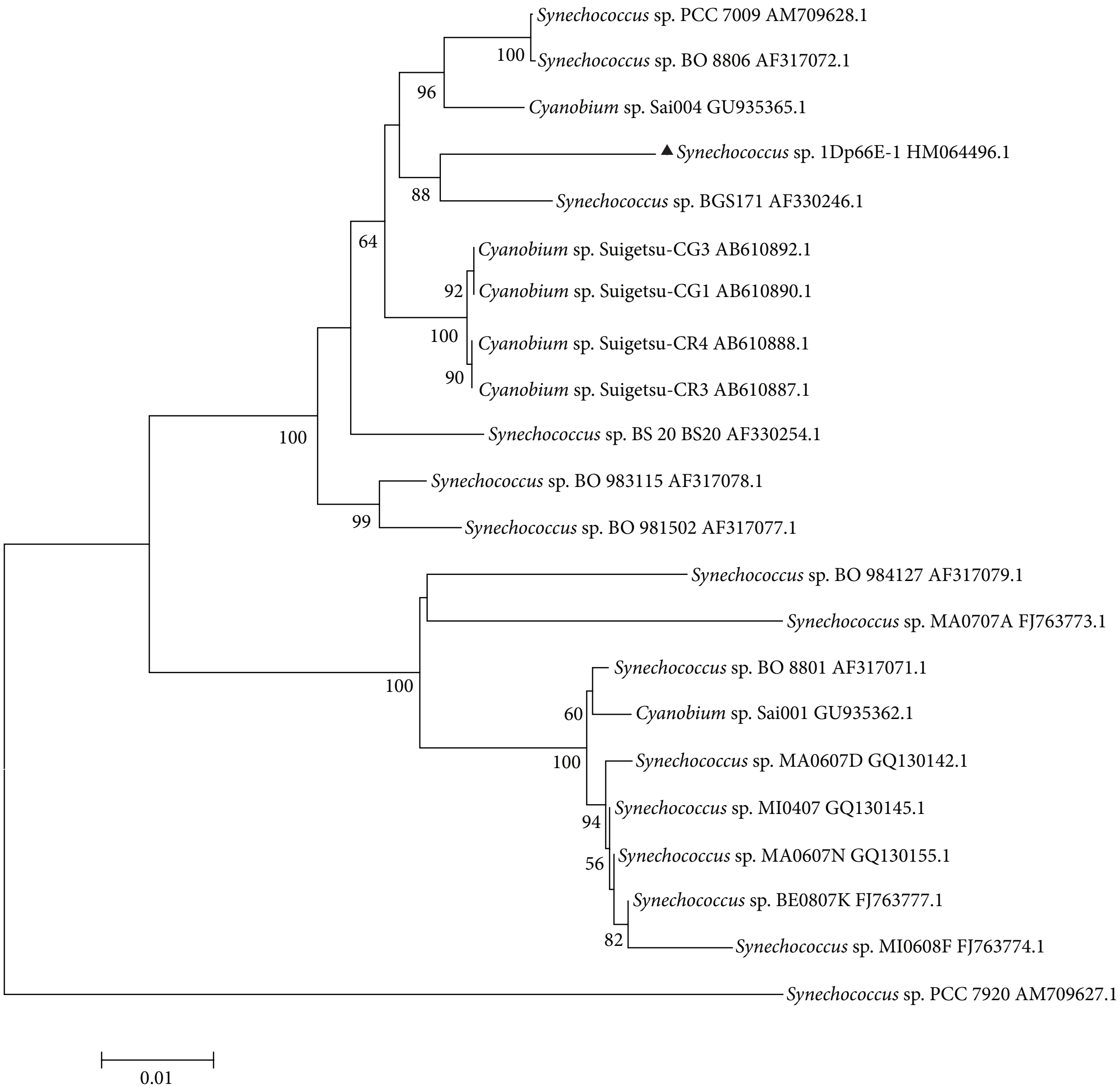

Synechococcus sp. PCC 7009 AM709628.1
100
Synechococcus sp. BO 8806 AF317072.1
96
Cyanobium sp. Sai004 GU935365.1
▲*Synechococcus* sp. 1Dp66E-1 HM064496.1
88
Synechococcus sp. BGS171 AF330246.1
64
Cyanobium sp. Suigetsu-CG3 AB610892.1
92
Cyanobium sp. Suigetsu-CG1 AB610890.1
100
Cyanobium sp. Suigetsu-CR4 AB610888.1
90
Cyanobium sp. Suigetsu-CR3 AB610887.1
100
Synechococcus sp. BS 20 BS20 AF330254.1
Synechococcus sp. BO 983115 AF317078.1
99
Synechococcus sp. BO 981502 AF317077.1
Synechococcus sp. BO 984127 AF317079.1
Synechococcus sp. MA0707A FJ763773.1
Synechococcus sp. BO 8801 AF317071.1
100
60
Cyanobium sp. Sai001 GU935362.1
100
Synechococcus sp. MA0607D GQ130142.1
Synechococcus sp. MI0407 GQ130145.1
94
56
Synechococcus sp. MA0607N GQ130155.1
Synechococcus sp. BE0807K FJ763777.1
82
Synechococcus sp. MI0608F FJ763774.1
Synechococcus sp. PCC 7920 AM709627.1
0.01

Figure 4: Phylogenetic relationships of *Synechococcus* sp. 1Dp66E-1 inferred under the neighbor-joining (NJ) criterion (MEGA 5.1 [14]) from gene for 16S rRNA, partial sequence; 16S-23S rRNA intergenic spacer, complete sequence; and 23S rRNA, partial sequence information. The numbers at the nodes indicate the level of bootstrap support based on neighbor-joining analysis of 1000 resampled datasets; only values higher than 50% are given. The scale bar indicates 0.01 substitutions per nucleotide position.

significance of the obtained dendrograms was calculated by bootstrap analysis by generation of 1000 permutations.

3. Results and Discussion

3.1. Morphological and Ultrastructural Characteristics and Identification of Cyanobacterial Isolates. Cyanobacterium 1Dp66E-1 was characterized by short rods (Table 2). These cells contained 2-3 thylakoids located in parallel to a cell surface. In the center of the cell there were a nucleoid, ribosomes, and carboxysomes (Figure 1). Visible polymers have been identified as granules of poly-β-hydroxybutyrate (Phb) and polyphosphate. Cyanophycin granules were absent. As it was possible to see after negative staining, cells of 1Dp66E-1 were characterized by a rugulose relief of the cell surface. Similar surface type has been observed in cases of no phototrophic Gram-negative bacteria [15] and of cyanobacteria *Synechococcus* sp. PCC 6301 and *Synechococcus* sp. PCC 7942 [16]. On ultrathin sections the outer leaflet of the outer membrane of the cell wall looked like a structure of high electron density, and it was merged with an additional layer consisting of orderly packed subunits. This additional layer was similar to bacterial S-layer glycoprotein. The cyanobacterium 1Dp66E-1 had characteristics typical to genus of *Synechococcus*.

The study by light and electron microscopy revealed that cyanobacterium 2Dp86E was filamentous (Table 2; Figure 2). Trichomes were often bound into bundles and were surrounded by layered common sheathes, which had in local

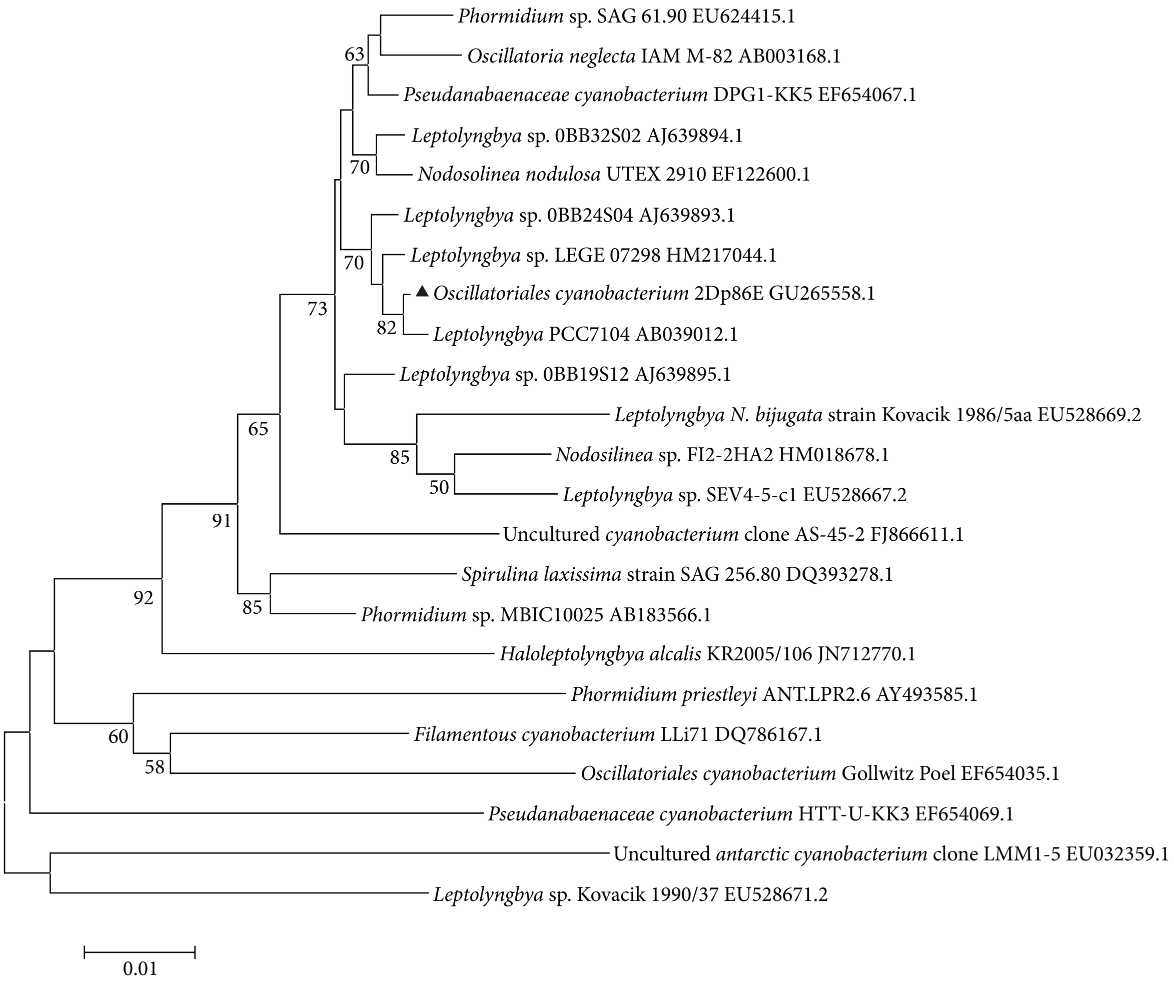

Phormidium sp. SAG 61.90 EU624415.1
Oscillatoria neglecta IAM M-82 AB003168.1
Pseudanabaenaceae cyanobacterium DPG1-KK5 EF654067.1
Leptolyngbya sp. 0BB32S02 AJ639894.1
Nodosolinea nodulosa UTEX 2910 EF122600.1
Leptolyngbya sp. 0BB24S04 AJ639893.1
Leptolyngbya sp. LEGE 07298 HM217044.1
▲ Oscillatoriales cyanobacterium 2Dp86E GU265558.1
Leptolyngbya PCC7104 AB039012.1
Leptolyngbya sp. 0BB19S12 AJ639895.1
Leptolyngbya N. bijugata strain Kovacik 1986/5aa EU528669.2
Nodosilinea sp. FI2-2HA2 HM018678.1
Leptolyngbya sp. SEV4-5-c1 EU528667.2
Uncultured cyanobacterium clone AS-45-2 FJ866611.1
Spirulina laxissima strain SAG 256.80 DQ393278.1
Phormidium sp. MBIC10025 AB183566.1
Haloleptolyngbya alcalis KR2005/106 JN712770.1
Phormidium priestleyi ANT.LPR2.6 AY493585.1
Filamentous cyanobacterium LLi71 DQ786167.1
Oscillatoriales cyanobacterium Gollwitz Poel EF654035.1
Pseudanabaenaceae cyanobacterium HTT-U-KK3 EF654069.1
Uncultured antarctic cyanobacterium clone LMM1-5 EU032359.1
Leptolyngbya sp. Kovacik 1990/37 EU528671.2
63
70
70
82
73
85
50
65
91
85
92
60
58
0.01

Figure 5: Phylogenetic relationships of *Oscillatoriales* cyanobacterium 2Dp86E inferred under the neighbor-joining (NJ) criterion (MEGA 5.1) from gene for 16S rRNA, partial sequence; 16S-23S rRNA intergenic spacer, complete sequence; and 23S rRNA, partial sequence information. The numbers at the nodes indicate the level of bootstrap support based on neighbor-joining analysis of 1000 resampled datasets; only values higher than 50% are given. The scale bar indicates 0.01 substitutions per nucleotide position.

Table 2: Morphological properties of cyanobacteria isolates.

Isolate name	Morphology, type of cell division, and differentiation	Subsection (order)*
1Dp66E-1	Unicellular, binary fission by constriction, cells sticks shaped 0.6–0.8 μm wide and 1.5–2 μm long, unsheathed.	I (*Chroococcales*)
2Dp86E	Filamentous, binary fission in one plane by septum formation, no cell differentiation, ensheathed, cells barrel or cylindrical shaped 1.5–2.5 μm long and 1.0–1.5 μm wide.	III (*Oscillatoriales*)
10Dp66E	Filamentous, binary fission in one plane by septum formation, cells round shaped 3.0–3.5 μm in diameter and are able to differentiate hormogonia and intercalary heterocysts and akinetes. Vegetative trichomes have their own sheaths and are joined by colonial mucus.	IV (*Nostocales*)

*Classification was performed according to Castenholz [25].

areas sleeve-like bulges (Figure 2). In the structure of many sheaths, there were revealed electron transparent sections. The sheaths of neighboring trichomes could be closely contacted. The end cells in trichomes usually are smaller and narrower at the terminal parts. Cells of this isolate had 4–6 thylakoids, located on the periphery of the cytoplasm and in parallel to a cell surface. Nucleoid, carboxysomes, polyphosphate, and Phb-like granules were located in the cell center (Figure 2). Clusters of ribosomes were concentrated around the nucleoid. Cyanophycin granules were rare.

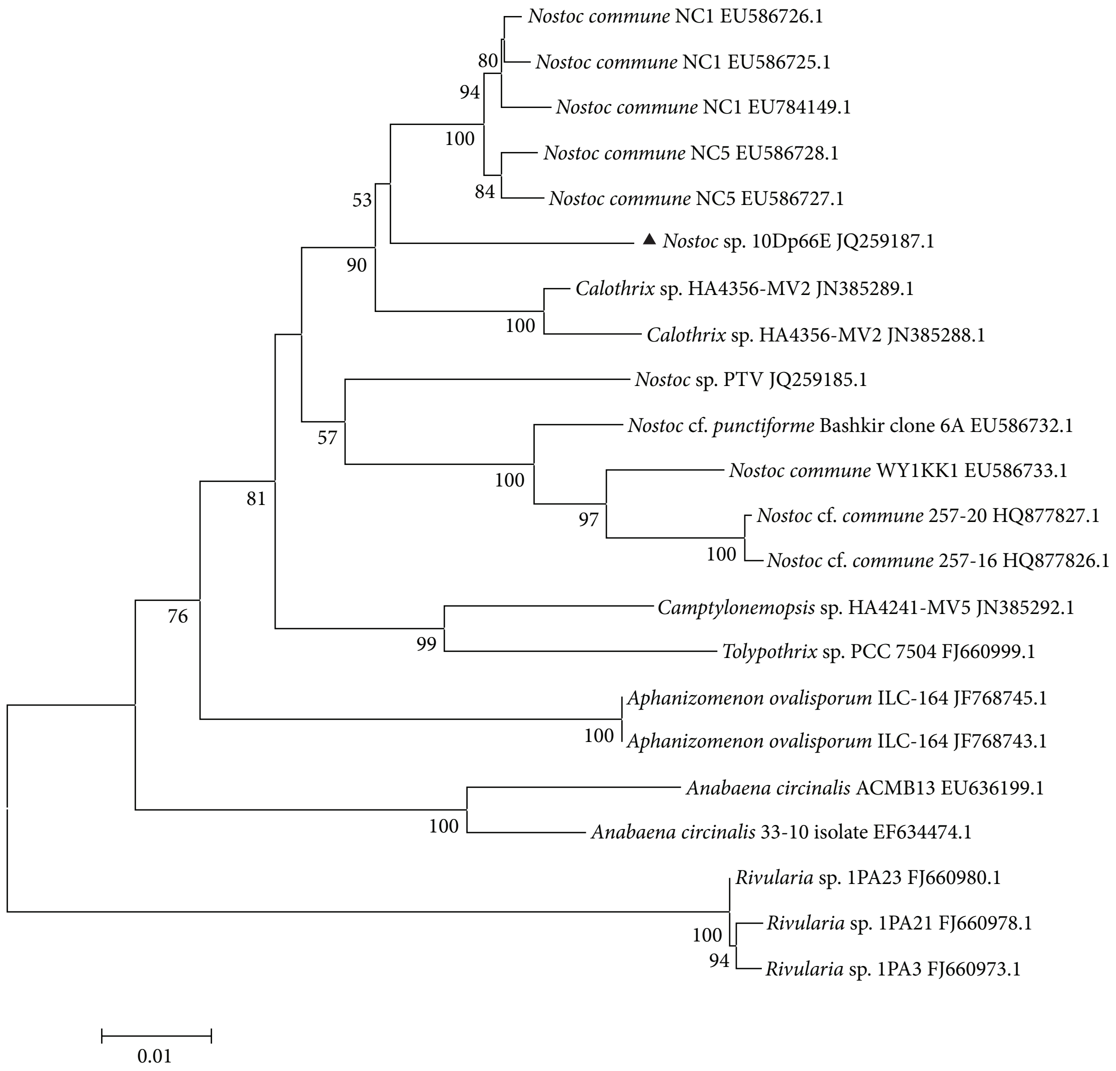

Nostoc commune NC1 EU586726.1
80
Nostoc commune NC1 EU586725.1
94
Nostoc commune NC1 EU784149.1
100
Nostoc commune NC5 EU586728.1
84
Nostoc commune NC5 EU586727.1
53
▲ Nostoc sp. 10Dp66E JQ259187.1
90
Calothrix sp. HA4356-MV2 JN385289.1
100
Calothrix sp. HA4356-MV2 JN385288.1
Nostoc sp. PTV JQ259185.1
57
Nostoc cf. punctiforme Bashkir clone 6A EU586732.1
100
Nostoc commune WY1KK1 EU586733.1
81
97
Nostoc cf. commune 257-20 HQ877827.1
100
Nostoc cf. commune 257-16 HQ877826.1
Camptylonemopsis sp. HA4241-MV5 JN385292.1
76
99
Tolypothrix sp. PCC 7504 FJ660999.1
Aphanizomenon ovalisporum ILC-164 JF768745.1
100
Aphanizomenon ovalisporum ILC-164 JF768743.1
Anabaena circinalis ACMB13 EU636199.1
100
Anabaena circinalis 33-10 isolate EF634474.1
Rivularia sp. 1PA23 FJ660980.1
100
Rivularia sp. 1PA21 FJ660978.1
94
Rivularia sp. 1PA3 FJ660973.1
0.01

Figure 6: Phylogenetic relationships of *Nostoc* sp. 10Dp66E inferred under the neighbor-joining (NJ) criterion (MEGA 5.1) from gene for 16S rRNA, partial sequence; 16S-23S rRNA intergenic spacer, complete sequence; and 23S rRNA, partial sequence information. The numbers at the nodes indicate the level of bootstrap support based on neighbor-joining analysis of 1000 resampled datasets; only values higher than 50% are given. The scale bar indicates 0.01 substitutions per nucleotide position.

Despite the fact that this strain was able to greatly accumulate phycocyanin [7], phycobilisomes on ultrathin sections were not detected due to the high density of the surrounding cytoplasm. From the combination of morphological and ultrastructural signs of cyanobacteria 2Dp86E had similarities with the representatives of the genus *Leptolyngbya*, which includes groups of filamentous cyanobacteria in LPP-group B that exhibit very thin trichomes (<3 μm) [17, 18].

Isolate 10Dp66E was a filamentous cyanobacterium (Figure 3, Table 2). Depending on the phase of the culture development in the colonies of cyanobacteria dominated hormogonia, long-row vegetative trichomes with intercalary heterocysts and akinetes, or rounded clusters formed by twisted trichomes. Many chains of cells were "dressed" by one- or multilayer fibrillar sheaths (Figure 3). Colonial mucus and extracellular matrix of clusters were composed from loosely packed reticular-fibrous and granular materials. Peripheral layer of mucous extracellular matrix was compacted and was common for the entire cluster. Bacteria of different morphotypes were localized in colonial mucus, in the intercellular matrix, and in sheaths of trichomes. A feature of this cyanobacterium is the differentiation of heterocysts even in presence of combined nitrogen in the medium. Photosynthetic apparatus of cyanobacterium 10Dp66E was represented by randomly distributed thylakoids in the cytoplasm. Intrathylakoid space could be increased and in some cases was filled by mesh substance. Nucleoid, ribosomes, and carboxysomes are located between thylakoids. In the 10Dp66E, cells were accumulated with glycogen, Phb, cyanophycin, protein granules, and lipid globules

Table 3: BLAST results obtained by querying the 16S-23S rRNA gene cluster of *Synechococcus* sp. 1Dp66E-1 with GenBank and geographical and ecological origins of the hits.

Closest GenBank relative	GenBank number	Query coverage %	Score %	Identity %	*E*-value	Origin of the strain and reference
Cyanobium sp. Suigetsu-CG3	AB610892.1	99	3469	95	0	Picocyanobacteria isolated from the surface layer permanent halocline of the saline (from 0.4% to 1.4%) meromictic Lake Suigetsu, (35°35′N, 135°52′E, coast of the Japan Sea in Fukui Prefecture, Japan) [26].
Synechococcus sp. PCC 7009	AM709628.1	99	3431	94	0	Originated from CA, USA. It was placed then in PCC collection.
Synechococcus sp. BGS171	AF330246.1	98	3422	93	0	Picocyanobacteria isolated from the fresh water Lake Constance (47°39′N 9°19′E) Europe [27].
Synechococcus sp. BO8806	AF317072.1	98	3388	94	0	Picocyanobacteria isolated from the fresh water Lake Balaton (46°50′N 17°44′E) Europe [27].
Synechococcus sp. BO983115	AF317078.1	98	3377	94	0	Picocyanobacteria isolated from the fresh water Lake Balaton (46°50′N 17°44′E) Europe [27].
Synechococcus sp. BO981502	AF317077.1	98	3341	94	0	Picocyanobacteria isolated from the fresh water Lake Balaton (46°50′N 17°44′E) Europe [27].
Synechococcus sp. BS20	AF330254.1	98	3321	94	0	Picocyanobacteria isolated from the Bornholm Sea (55°07′N 14°55′E (salinity 9 g L^{-1})), Baltic Sea, Europe [27].
Cyanobium sp. Sai004	GU935365.1	97	3296	94	0	Cyanobacteria isolated from the drinking water of the dam in the Saidenbach Saxony (Germany), Europe [Schumann, unpublished].

(Figure 3). By morphological criteria and ultrastructural features, cyanobacterium 10Dp66E was similar to representatives of the genus *Nostoc*.

3.2. Phylogenetic Analysis of the Ribosomal RNA Operons. To identify and to determine the phylogenetic positions of the new cyanobacterial isolates purified from native samples of the sublittoral hydroid *D. pumila,* we used 16S ribosomal RNA gene cluster as a molecular marker. Cloning and sequencing of the DNA fragment, containing 16S ribosomal RNA gene and 16S-23S ribosomal RNA intergenic spacer sequences, have been performed for all three isolated strains using appropriate primers (Table 1). Comparison of rRNA gene cluster sequence (2277 bp) of the cyanobacterium 1Dp66E-1 revealed that 1Dp66E-1 shows the highest similarity with several strains of *Synechococcus* (Table 3). Phylogenetic analysis (Figure 4) revealed that the cyanobacterium 1Dp66E-1 forms the unique evolutionary branch and belongs to a cluster of picocyanobacteria *Synechococcus*, which includes fresh- and seawater strains. One group of closely related organisms had been isolated from freshwater European lakes (Table 3). The closest relatives of the 1Dp66E-1 are *Synechococcus* sp. BO981502, isolated from Lake Balaton, and *Synechococcus* sp. BGS171, isolated from Lake Constance (Bodensee). Another strain of the genus *Synechococcus* (*Synechococcus* sp. BS20) was isolated from the Baltic Sea near the Danish island of Bornholm. The strain *Synechococcus* sp. PCC 7009 was originated from the USA, whereas the *Cyanobium* sp. Suigetsu-CG3 was purified from a lake in Japan (Table 3). Thus, the majority of *Synechococcus* strains that are closely related to 1Dp66E-1 have been isolated from marine and freshwater habitats of Europe, as well of the USA and Japan. These organisms exist in the form of plankton. However, the strain of *Synechococcus* sp. PCC 7009 was clustered with chromatophore (photosynthetic endosymbiont) of the freshwater filose amoeba *Paulinella chromatophora*. This endosymbiont has cyanobacterial origin, and it is closely related to free-living *Prochlorococcus* and *Synechococcus* species [19]. Here, we demonstrate that the picocyanobacterium *Synechococcus* sp. 1Dp66E-1 is associated with the invertebrate of White Sea.

DNA fragment of 1732 bp (GU265558.1) containing almost complete sequence of the genes coding for 16S rRNA, internal transcribed spacer 16S-23S rRNA, and part of the gene for 23S rRNA of the cyanobacterium 2Dp86E (GeneBank accession no. GU265558.1) has the highest similarity with analogous sequences of different *Leptolyngbya*

Table 4: BLAST results obtained by querying the 16S-23S rRNA gene cluster of *Oscillatoriales* cyanobacterium 2Dp86E with GenBank, and geographical and ecological origins of the hits.

Closest GenBank relative	GenBank number	Query coverage %	Score %	Identity %	E-value	Origin of the strain and reference
Leptolyngbya PCC7104	AB039012.1	66	2035	99	0	Long Island, NY, USA, Montauk Point [28].
Leptolyngbya sp. LEGE 07298	HM217044.1	65	2013	99	0	Cyanobacteria from three Portuguese temperates [29].
Leptolyngbya sp. 0BB24S04	AJ639893.1	67	2082	99	0	Bubano Basin, Imola, Italy [30].
Leptolyngbya sp. 0BB19S12	AJ639895.1	67	2046	99	0	Bubano Basin, Imola, Italy [30].
Leptolyngbya sp. 0BB32S02	AJ639894.1	67	2033	99	0	Bubano Basin, Imola, Italy [30].
Nodosilinea sp. FI2-2HA2	HM018678.1	96	2596	95	0	Fort Irwin, California, Mojave Desert, soil [31].
Leptolyngbya sp. Kovacik 1986/5 aa	EU528669.2	99	2475	92	0	Poland: Lake Piaseczno littoral region [Casamatta, unpublished].
Leptolyngbya sp. SEV4-5-c1	EU528667.2	76	2152	97	0	Desert soil [Casamatta, unpublished].
Nodosolinea nodulosa UTEX 2910	EF122600.1	64	1954	99	0	South China Sea (marine), plankton tow, 10 m depth [32].

Table 5: BLAST results obtained by querying the 16S-23S rRNA gene cluster of *Nostoc* sp. 10Dp66E with GenBank and geographical and ecological origins of the hits.

Closest GenBank relative	GenBank number	Query coverage %	Score %	Identity %	E-value	Origin of the strain	Reference
Nostoc commune NC1 clone 10	EU586726.1	92	2693	94	0	Unknown	N. commune Vaucher exBornet et Flahault [Johansen J. R. et al., unpublished].
Nostoc commune NC5 clone 11	EU586728.1	92	2666	94	0	Unknown	N. commune Vaucher exBornet et Flahault [Johansen J. R. et al., unpublished].
Nostoc commune NC1	EU784149.1	92	2681	94	0	Unknown	Isolated from meadow soil, close Vrbensky pond, Ceske Budejovice, Czech Republic [Rehakova K., unpublished].
Calothrix sp. HA4356-MV2	JN385289.1	99	2634	92	0	Cave wall scraping, Maniniholo cave near Haena, Hawaii	[33]
Calothrix sp. HA4356-MV2	JN385288.1	81	2634	96	0	Isolated from the island of Oahu (21°28′N157°59′W), Hawaii	[33]

strains (Table 4), and it is clustered with the representatives of the genus *Leptolyngbya* (Figure 5). The latter have been isolated from freshwater lakes in Italy and Poland (European lakes), estuaries of benthic temperate of Portugal, and from the desert soils of North America (Table 4). All these closely related organisms of the 2Dp86E have been isolated from freshwater and soil habitats. There is only one representative of the marine species which is related to the 2Dp86E. Until now, eight species of *Leptolyngbya* have been found in White Sea: they have been isolated from fouling stones, or, as *Leptolyngbya mucicola*—the endofit from colonies of *Rivularia coadunata* [20]. In addition, several *Leptolyngbya* strains have been identified in White Sea, which inhabit areas of varying salinity [21]. However, all mentioned strains have not been analyzed with molecular methods.

Analysis of the nucleotide sequence (1902 bp) of the 16S rRNA gene cluster of the *Nostoc* sp. 10Dp66E (JQ259187.1) revealed that it has no identity to any of previously studied organisms. Cyanobacterium 10Dp66E has similarity with several strains of *Nostoc commune* (Table 5). It forms the unique branch and belongs to a cluster of *Nostoc*, with the closest relative of *Nostoc commune* isolates (Figure 6). There are only two strains of *Nostoc*, namely, *Nostoc commune* Vaucher ex Bornet et Flahault and *Nostoc zetterstedtii* Aresch. ex Bornet et Flahault, that had been detected in White Sea so far [20]. No information is available on the molecular typing of these strains. However, it is possible that *Nostoc commune* Vaucher ex Bornet et Flahault might represent the same strain as *Nostoc commune* NC1 clone 10 and *Nostoc commune* NC5 clone 11, that were defined in the GenBank (Table 5). *Nostoc commune* NC1 (EU784149.1) was isolated from meadow soils, close to Vrbensky ponds, Ceske Budejovice, Czech Republic.

Cyanobacterial mucous sheaths provide the specific econiches—environments for numerous heterotrophic bacteria. In this work, using 16S rRNA gene, we first identified the heterotrophic bacterium associated with *Nostoc* sp. 10Dp66E. This bacterium is a member of the genus *Gemmatimonas* (JX437625.1). The Gemmatimonadetes represents recently described bacterial group whose members are widespread in nature. These bacteria which are phylogenetically novel Gram-negative aerobic heterotrophs are capable of accumulating polyphosphates [22]. *Gemmatimonas* was found in associations with sponge [6].

4. Conclusions

Based on phenotypic and genotypic characteristics, we identified new cyanobacteria from association with the hydroid *Dynamena pumila* living in White Sea. Two of them form the unique branches in the corresponding phylogenetic trees. Among species clustered with the new cyanobacterial isolates there are fresh- and seawater strains from Europe, America, and Japan. The identified strains are mostly photoautotrophic, but some of them are diazotrophic bacteria. Future biochemical and genetic experiments will help to understand metabolic relations between these bacterial isolates and the hydroid *Dynamena pumila* in the association.

Acknowledgment

This work was supported by Grants nos. 09-04-00675 and 11-04-00774 of the Russian Foundation for Basic Research.

References

[1] E. J. Carpenter and R. A. Foster, "Marine cyanobacterial symbiosis," in *Cyanobacteria Symbiosis*, A. N. Rai, B. Bergman, and U. Rasmussen, Eds., pp. 11–17, Academic Kluwer, Dordrecht, The Netherlands, 2002.

[2] M. P. Lesser, C. H. Mazel, M. Y. Gorbunov, and P. G. Falkowski, "Discovery of symbiotic nitrogen-fixing cyanobacteria in corals," *Science*, vol. 305, no. 5686, pp. 997–1000, 2004.

[3] R. A. Foster, E. J. Carpenter, and B. Bergman, "Unicellular cyanobionts in open ocean dinoflagellates, radiolarians, and tintinnids: ultrastructural characterization and immuno-localization of phycoerythrin and nitrogenase," *Journal of Phycology*, vol. 42, no. 2, pp. 453–463, 2006.

[4] T. Romagnoli, G. Bavestrello, E. M. Cucchiari et al., "Microalgal communities epibiontic on the marine hydroid *Eudendrium racemosum* in the Ligurian Sea during an annual cycle," *Marine Biology*, vol. 151, no. 2, pp. 537–552, 2007.

[5] A. A. Venn, J. E. Loram, and A. E. Douglas, "Photosynthetic symbioses in animals," *Journal of Experimental Botany*, vol. 59, no. 5, pp. 1069–1080, 2008.

[6] M. W. Taylor, R. Radax, D. Steger, and M. Wagner, "Sponge-associated microorganisms: evolution, ecology, and biotechnological potential," *Microbiology and Molecular Biology Reviews*, vol. 71, no. 2, pp. 295–347, 2007.

[7] O. Gorelova, I. Kosevich, O. Baulina, T. Fedorenko, A. Torshkhoeva, and E. S. Lobakova, "Associations between the white sea invertebrates and oxygen-evolving phototrophic microorganisms," *Moscow University Biological Sciences Bulletin*, vol. 64, no. 1, pp. 16–22, 2009.

[8] O. A. Gorelova, O. I. Baulina, I. A. Kosevich, and E. S. Lobakova, "Associations between the White Sea colonial hydroid *Dynamena pumila* and microorganisms," *Journal of the Marine Biological Association of the UK*, vol. 93, no. 1, pp. 69–80, 2013.

[9] R. Y. Stanier, R. Kunisawa, M. Mandel, and G. Cohen-Bazire, "Purification and properties of unicellular blue-green algae (order Chroococcales)," *Bacteriological Reviews*, vol. 35, no. 2, pp. 171–205, 1971.

[10] E. S. Reynolds, "The use of lead citrate at high pH as an electron-opaque stain in electron microscopy," *The Journal of Cell Biology*, vol. 17, pp. 208–212, 1963.

[11] O. Koksharova, M. Schubert, S. Shestakov, and R. Cerff, "Genetic and biochemical evidence for distinct key functions of two highly divergent GAPDH genes in catabolic and anabolic carbon flow of the cyanobacterium *Synechocystis* sp. PCC 6803," *Plant Molecular Biology*, vol. 36, no. 1, pp. 183–194, 1998.

[12] D. Papaefthimiou, P. Hrouzek, M. A. Mugnai et al., "Differential patterns of evolution and distribution of the symbiotic behaviour in nostocacean cyanobacteria," *International Journal of Systematic and Evolutionary Microbiology*, vol. 58, no. 3, pp. 553–564, 2008.

[13] J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York, NY, USA, 2nd edition, 1989.

[14] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[15] L. O. Zwillenberg, "Electron microscopic features of gram-negative and gram-positive bacteria embedded in phosphotungstate," *Antonie van Leeuwenhoek*, vol. 30, no. 1, pp. 154–162, 1964.

[16] O. A. Gorelova, O. I. Baulina, U. Rasmussen, and O. A. Koksharova, "The pleiotropic effects of *ftn2* and *ftn6* mutations in cyanobacterium *Synechococcus* sp. PCC 7942: an ultrastructural study," *Protoplasma*, 2013.

[17] R. Rippka, J. Deruelles, and J. B. Waterbury, "Generic assignments, strain histories and properties of pure cultures of cyanobacteria," *Journal of General Microbiology*, vol. 111, no. 1, pp. 1–61, 1979.

[18] L. Bruno, D. Billi, S. Bellezza, and P. Albertano, "Cytomorphological and genetic characterization of troglobitic *Leptolyngbya* strains isolated from Roman hypogea," *Applied and Environmental Microbiology*, vol. 75, no. 3, pp. 608–617, 2009.

[19] B. Marin, E. C. M. Nowack, G. Glöckner, and M. Melkonian, "The ancestor of the *Paulinella chromatophore* obtained a carboxysomal operon by horizontal gene transfer from a Nitrococcus-like γ-proteobacterium," *BMC Evolutionary Biology*, vol. 7, article 85, 2007.

[20] R. N. Beliakova, "Cyanophyta reservati kandalakshnsis," in *Novitates Systematicae Plantarum Non Vascularium*, K. L. Vinogradov, Ed., vol. 31, pp. 9–16, Nauka, St. Peterburg, Russia, 1996.

[21] A. Ulanova, *Algae of ponds with unstable salinity coasts of the White and Barents Seas [Ph.D. thesis]*, St. Peterburg, Russia, 2003.

[22] H. Zhang, Y. Sekiguchi, S. Hanada et al., "*Gemmatimonas aurantiaca* gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum *Gemmatimonadetes* phyl. nov," *International Journal of Systematic and Evolutionary Microbiology*, vol. 53, no. 4, pp. 1155–1163, 2003.

[23] U. Nübel, F. Garcia-Pichel, and G. Muyzer, "PCR primers to amplify 16S rRNA genes from cyanobacteria," *Applied and Environmental Microbiology*, vol. 63, no. 8, pp. 3327–3332, 1997.

[24] A. Taton, S. Grubisic, E. Brambilla, R. De Wit, and A. Wilmotte, "Cyanobacterial diversity in natural and artificial microbial mats of Lake Fryxell (McMurdo Dry Valleys, Antarctica): a morphological and molecular approach," *Applied and Environmental Microbiology*, vol. 69, no. 9, pp. 5157–5169, 2003.

[25] R. W. Castenholz, "Phylum BX. Cyanobacteria. Oxygenic photosynthetic bacteria," in *Bergey's Manual of Systematic Bacteriology*, D. R. Boone and R. W. Castenholz, Eds., vol. 1, pp. 473–599, 2nd edition, 2001.

[26] K. Ohki, K. Yamada, M. Kamiya, and S. Yoshikawa, "Characterization of picocyanobacteria isolated from the halocline of the saline meromictic lake, Lake Suigetsu, Japan," in *Proceedings of the 13th International Symposium on Phototrophic Prokaryotes Montréal*, p. 50, August 2009.

[27] A. Ernst, S. Becker, U. I. A. Wollenzien, and C. Postius, "Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis," *Microbiology*, vol. 149, no. 1, pp. 217–228, 2003.

[28] T. Ishida, M. M. Watanabe, J. Sugiyama, and A. Yokota, "Evidence for polyphyletic origin of the members of the orders of Oscillatoriales and Pleurocapsales as determined by 16S rDNA analysis," *FEMS Microbiology Letters*, vol. 201, no. 1, pp. 79–82, 2001.

[29] V. R. Lopes, V. Ramos, A. Martins et al., "Phylogenetic, chemical and morphological diversity of cyanobacteria from Portuguese temperate estuaries," *Journal of Marine Environmental*, vol. 73, pp. 7–16, 2012.

[30] B. Castiglioni, E. Rizzi, A. Frosini et al., "Development of a universal microarray based on the ligation detection reaction and 16S rRNA gene polymorphism to target diversity of cyanobacteria," *Applied and Environmental Microbiology*, vol. 70, no. 12, pp. 7161–7172, 2004.

[31] R. B. Perkerson, J. R. Johansen, L. Kovacik, J. Brand, J. Kastovsky, and D. A. Casamatta, "A unique Pseudanabaenalean (Cyanobacteria) genus *Nodosilinea* gen. nov. based on morphological and molecular data," *Journal of Phycology*, vol. 47, no. 6, pp. 1397–1412, 2011.

[32] Z. Li and J. Brand, "*Leptolyngbya Nodulosa* sp. nov. (Oscillatoriaceae), a subtropical marine cyanobacterium that produces a unique multicellular structure," *Phycologia*, vol. 46, no. 4, pp. 396–401, 2007.

[33] M. A. Vaccarino and J. R. Johansen, "Scytonematopsis contorta sp. nov. (Nostocales), a new species from the Hawaiian Islands," *Fottea*, vol. 11, no. 1, pp. 149–161, 2011.