Molecular approaches to the study of ecdysozoan evolution

Omar Rota Stabelli

Research Department of Genetics, Evolution and Environment

UCL

Submitted for the Degree of Doctor of Philosophy

September 2009

This work is dedicated to…

"Hurricane Pete" Pietro who did everything in his power to prevent me from writing this thesis and Maura who patiently took care of both.

Max, who gave me the opportunity to study the fabulous world of arthropods and Davide, who is giving me an other opportunity.

Racco and Marie which are sadly gone.

Declaration

I, Omar Rota Stabelli, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, this has been indicated in the thesis. Omar Rota Stabelli

Abstract

The Ecdysozoa is a large clade of animals comprising the vast majority of living species and some of the most studied invertebrate models, including fruitflies and nematodes. Some of the relationships between major ecdysozoan groups remain uncertain, however, undermining comparative studies and impairing our understanding of their evolution. One hotly debated problem is the position of myriapods which have been recently grouped according to molecules with chelicerates and not with insects and crustaceans as predicted by morphological evidence. Other disputed problems are the position of tardigrades, the position of hexapods within the crustaceans as well as the mutual affinities of the nematodes and priapulid worms. Molecular systematics of the ecdysozoans is complicated by rapid divergence of the main lineages (possibly evidenced in the Cambrian explosion) followed by a subsequent long period of evolution. This may have resulted in a dilution of the historical phylogenetic signal and an increased likelihood of encountering systematic errors of tree reconstruction. This problem is exacerbated by many lineages being poorly represented in current molecular datasets, as sequencing efforts have been biased toward lab models and economically relevant species.

In order to overcome problems of systematic error, I have assembled various large mitochondrial and phylogenomic datasets, including new data from undersampled tardigrades, onychophorans and especially myriapods. I analysed these datasets using the most recent evolutionary models. I have developed two new models in order to describe the evolutionary processes of metazoan mitochondrial proteins more accurately. My analyses of multiple datasets suggest that the grouping of myriapods plus chelicerates found by previous authors is likely to be the result of systematic errors; I find support for a closer relationships between myriapods and a group of insects plus crustaceans (the Mandibulata hypothesis). My analyses also support a paraphyletic origin of Cycloenuralia (nematodes and priapulids) and a sister group relationships between tardigrades, onychophorans and euarthropods in accordance with a single origin of legged ecdysozoans, the Panarthropoda. Finally, results support a monophyletic group of hemimetabolan insects. The majority of the results reconcile molecules and morphology, while others shade new light onto arthropod systematics. The evolutionary implications of these systematic findings as well as methodological advances are discussed.

Thanks to…

Max Telford for his great supervision, guidance and help. For let me go independently in my research, although being always there to prevent me from going too far and get lost in what he once defined a “multidimensional parametric nightmare”.

Andrew Economou, Josh Coulcher and Niko Pripc for morphological and evo-devo hard-core discussions.

Sarah Bourlat for patiently teaching me some of the mysteries of the wet lab.

Hervè Philippe and Henner Brinkmann (the lords of the LBA) for their experience, their teaching and especially their being a lifestyle example.

Greg Edgcombe and Alison Daley for trusting a young molecular phylogeneticist and for their incredible enthusiasm for everything which possess jointed appendages.

Dennis Lavrov and Mark Blaxter for their trust and patience.

David Horner and Peter Foster for their guidance, collaboration and patience when I was an undergraduate. I hope to pay back some day.

Davide Pisani for hosting me in his lab, for the many ideas and for the “cladistic” discussion.

Stuart Longhorn for fruitful suggestions over insect systematics and some proof reading

Maura Parazzoli for proof reading.

The ZOONET fellows, PIs and colleagues for the great science and the good time together

The Marie Curie Action for funding my studies.

Daniel Papillon, for teaching me that there is always an alternative way.

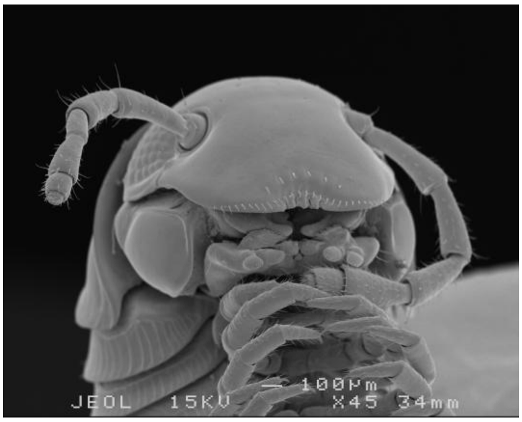
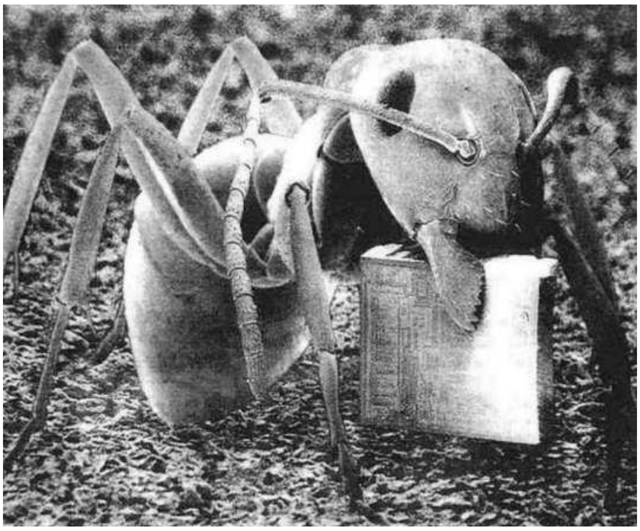


Table of Contents

List of figures and tables



Chapter 1

Introduction

1.1 The Ecdysozoa and the demise of a systematic establishment.

The Ecdysozoa is a widely recognised clade of moulting animals comprising, among others, insects, crustaceans and the nematode worms (Aguinaldo et al. 1997). Doubtless, it represents the most diverse and successful group of animals on earth. It has been estimated that the insects alone, which account for almost 80% of the documented animal biodiversity, may count as many as 10 million species (Chapman 2005, Novotny et al 2007). Above and beyond this, the majority of zooplankton species are crustaceans, making the latter "the most abundant type of multicellular animal on earth" in terms of number of individuals (Martin & Davis 2001). This primacy has probably to be shared with the nematode worms, which parasitize most multicellular creatures as well inhabit probably every soil and every body of water, to the extent that they account for 90 % of animal life on the sea-floor (Atkinson 1973).

The Ecdysozoa, and arthropods in particular, successfully adapted to all ecological niches from ocean trenches to fresh waters, from remote tropical caves to the Polar regions. Throughout their evolutionary history the Ecdysozoa developed an incredible variety of body forms and underwent extensive size variation. This is well exemplified by the crustaceans, which include planktonic forms of less then a millimeter in total length to the japanese spider crab with its four meter leg span. Some of the most striking varieties of Ecdysozoa have probably been irremediably lost as a consequence of extinctions, particularly the soft bodied species, that hardy fossilise. Many ecdysozoan fossils have, however, been found, such as the giant *Arthropleurid*, a meters long myriapod which crawled the carbonifeorous forests, the enigmatic *Anomalocaris*, which dominated Cambrian seas and probably preyed on trilobites, *Opabinia* and *Hallucigenia*, some of the most puzzling fossils ever discovered. These and many other fossils, especially from the Burgess Shale and the Chengjiang lagerstätte, suggest that the extant ecdysozoan diversity is just the tip of what has been produced by numerous, failed adaptive attempts.

1.1.1 The eight ecdysozoan phyla

There are eight extant ecdysozoan phyla: the Arthropoda, Tardigrada, Onychophora, Nematoda, Nematomorpha, Loricifera, Priapulida and Kinorhyncha. They possess extremely different body-plans and unique morphological features, but they can be at first sight divided in two groups on the basis of their bodyplan.

Arthropods, onychophorans and tardigrades (depicted in figure 1.1 A, B and C) possess a distinctive "arthropod-like" bodyplan to the extent that they have been tentatively grouped in the Panarthopoda clade (Nielsen 2001). This group is characterised by a segmented coelomated body bearing paired, ventrolateral walking appendages. The naturalness of this group is further reinforced by the parasegmental expression of the segment polarity gene *engrailed* (Gabriel and Goldestein 2007).

More in details, the arthropods, in some cases named Euarthropoda, are probably the most diverse animals on earth (Nielsen 2001). They have adapted to almost all niches on the planet, and invaded the continental landmasses independently at least three times, with the arachnids, myriapods and insects. Arthropods are characterised by jointed appendages (hence their name) and a hard exoskeleton. There are four main arthropod groups: the hexapods (including the insects), the myriapods (e.g. millipedes and centipedes), crustaceans (e.g. lobsters and woodlice) and chelicerates (e.g. arachnids and the horseshoe crabs). After many years of debate, a consensus has emerged that these four sub-phyla (or classes) plus the extinct Trilobita form a monophyletic group called the Euarthropoda (see Figure 1.2B and section 1.1.3 for details). Each group, however, possesses a typical body plan with specific tagmosis (compare the copepod and the centipede in figure 1.1A), suggesting that versatility of modularity was a key aspect of arthropod evolution and probably a major contributor to their success (Yang 2001).

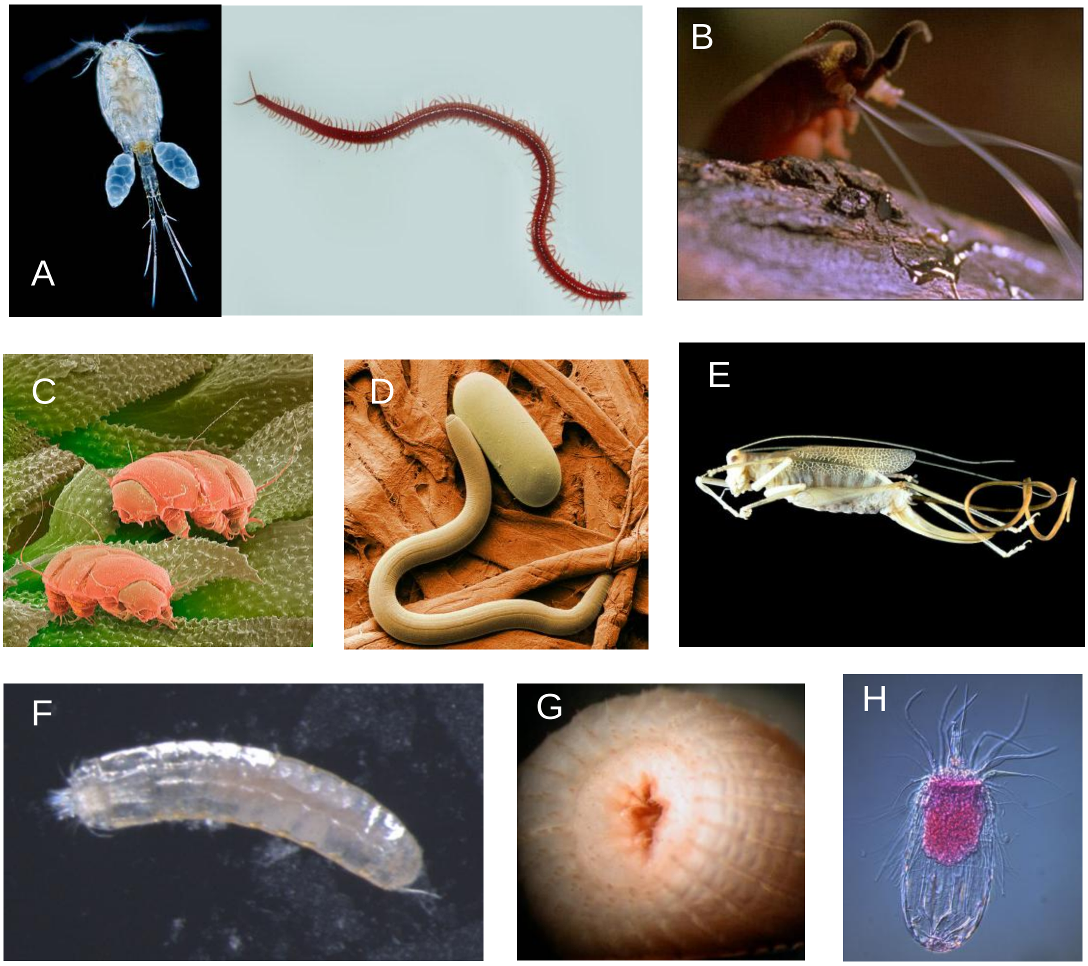


Figure1.1. The eight ecdysozoan phyla. In figure A two examples of arthropods, a copepode on the left and the centipede *Strigamia maritima* on the right. B: an onychophoran squirting adhesive slime to a prey. C: false colour electron microscopy of two tardigrades. D: electron microscopy of a soybean nematode and its egg. E: a nematomorph extruding from a cricket. F: a kinorhynch with the introvert partially everted. G: detail of the fully everted introvert of a priapulid worm. H: a loriciferans. All images are from wikicommons except for A which is from www.nathistoc.bio.uci.edu, B from www.news.bbc.co.uk, C from www.focus.it, F is from the author of this thesis and H is from www.tiefsee.senckenberg.de.

Onychophorans, literally "claw-bearer", are fascinating predators of about 10 cm in length. Their cuticle is covered with tiny scales, giving them a velvety appearance and their common name, velvet worms. They are limited to humid environments such as tropical forests (Peripatidae family) and temperate austral regions (Peripatopsidae family), because their respiratory tracheae do not close and may lead the animal to desiccation. Onychophorans possess a pair of antennae and typical conical walking appendages, which are unjointed and bear retractable sclerotised claws. One interesting characteristic of onychophorans are the oral papillae, two glands that squirt a sticky slime used to immobilize prey (as depicted in figure 1.1B).

Tardigrades, also known as water-bears and literally “slow-walker” due to their reminiscing bear’s gait, are tiny creatures of up to 1 mm. They are ubiquitous animals, but they need moisture for living. They can, however, survive any environmental conditions through cryptobiosis, usually by loosing 99% of their water and changing their body structure. It has been reported that tardigrades can survive exposure to almost absolute zero (-272 C), pressure of 6 times the deepest sea and lethal radiations in the outer space (Ingemar-Jonsson 2008). Similarly to onychophorans, tardigrades posses unjointed clawed appendages (typically 4 pairs).

The second group comprises Nematodes, nematomorphs, and the “gloriously obscure marine worms” kinorhynchs, priapulids and loriciferans (depicted in figure 1.1D to H) and is characterised by a “worm-like” bodyplan (Budd 2004). They inhabit aquatic niches only or depend on moist environments. These animals lack a true coelom, do not possess walking appendages nor locomotory cilia, posses a frontal mouth and are usually refered to as Cycloneuralia on the basis of their typical circular brain that forms a collar around their pharynx (Nielsen 2001). At least some members of all the Cycloneuralia posses an eversible anterior end (called introvert), which usually bears spines or teeth and gives the Cycloneuralia their alternative name of Introverta.

Nematodes and nematomorphs share many morphological characters, such as a collagenous cuticle and lack of circular muscles, and have been grouped in the Nematoida clade (Schmidt-Rhaesa 1996). Nematodes, or roundworms, are “thread like” creatures according to the origin of their name. Like the arthropods, they have adapted to almost all niches in the planet, but retained a similar body plan, to the extent that their classification is extremely complex. They vary extremely, however, in length ranging form 1 mm in *C. elegans* to eight meters in *Placentonema gigantissimum,* a nematode which parasites the placenta of whales. While the majority of nematodes are free living, approximately a quarter of them are parasites of almost all the other living creatures that are big enough to contain them. Nematomorphs are known as horsehair worms and their Latin name suggest a close morphological similarity with the nematodes. They are however obligate parasites: while the adult is free living in fresh waters, the larva parasites mainly arthropods and uses its introvert to penetrate the host body. During mating, nematomorphs characteristically aggregate in to a “Gordian knot” which gives their alternative name of Gordian worms.

Kinorhynchs, priapulids and loriciferans are linked by the presence of scalids (spiny appendages) on the introvert which can be everted for locomotion or to gather food (hence the name Scalidophora, Schmidt-Rhaesa 1998). Kinorhynchs, also called mud dragons or literally “snout-mover”, are extremely small (< 1 mm) meiobenthic animals which live in the interstices of costal sands where they prey on small diatoms and debris. They use their scalids and various spines on the trunk for locomotion: they withdraw the intorvert and push forward, then evert the introvert, hold with the spines and draw up the body (Brusca and Brusca 2001). Loriciferans, literally “armor-bearing”, are the most recently recognised phylum of ecdysozoans (Kristensen 1983). They are extremely small sediment-dwelling animals as the kinorhynchs, but they inhabit subtidal marine and freshwater intertidal space. They are characterised by the “lorica” which is a series of protective cuticular plates and by long, curved scalids on the introvert (evident in figure 1.1H). Like the kinorhynchs, they possess a conical mouth surrounded by stylets, suggesting that they may pierce and suck fluids from other animals. Finally, the priapulids take their name from the fertility divinity Priapus and are generally known as penis worms. They are carnivorous marine benthic animals and burrow tunnels using their large introvert. They are much larger than kinorhynchs (up to 10 cm) and, probably for this reason, priapulids have left a variety fossils, in particular the middle Cambrian *Ottoia*. The priapulid *Priapulus caudatus* is rapidly becoming a model organism as a natural outgroup to the Arthropoda. This is principally because in comparison to the nematodes, *P. caudatus* is characterised by slower evolving molecules and less derived morphology (Webster et al. 2006).

1.1.2 The status quo ante: Coelomata and Articulata

The notion of the Ecdysozoa as a clade is recent: it was formally proposed at the end of the last century on the basis of ribosomal molecular studies (Aguinaldo et al.1997). Prior to that time, our understanding of arthropods (and animals) evolution was extremely different.

Arthropods have been grouped with chordates, echinoderms, annelids and molluscs in the Coelomata, a clade characterised by a body cavity of mesodermal origin (figure 1.2A, Brusca and Brusca 2001). Coelomates have been generally distinguished from acoelmates (eg. Platyhelminthes or flat worms) which do not posses the body cavity and pseudocoelomates (eg. nematodes), which posses a “false” body cavity originated from the blastocoel. Intriguingly, the coelomate/acoelomate scenario reflects the gradualistic idea that animals evolve through a subsequent improvement of their body forms, moving from "basal" flattened acoelomate state to more complex coelomated one. However, it is now clear that evolution does not always proceed gradually and that some lineages may have undergone a drastic simplification of their bodyplan as a result of a adaptive selection (as in the case of intertidial animals for examples).

Within the Coelomata, the arthopods were long thought to be closely related to annelids in a clade of segmented invertebrates, the Articulata (Anderson 1973, see figure 1.2A), implying a common segmented ancestor in the invertebrates. Within the arthropods, myriapods were thought to be closely related to the hexapods (Atelocerata or Tracheata hypothesis, Heymons 1901) in some schemes with the addition of Onychophora (Uniramia, Manton 1977). From a morphological point of view, myriapods and hexapods share a distinctive head composed of five segments distinguished by their unique appendages – the antennal, intercalary (appendage-less), mandibular, and usually two pairs of maxillae (the second being the insect labium). Crustaceans, on the other hand, differ in having a second antennal rather than an intercalary segment. Further characteristics of the Atelocerata are tracheal breathing (where crustaceans have gills) and the possession of malpighian tubules for excretion. The Atelocerata/Uniramia hypothesis implied a paraphyletic origin of the arthropods and independent evolution of the "arthropod grade of organisation" in the atelocerates, crustaceans, chelicerates and extinct trilobites from primitive annelid-like ancestors (Nielsen 2001 and figure 1.2A).

1.1.3 The advent of molecular systematics and the new animal phylogeny.

The Ecdysozoa as a monophyletic group was formally proposed by a phylogenetic study based on the small nuclear ribosomal subunit (18S or SSU, Aguinaldo et al. 1997). In this study, the authors addressed a classical problem of phylogenetic reconstruction, Long Branch Attraction (LBA, Felsenstein 1978), which is responsible for the grouping of unrelated lineages that share either accelerated or reduced evolutionary rates. The authors showed that the basal position of nematodes within the Bilateria – as supported by the grouping of fast evolving *Caenorhabditis elegans* with distant outgroup sequences and in accordance with the Coelomata - was likely a LBA effect, as the use of slower evolving nematodes resulted in a group of arthropods, priapulids and nematodes. The authors named this group the Ecdysozoa on the basis of the periodic moulting (ecdysis) of a similar trilayered cuticle, which is influenced by ecdysteroid hormones. Other synapomorphies uniting the Ecdysozoa have been noted, such as a terminal mouth (seen in fossil arthropods), a lack of locomotory cilia (widely present in other protostomes) and the formation of the epicuticle from the tips of epidermal microvilli (Schmidt-Rhaesa et al. 1998). Earlier evidence in favour of the Ecdysozoa was, however, proposed in the 1992 by Eernisse and colleagues (Eernisse et al. 1992) on the basis of a cladistic analyses of morphological characters. Intriguingly, this contribution has been overlooked by the scientific community, partially because morphological comparisons are complicated by the extremely derived nature of some of the ecdysozoans and, most likely, because this work challenged the very well established Coelomata hypothesis (eg: vertebrates + arthropods, compare trees in figure 1.2).

The Ecdysozoa, which groups among others coelomate arthropods and pseudocoelomate nematodes, implies either that the nematodes have lost their coelomic cavity as a consequence of (at least primitive) miniaturisations and parasitic lifestyle, or that the coelom may have arisen independently in the arthropods and in other coelomate groups, such as chordates. In any case the scenario is less parsimonious than assuming a monophyletic origin of the coelomate lineages

The Ecdysozoa also challenges the Articulata, which groups segmented arthropods and annelids. The new scenario, as suggested by molecules, suggests instead that arthropods are ecdysozoans and that annelids are lophotrochozoans (Eernisse et al. 1992, Halanych 1995 and see figure 1.2B), implying either that segmentation in invertebrates arose at least two times independently or that the common ancestor of the protostomes was segmented and that segmentation has been repeatedly lost.

The advent of molecular systematics also challenged our interpretation of arthropod relationships. Virtually all molecular (Friedrich and Tautz 1995, Boore et al. 1998, Dunn et al. 2008 among the others) and some morphological (Kadner et al. 2004) analyses provided in the last decade convincing evidence that hexapods group with (and probably within) the crustaceans and not with the myriapods as traditionally believed (Atelocerata hypothesis, compare top of trees in figure 1.2). This new clade has been named Pancrustacea or, more correctly, Tetraconata on the basis of their shared ommatidial structure (Dohle 1997 and 2001, Firedirch and Tautz 1995, Telford 1995). The new scenario as suggested by molecules, implies a convergent acquisition of some characters in the hexapods and the myriapods as a consequence of the adaptation to life on lands. An impendent origin of arthropodisation as suggested by Manton (1977) has been disproved by virtually all molecular markers (Telford et al. 2008).

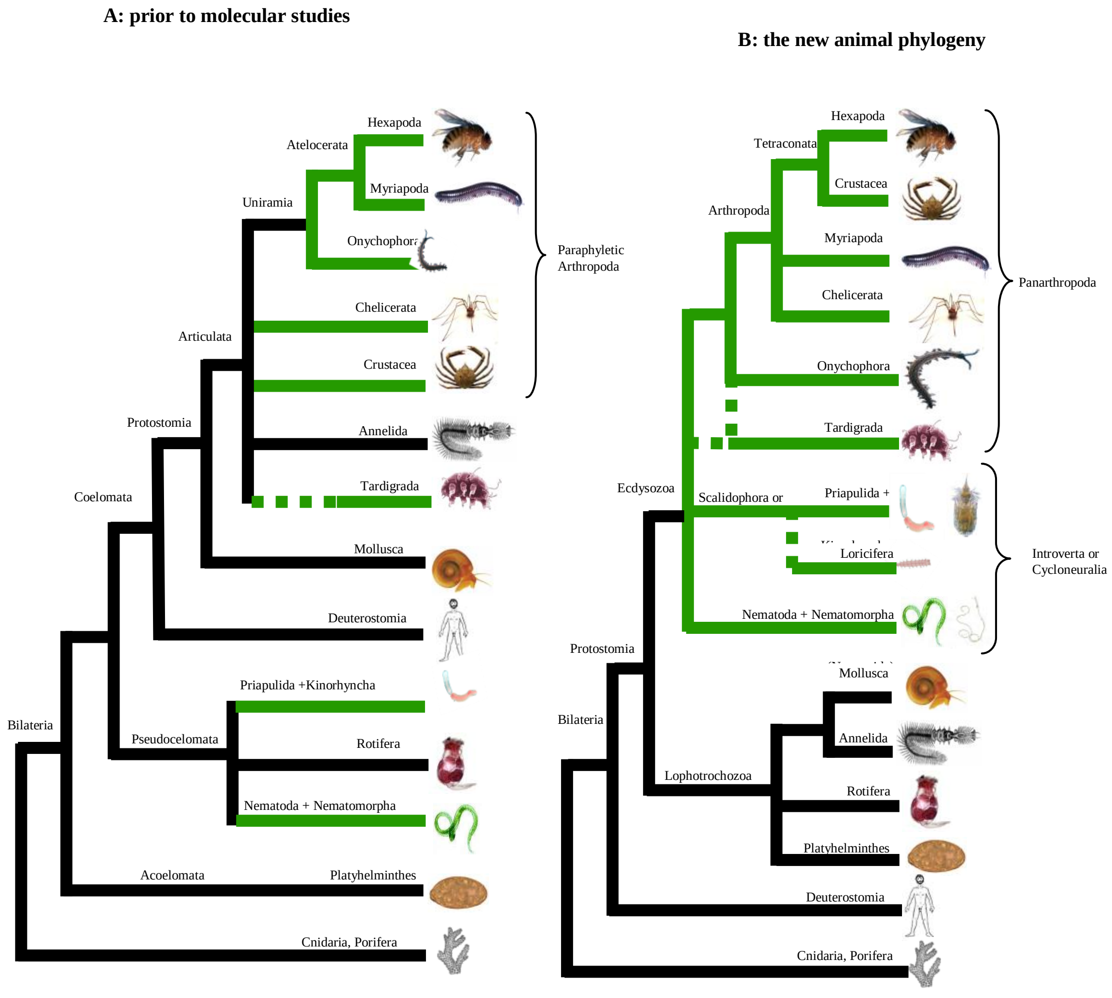


Figure 1.2. The old and new metazoan phylogeny. A: the view of animal relationships prior to molecular phylogenetics and in accordance with the Coelomata and Articulata hypotheses. Animals evolved gradually from a non-bilaterian to a coelomated form, through the intermediate acelomated (eg: Platyhelminthes) and pseudocelomated (eg: Nematoda) state. Arthropods are paraphyletic and closer related to segmented Annelida. Groups which now form the Ecdysozoa are in green. **B:** A consensus tree of metazoan relationships as supported by molecular studies. Nematoda and other Intorverta phyla are closely related to monophyletic Arthropoda in the Ecdysozoa clade. The Annelida joins the Mollusca and other phyla in the Lophotrochozoa clade.

The ecdysozoan hypothesis has always been difficult to accept from a morphological point of view (but see Eernisse et al. 1992), as its existence implies a complex evolutionary scenario with either a secondary loss of some characters (coelom, segmentation) or their independent gain in unrelated lineages. The Ecdysozoa is, however, of primary importance in biology, because the way the three primary animal models – vertebrates, nematode worms and fruitflies – are related has serious implications for genetic, genomic and evolutionary studies. Interpretations of comparative analyses rely on how the three groups are related. Many contributions have been published in support or against the Ecdysozoa hypothesis and, as discussed in the next section, only recently the Ecdysozoa have eventually prevailed.

1.1.4 Ten years of scientific debate

The last decade has been sparkled by a vigorous scientific debate over the existence of the Ecdysozoa. In the years which followed the seminal study of Aguinaldo and colleagues, the Ecdysozoa hypothesis has been validated by various molecular studies based on ribosomal subunits, nuclear genes and antigenic evidences (de Rosa et al. 1999, Haase et al. 2001, Mallat and Winchell 2002, Ruiz-Trillo et al. 2002). While the first molecular evidence for a clade of Ecdysozoa comprised only arthropods, priapulids and nematodes (Aguinaldo et al. 1997), following analyses successively added the remaining phyla (Telford et al. 2008).

Unexpectedly, the advent of phylogenomics – the phylogenetic approach based on whole genome sequences or large EST assemblies – supported a group of arthropods plus chordates with the exclusion of nematodes - as predicted by the Coelomata hypothesis (Blair et al. 2002, Wolf et al. 2004, Philip et al. 2005). It became rapidly clear, however, that these phylogenomic analyses were dependent on artefacts related to the LBA, as in the earlier Aguinaldo 1997 study. First, the extremely derived nature of nematodes had been shown to be responsible for secondary loss of many markers (protein families) resulting in an unspecific signal uniting nematodes and the distant outgroups (in which the markers were primarily absent) (Copley et al. 2004). Second, detailed exploration of signal in large datasets has shown that the grouping of arthropods and chordates was most likely a consequence of LBA due to suboptimal taxon sampling (Philippe et al. 2005). Similar explanations also clarified why rare amino acid replacements observed along the genomes apparently supported Coelomata and not Ecdysozoa (Rogozin et al. 2007, Irimia et al. 2007).

Further evidence against Coelomata has recently come from EST based phylogenomic analyses, which gave clear support in favour of the Ecdysozoa. (Dunn et al. 2008, Lartillot and Philippe 2008, Marletaz et al. 2008). These studies used a large sample of nematodes which may have effectively reduced the length of the nematodes stem branch and lessened the effect of possible LBA artefacts. Furthermore, a large phylogenomic analysis based on 42 metazoan complete genomes supported Ecdysozoa (Holton and Pisani, submitted). A final unquestionable proof comes from the comparative analysis of two adjacent fragments in the mitochondrial coded subunit Nad5 of the respiratory complex 1 (Papillon et al. 2004, Telford et al. 2008). In Nad5 there is a clear signature involving several amino acids which are mutated (and conserved) throughout all the protostomes while different mutations characterise the deuterostome and the non bilaterian outgroup sequences. Clear implications are that (i) protostomes (including nematodes and arthropods) are monophyletic, (ii) Coelomata is not a clade and (iii) phylogenomic studies supporting Coelomata are therefore wrong.

The Ecdysozoa/Coelomata dispute is a clear example of problems, such as LBA, correlated with the molecular inference of phylogeny. Some of these problems and possible solutions will be addressed in section 1.3 of this chapter.

1.2 Open questions in ecdysozoan systematics

While a monophyletic origin of the ecdysozoans is now widely accepted, relationships amongst the eight extant ecdysozoan phyla, as well as the affinities of major arthropod groups are still disputed.

1.2.1 Monophyly of Cycloneuralia?

There is a high uncertainty over the affinities of the Cycloneuralia, the group comprising Nematoida (nematodes and nematomorphs) and Scalidophora (priapulids, loriciferans and kinorhynchs, see figure 1.2B). Many characters unite them, such as the oral circular brain (hence the name Cycloenuralia), the absence of locomotory cilia and the presence of the eversible anterior introvert (Introverta). However, neither the introvert nor the collar-brain seem to be unique synapomorphies of the Cycloneuralia, reducing the possibility of a single acquisition of the two characters. The lophotrochozoan Gastrotricha possess an oral-circular brain, to the extent that they have been grouped with the "ecdysozoan cycloneuralian" (Ruppert, Fox and Barnes 2004, Nielsen 2001). The tardigrades are also characterised by a group of ganglia which completely surround the mouth opening, although they also possess typical lateral brain lobes which resemble the arthropods. Still, the introvert is found only in very few nematodes, but also in the Sipuncula, which are Lophotrochozoa.

According to molecules, the scenario of monophyletic Cycloneuralia is even more unclear: combined ribosomal subunits analyses support a basal position of the Scalidophora (Mallat and Giribet 2006 and figure 1.3 A), while larger phylogenomic support monophyly of Cycloneuralia (Dunn et al. 2008, figure 1.3 C). As suggested by Telford and colleagues (2008) determining support either for a paraphyletic or a monophyletic origin of the Cycloneuralia is extremely important for drawing a picture of the ancestral ecdysozoan (figure 1.4 A). If Cycloneuralia are paraphyletic then their common ancestor is also the ecdysozoan ancestor and probably possessed a collar-brain and an introvert, characters which have been lost in the arthropods.

1.2.2 Tardigrada

Morphology strongly supports a common origin of the three panarthropod phyla – arthropods, tardigrades and onychophorans, but this has found little molecular support (Nielsen 2001). An arthropod affinity of the velvet worms (onychophorans) is now widely accepted (Dunn et al. 2009, Mayer and Withington 2009). The complete mitochondrial genomes of two onychophorans have been sequenced, but analyses of these are questionable from a morphological point of view, as they do not support

Panarthropoda, but place onychophorans sister to a group composed of arthropods plus Priapulida (Podsiadlowski et al. 2008).

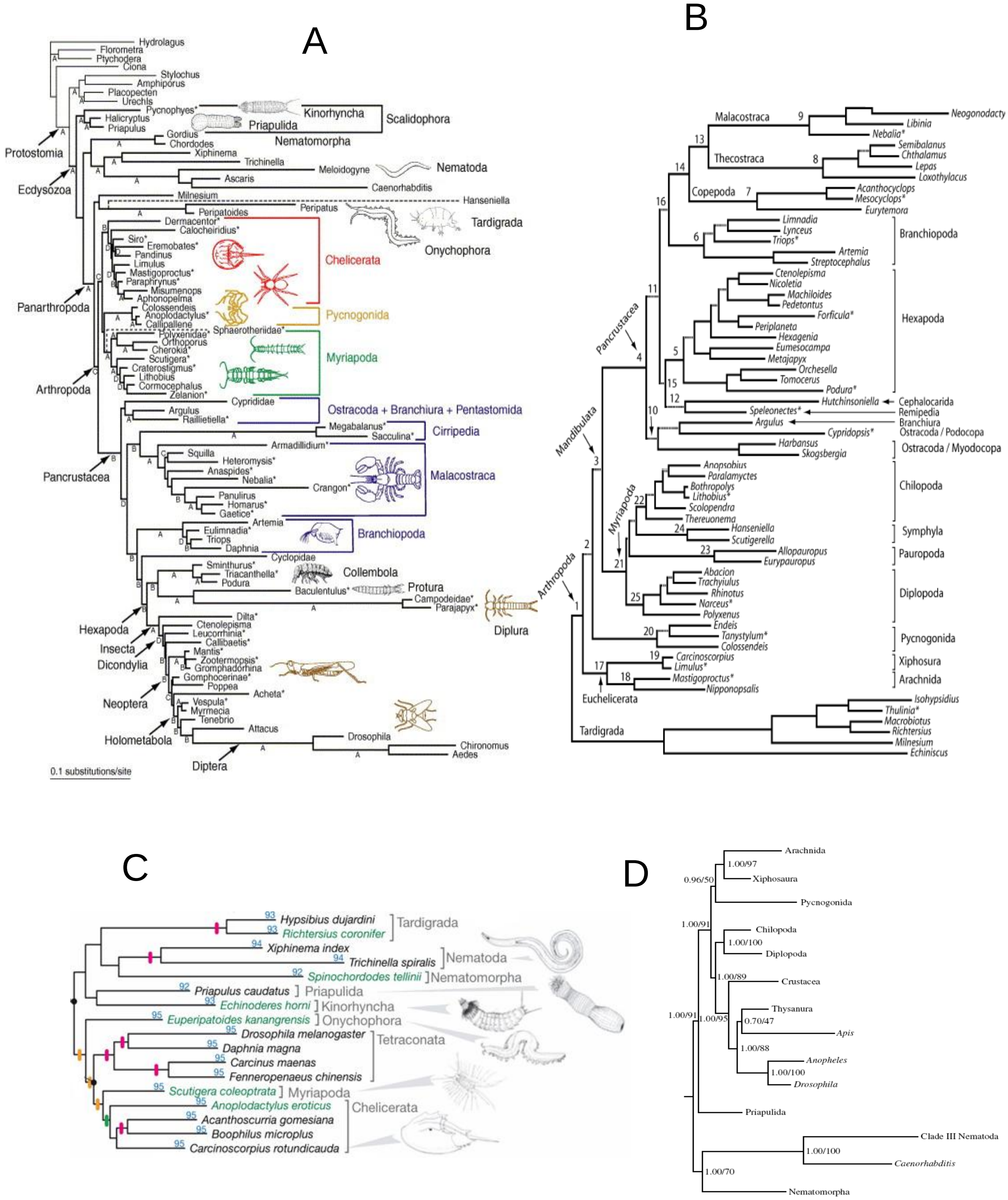


Figure 1.3 Great uncertainty over affinity of the myriapods. While ribosomal (A, from Mallat and Giribet 2006) and phylogenomic (C, from Dunn et al. 2008) studies support a clade of myriapods plus chelicerates (Myriochelata), combined marker analysis (D, from Bourlat et al. 2008) support a group of myriapods plus Pancrustacea (Mandibulata) more in accordance with morphological and developmental observations. Phylogeny based on nuclear markers (B, from Regier et al. 2008, which is the updated analysis of Regier et al. 2005) failed to support either hypotheses or gave modest support for Mandibulata.

On the other hand, the position of tardigrades is equally unclear (figure 1.4 B). Ribosomal sequences (Mallat and Giribet 2006, figure 1.3 A) support a group of tardigrades plus onychophorans as sister to the arthropods, while EST data have challenged the morphological view, linking tardigrades and nematodes (Lartillot and Philippe 2008). In these analyses, tardigrades and nematodes are characterized by long branches, suggesting that the tardigrade plus nematode clade could represent a phylogenetic artifact. This inference is reinforced by the recent phylogenomic analyses of Dunn and colleagues (2008), which suggested that tardigrade affinity may be model-dependent: analyses using the WAG matrix (Figure 1.3 C) support a nematode affinity of the tardigrades, while analyses performed using the CAT model (Lartillot and Philippe 2004) support tardigrades as basal to a group of onychophorans plus arthropods.

1.2.3 Basal arthropod relationships: Mandibulata versus Myriochelata.

While monophyly of (Eu)arthropoda is well established, one intriguing aspect currently under strong debate, and a central theme of this thesis, is the position of the myriapods.

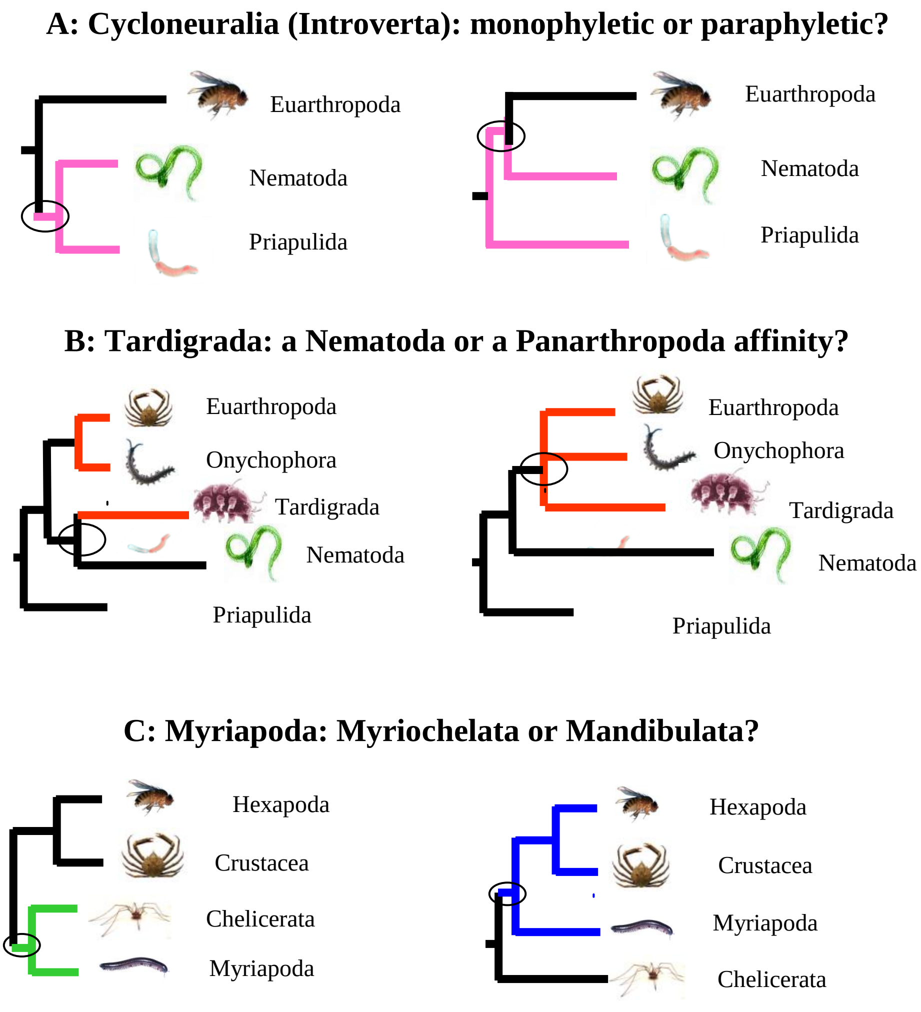
Chelicerates, compared to Tetraconata (hexapods and crustaceans) and myriapods, have a radically different arrangement of head appendages. They possess chelicerae and pedipalps in place of first and second antennae and walking legs in place of mandibles, maxillae/labia. When compared to chelicerates, the detailed similarities of the arrangement of head segments and associated appendages in Tetraconata and myriapods strongly supports their sister group relationship in a clade which has been named the Mandibulata in recognition of the similarity of their biting mouthparts, the mandibles. (Edgecombe et al. 2003, see figure 1.4 C) In crustaceans, insects and myriapods mandibles are all located on the first post-tritocerebral segment, and are followed by a further two pairs of feeding appendages: the maxillae. Expression patterns of the genes *Distal-less* and *dachshund* in mandibles of the three groups have been interpreted as showing that all three are gnathobasic structures formed from the coxal (proximal) leg segment and in all three groups the gnathal part of the mandible is subdivided into strikingly similar parts. Notably, the homologs of the mandibular and maxillary segments in chelicerates bear walking legs. These appendages represent a primitive character state rather than a derived one. In addition to the complex similarities of head structure, likely synapomorphies of Mandibulata include arrangements of midline neuropils in the brain, correspondences in cell numbers and specialised cell types in the ommatidia, similar sternal buds in the stomodeal region, and specific arrangements of serotonin-reactive neurons in the nerve cord (a detailed list of characters in appendix 1).

Considering the complex shared features of myriapod and tetraconatan head morphology, it is surprising that the majority of molecular markers do not support the Mandibulata, instead placing the myriapods as the sister group of the chelicerates in an assemblage that has been named the Myriochelata or Paradoxopoda (figure 1.4 C). Support for the Myriochelata clade was first obtained using mitochondrial protein coding sequences (Hwang et al. 2001, Pisani et al. 2004) and supported by analysis of small subunit rRNAs (Mallatt et al. 2004), although lessened by updated analyses (Mallatt and Giribet 2006, figure 1.3A). On the other hand, although this is clearly not independent of purely morphological analyses, work based on mixed morphological and molecular character sets (Giribet et al. 2001) supports the Mandibulata concept. Mandibulata is also supported by a recent analysis of mixed molecular markers (Bourlat et al. 2008, figure 1.3 D, but see Paps et al. 2009), while analyses of nuclear coding genes (Regier et al. 2005, figure 1.3 B) support neither hypothesis, but rather link the chelicerates to Tetraconata. In an effort to minimise stochastic error, the work of Regier and colleagues (2005) has been recently expanded to a large dataset of 62 gene from 13 species: their results gave some support for Mandibulata, although this is conditioned by the use of certain analytical conditions (Regier et al. 2008). Finally, the largest scale study of metazoan relationships (Dunn et al. 2008, figure 1.3 C) involving 21152 amino acids from 150 genes, supports Myriochelata with greater than 90% bootstrap support, although the taxonomic sampling included only 11 panarthropods.

Interestingly, internal branches leading either to Myriochelata or Mandibulata are short in all of the phylogenetic reconstructions mentioned and in some cases are poorly supported implying a weak phylogenetic signal (see the nodes in figure 1.3). It has also been shown that support for either of the two hypotheses may depend on the nature of the outgroup used (Rota-Stabelli and Telford 2008), exclusion of sites (Pisani 2004) or method of phylogenetic inference (Regier et al. 2008), suggesting that signal at this node is weak and that phylogenetic conclusions may be prone to systematic errors.

The only morphological character which has been cited in support of Myriochelata involves the mechanism by which neurons arise from clusters of cells which migrate from the neuroectoderm (Stollewerk and Chipman 2006 for a review). This character has been found in myriapods and chelicerates but not in Tetraconata in which single cells are segregated from the neuroectoderm. However, the absence of a similar study in a close outgroup, has always prevented strong conclusions being drawn, as this character may either be a synapomorphy (uniting myriapods and chelicerates) or a symplesiomorphy (shared by myriapods, chelicerates and the outgroup, but absent in the Tetraconata). Recently, Georg Mayer (Mayer and Whitington 2009) has been able to polarise this character as a true synapomorphy of the Myriochelata; the onychophoran outgroup possesses a process of neurogenesis more closely resembling that of Tetraconata than that of the myriapods and the chelicerates. The study of Mayer also found an additional synapomorphy of the Myriochelata, based on the presence of a ‘cumulus’ of mesenchymal cells which determine the dorsal region in chelicerates and myriapods. The cumulus is clearly absent in the onychophorans and has been never observed in Tetraconata. However, the Myriochelata hypothesis either implies that the many similarities seen between the myriapod and insect/crustacean heads evolved convergently, or that the head structures in the mandibulate groups are indeed homologous, and the walking legs seen in homologous segments in chelicerates are a reversion to the ancestral state.

Monophyly of myriapods and of chelicerates has also been challenged by molecular and morphological studies, in some cases placing Pycnogonida as basal to all other arthropods, a clade known as Cormogonida, in some other supporting paraphyletic chelicerates (Negrisolo et al 2004, Giribet Edcombe and Wheeler 2001; Maxmen et al 2005; Mallat et Giribet 2006). Mitochondrial studies have reported an affinity between Acaria and Pycnogonida, which is believed to be the effect of a systematic error due to Pycnogonida being fast evolving (Podsiadlowski and Braband 2006; Park et al 2007). This finding is reinforced by the recent multi gene analysis of Regier and colleagues (2008) which show that position of Pycnogonida is parameter dependent.



A: Cycloneuralia (Introverta): monophyletic or paraphyletic?

B: Tardigrada: a Nematoda or a Panarthropoda affinity?

C: Myriapoda: Myriochelata or Mandibulata?

Figure 1.4: **Three systematic problems addressed in this thesis.** (A) Are the Cycloneuralia a monophyletic group? (B) Are the tardigrades more related to nematodes or to the arthropods (Panarthropoda)? (C) Are myriapods closer related to chelicerates (Myriochelata hypothesis) or to hexapods and crustaceans (Mandibulata hypothesis)?

1.2.4 Are Crustacea paraphyletic?

Relationships among the crustacean classes, as well as their monophyly, have also been questioned. Crustaceans encompass at least six classes: Branchiopoda (brine shrimp, water flea), Malacostraca (crabs, shrimps), Ostracoda (seed shrimps), Remipedia, Cephalocarida (horseshoe shrimps) and Maxillopoda (barnacles, copepods). The Maxillopoda possibly being paraphyletic and implying additional classes: Thecostraca, Copepoda, Branchiura, Penatastomida, Mystacocarida and Tantulocarida (figure 1.5). The majority of molecular analyses have suggested a paraphyletic origin of the crustaceans with the hexapods being in effect a group of terrestrial crustaceans, but there is little consensus as to how the crustacean classes are related to each other and where the hexapods fit into the Tetraconata assemblage (Regier et al 2008, Mallat and Giribet 2006, Carapelli et al. 2007, Dunn et al. 2008,).

Ribosomal and phylogenomic markers tend to support Branchiopoda as sister to the hexapods (Mallat and Giribet 2006, Lartillot and Philippe 2007, Dunn et al. 2008, Roeding et al. 2009), although certain morphological characters group malacostracans (and remipedes where sampled) with hexapods (Harzsch 2002, Friedirch et al. 2004). Mitochondrial studies have failed to resolve this problem unambiguously, although under certain conditions of analysis a group of Malacostraca plus Branchiopoda and Cephalocarida (Thoracopoda hypothesis) is supported (Carapelli et al 2007). A reasonable alternative is the Entomostraca hypothesis which groups all crustacean classes with the exception of the Malacostraca (Hessler 1992), and has found only poor support from molecules (Giribet et al. 2005).

1.2.5 Relationships within insects.

It has been broadly accepted that insects (Ectognatha), together with collembolans, diplurans and proturans (the three latter being Enthognatha - with internal mouthparts) form the Hexapoda, a subphylum characterised by a six-legged bodyplan. Monophyly of hexapods has, however, been questioned on the basis of mitochondrial studies (Nardi et al. 2003, Carapelli et al 2007) although the majority of other markers support a common origin of the hexapods (Regier et al. 09, Dunn et al 2008, Mallat and Giribet 2006). See figure 1.5 for a consensus of current systematics of Tetraconata.

The vast majority of insects are neopterans and can fold their wings over the abdomen, while palaeopterans (dragonflies, mayflies and extinct Dyctioneuida) are characterised by unfoldable wings. Among the neopterans, the Holometabola (flies and bees amongst others) is by far the most successful and radiated group of insects. This is partially explained by their capability of niche diversification and by their ontogenic strategy (Hunt et al. 2007, Yang 2001). Holometabolans undergo complete metamorphosis and develop wings internally during their pupal stage (hence the alternative name of Endopterygota), while remaining winged insects (informally Hemimetabola) posses an incomplete metamorphosis which passes through gradual changes and develop wings externally (hence Exopterygota). The relationships of holometabolan orders are disputed and some markers even failed to recover their monophyly (Whiting et al 1997, Mallat and Giribet 2006, Carapelli et al 2007, Timmermans et al. 2008, Cameron et al. 2004). Traditionally, morphologists have placed the Coleoptera (beetles) at the base of the Holometabola, with the Hymenoptera (bees, wasps and ants) closer to Diptera (flies) and Lepidoptera (butterflies and moths) (Kristensen 1981). Recent phylogenies, however, have suggested that the Hymenoptera may be the basal holometabolan clade, either as sister to the remaining holometabolans (Savard et al 2007, Wiegmann et al.2009) or as sister to the Coleoptera (Timmermans et al. 2008).

Developmental strategies of the Hemimetabola vary extensively among different orders and even within orders, varying from "pseudometaboly", which is characterised by a reduced ontogenetic process, to "neometaboly" in some bugs and thrips, which is characterised by a holometabolan-like development (Heming 2003). The ontogenetic variety of the hemimetabolans is reflected by an extreme uncertainty over their phylogenetic affinities. Two assemblages are widely recognised: the Hemipteroidea (or Paraneoptera), a clade encompassing bugs, booklice, lice and thrips, and the Orthopteroidea (or Polyneoptera), which groups remaining hemimetabolans, except for stoneflies (Plecoptera) (Grimaldi and Engel 2005, Brusca and Brusca 2003 amongst others). Among the orthopteroids, cockroaches, termites and mantids are grouped in the monophyletic Dictyoptera (Kristensen 1975, Nichols 1989, Ma et al 2009, Cameron et al. 2006, Lo et al. 2000). It is a common view that the hemipteroids are sister to the holometabolans in a clade named Eumetabola, (Wheeler et al 2001, Kristensen 1991 and 1995, Grimaldi and Engel 2005, Hamilton 1972, various chapters in Fortey and Thomas 1998), an hypothesis which found some evidence in the complex hemipteran and thysanopteran nymphal ontogeny, but poor morphological support. It has been suggested that hemipteroids and holometabolan larvae lack frontal ocelli (Paulus 1979) and that the adults share an "R plus M forewing media fusion, the presence of a "jugal bar", a "holometabolan" type mesotrochantin and cryptosterny (Wheeler et al. 2001).

.

Molecular markers are discordant over the hemipteroids position, however, either supporting Eumetabola (Hovmöller et al 2002, Kjer 2004), paraphyletic Hemimetabola (Mallat and Giribet 2006, Timmermans et al. 2008, Roeding et al. 2009) or, although the result is model dependent, a sister relationship between hemipteroids and orthopteroids (Lartillot and Philippe 2008). Various mitogenomic studies have also addressed relationships of non holometabolan insects, but have reached ambiguous conclusions and have been unable to recover monophyly of some commonly accepted orders such as the hemipterans (Cameron et al 2005, Hassanin et al. 2005, Carapelli et al 2007).

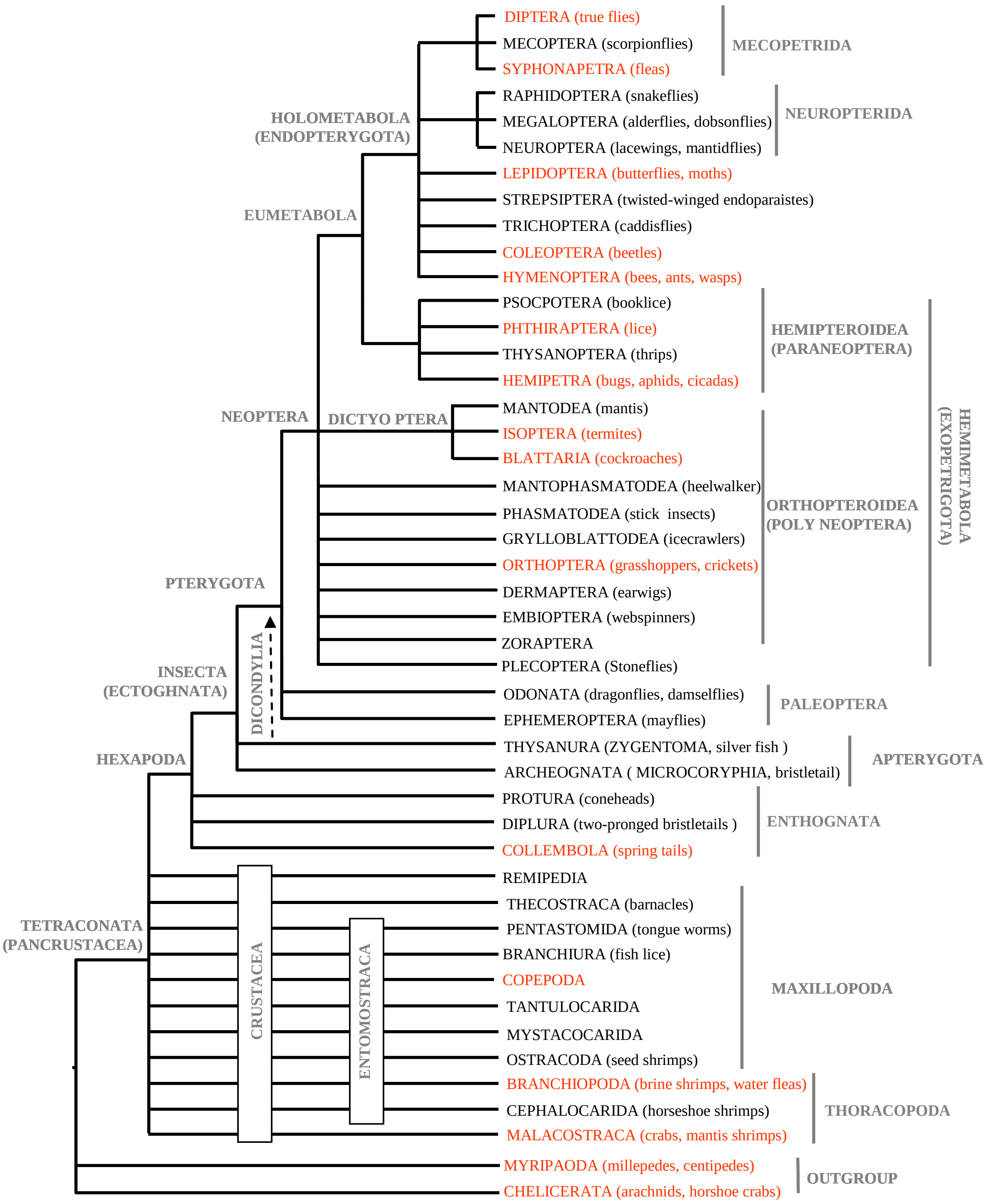

DIPTERA (true flies)
MECOPTERA (scorpionflies)
SYPHONAPETRA (fleas)
MECOPETRIDA
RAPHIDOPTERA (snakeflies)
MEGALOPTERA (alderflies, dobsonflies)
NEUROPTERA (lacewings, mantidflies)
NEUROPTERIDA
HOLOMETABOLA
(ENDOPTERYGOTA)
LEPIDOPTERA (butterflies, moths)
STREPSIPTERA (twisted-winged endoparaistes)
EUMETABOLA
TRICHOPTERA (caddisflies)
COLEOPTERA (beetles)
HYMENOPTERA (bees, ants, wasps)
PSOCPOTERA (booklice)
PHTHIRAPTERA (lice)
THYSANOPTERA (thrips)
HEMIPETRA (bugs, aphids, cicadas)
HEMIPTEROIDEA
(PARANEOPTERA)
NEOPTERA
DICTYO PTERA
MANTODEA (mantis)
ISOPTERA (termites)
BLATTARIA (cockroaches)
MANTOPHASMATODEA (heelwalker)
PHASMATODEA (stick insects)
GRYLLOBLATTODEA (icecrawlers)
ORTHOPTERA (grasshoppers, crickets)
ORTHOPTEROIDEA
(POLY NEOPTERA)
HEMIMETABOLA
(EXOPETRIGOTA)
PTERYGOTA
DERMAPTERA (earwigs)
EMBIOPTERA (webspinners)
ZORAPTERA
PLECOPTERA (Stoneflies)
INSECTA
(ECTOGHNATA)
DICONDYLIA
ODONATA (dragonflies, damselflies)
EPHEMEROPTERA (mayflies)
PALEOPTERA
HEXAPODA
THYSANURA (ZYGENTOMA, silver fish)
ARCHEOGNATA (MICROCORYPHIA, bristletail)
APTERYGOTA
PROTURA (coneheads)
DIPLURA (two-pronged bristletails)
COLLEMBOLA (spring tails)
ENTHOGNATA
REMIPEDIA
TETRACONATA
(PANCRUSTACEA)
THECOSTRACA (barnacles)
PENTASTOMIDA (tongue worms)
BRANCHIURA (fish lice)
COPEPODA
TANTULOCARIDA
MYSTACOCARIDA
OSTRACODA (seed shrimps)
MAXILLOPODA
CRUSTACEA
ENTOMOSTRACA
BRANCHIOPODA (brine shrimps, water fleas)
CEPHALOCARIDA (horseshoe shrimps)
MALACOSTRACA (crabs, mantis shrimps)
THORACOPODA
MYRIPAODA (millepedes, centipedes)
CHELICERATA (arachnids, horshoe crabs)
OUTGROUP

Figure 1.5. Tetraconata relationships. The cladogram is a schematic representation of current knowledge of Tetraconata relationships. In brackets are the English common names of some representative of the 30 orders of insects and the putative 11 classes of crustaceans. Major commonly accepted clades are in grey. Lineages sampled in the analysis of chapter 6 are in red.

1.3 Current molecular phylogenetics

After many years of methodological improvements in the field of molecular systematics, it is now possible to use sophisticated models of evolution which account for example for heterogeneity of the substitution process among sites (Lartillot and Philippe 2004). Methods of phylogenetic inference also significantly improved to the extent that Bayesian (Huelsenbeck and Ronquist 2001) and fast maximum likelihood methods (Stamatakis 2006) allow the analysis of large and dense molecular datasets. It has however became clear that molecular phylogeny may be complicated by reconstruction artefacts such as Long Branch Attraction (responsible for Coelomata as discussed in section 1.1.4). Some of these problems will be addressed in this section.

Throughout this thesis I will explore the phylogenetic relationships of the ecdysozoans using different molecular markers (in particular mitochondrial), various methods of phylogenetic inference and a variety of models of evolution. Majority of the analyses will be carried out at the amino acid level for which I have developed new models of evolution aimed to generate more reliable phylogenies (chapter 2). Accordingly, in this section I address some up to date problems and methods in phylogenetic reconstruction, focusing attention on models of amino acid evolution and inference of phylogeny using mitochondrial sequences.

1.3.1 Systematic and stochastic errors in molecular phylogeny

One of the possible explanations of the great level of uncertainty over ecdysozoan relationships (see in particular section 1.2.3) is the lack of suitable molecular datasets. Taxonomically broad datasets, such as the mitogenomic and ribosomal ones, suffer from being limited in their number of positions, allowing space for possible stochastic errors due to a lack of enough phylogenetic signal (for example Mallat and Giribet 2006 in figure 1.3A and Regier et al. 2005 in figure 1.3 B). On the other hand, larger datasets (for example of Regier et al 2008, Lartillot and Philippe 2008, Roeding et al. 2009, Dunn et al. 2008 in figure 1.3 C) suffer from being poorly taxonomically sampled at some nodes of interest, a condition which may lead to systematic errors, such as long branch attraction (LBA) artefacts (Felsenstein 1978). These problems may be exacerbated by rapid divergence of the main lineages, followed by subsequent long period of within lineage changes (autapomorphies), which may have diluted the historical signal (Whitfield and Kjer 2008, Rokas and Carroll 2006). Some of these problems can be alleviated by using a large phylogenomic dataset which is able to provide more phylogenetic signal (due to more genes) and reduce the number of autapomorphic and/or homoplastic changes observed (due to more taxa and shorter internal branches, Philippe and Telford 2006).

Probably the most widely recognised systematic error is LBA, which arises from unequal rates of evolution among lineages. LBA is particularly marked when analysing lineages which are the result of close speciation events or have differentiated in ancient times. In both cases a small number of informative substitutions (those that happened before the split of two lineages) may be diluted by a large number of homoplastic substitutions (those happened after the split of the two lineages), which can be responsible for “non-phylogenetic signal” (Baurain, Brinkmann and Philippe 2006). This seems likely to be the case in the myriapod lineage, as branches describing their affinity are extremely short in all the molecular phylogenies published so far (figure 1.3), suggesting a lack of informative signal and a likelihood of encountering systematic and/or stochastic errors.

1.3.2 Models of amino acid evolution: from homogeneity to heterogeneity of the replacement process.

Systematic errors in phylogeny come from model violation: the model of evolution may incorrectly interpret the multiple substitutions occurring at a given position. This problem is exacerbated when different lineages posses unequal rates of evolution (the LBA artefact) and when the signal is subtle due to fast radiation of lineages, as it may be the case of myriapods (discussed in 1.2.3). A proven way to overcome the non-phylogenetic signal is to use better evolutionary models (Whelan, Liò and Goldman 2001; Felsenstein 2004, Philippe et al. 2005).

The history of models of amino acid evolution is intimately linked with the history of molecular systematics. The first attempts to obtain phylogenetic information from molecules were indeed based on amino acid sequences, as they were the only sequences available in the early sixties, thanks to Edman degradation sequencing, developed a decade before (Edman 1950). In their seminal work, Dayhoff and Eck (1966) analysed proteins on the basis of a symmetrical matrix (20 X 20), in which all the possible replacement between amino acids had the same probability to occur. Their approach was parsimonious, so that they inferred the tree minimising the number of steps observed along the tree. The analysis of Dayhoff wasn’t the first computational approach to systematic studies, but the first to use molecules. In previous years Cavalli-Sforza and Edwards already analysed gene frequency polymorphisms in human populations and, incredibly, introduced in a single paper both the parsimony and the likelihood methods (Edwards and Cavalli-Sforza 1964).

It became rapidly clear, however, that the replacement probability was not the same for each pair of amino acids and in the following years Dayhoff and Eck (1968) proposed an empirical model, the PAM (probability of accepted mutation) based on the parsimonious counts of amino acid changes observed in various sets of related proteins. In the PAM1 model, each value of the 20 X 20 matrix is the probability of changing from one amino acid to another when 1% of the amino acids of the alignment are expected to change. Although the PAM matrix is no longer used in its original version, it was the first attempt to account for the amino acid heterogeneity of the replacement process.

In the following years, other replacement matrices have been proposed based on a more accurate calculation and larger datasets. Jones, Taylor and Thornton (1992) proposed the JTT matrix, which was based on transmembrane proteins, suggesting that the secondary structure of proteins plays a significant role in determining amino acid composition and the replacement probabilities between them. A significant improvement has been made by ameliorating the way in which the replacement matrices are calculated. While PAM or JTT have been estimated counting substitutions according to a parsimony criterion, Adachi and Hasegawa (1996) used a maximum likelihood approach to estimate MtREV from a mitochondrial protein dataset, as explained in more detail below. This approach, based on the reversibility of the evolutionary process, has the advantage of partially accounting for saturation and has been applied successfully to large nuclear datasets as in the case of the WAG model (Whelan and Goldman 2001).

1.3.3 Empirical and mechanistic models

Although nucleotides have been preferred in the last two decades for computational reasons, the majority of current "deep" phylogenetic analyses are carried out at the amino acid level (Rota-Stabelli et al 2009). The reason is that nucleotide sequences are more susceptible to substitutional saturation. Coding sequences can also be analyzed at the codon level using a variety of mechanistic (Yang and Nielsen, 2008) or, as recently proposed, empirical models (Kosiol et al., 2007). However these models are still too computationally demanding for phylogenomic studies and are not indicated for deep level mitogenomic studies because mtDNA genetic codes vary in different metazoan lineages.

Mechanistic means that the replacement rates are estimated directly from the dataset during the tree search (and not taken from a pre-existing empirical matrix, such as PAM, JTT or WAG) (Lanave et al., 1984; Yang et al., 1998). Although computationally demanding, amino acid substitutions can also be described by a mechanistic General Time Reversible model (GTR, next paragraph for more details). The mechanistic approach, usually simply refered to as GTR, is often applied in nucleotide studies as the corresponding replacement matrix contains only 8 values (half of a 4 X 4 matrix). When the mechanistic GTR approach is applied to amino acids, it risks introduction of stochastic errors in the estimation of replacement rates due to the relatively limited quantity of information present in most datasets. Reliable estimation of the amino acid replacement rates needs a significant amount of substitutional information from the dataset and small datasets typically used in phylogenetic analyses may not contain sufficient information. Additionally, a clear problem in this procedure is the large size of the amino acid alphabet, which makes the estimation of all the parameters a demanding computational task (the matrix in this case is (20 X 20)/2). Consequently, the majority of available models are empirically derived, such that replacement rates (r) and amino acid frequencies (π) are stored in matrices that have been pre-estimated from large, well-curated datasets.

While empirical models of nuclear protein evolution such as the aforementioned JTT and WAG or the new LG are estimated from taxonomically varied datasets (Jones et al. 1992, Whelan and Goldman, 2001, Le and Gascuel, 2008), models of mitochondrial amino acid evolution have been estimated from phylogenetically restricted datasets: MtREV, Mtmam, MtArt and MtPan, are based on the analysis of only vertebrates, mammals, arthropods and Tetraconata respectively (Adachi and Hasegawa, 1996, Yang et al., 1998, Abascal et al., 2007, Carapelli et al., 2007).

Empirical models can be estimated within a maximum likelihood framework, which calculates the evolutionary replacement matrix and stationary amino acid frequencies that best explains how the observed data (amino acid sequences) evolved accordingly to their phylogenetic tree. In the case of proteins, given an amino acid alignment of N species and the corresponding phylogenetic tree, it is possible to estimate a 20 X 20 amino acid replacement matrix (R) and the frequencies of the 20 amino acid at stationarity (π_j, for any j 20 amino acids) (Adachi and Hasegawa, 1996). The values in the R matrix are called replacement rates (r_{ij}) and are multiplied by the stationary frequencies (π_j) to obtain the corresponding exchangeability rates $q_{ij} = \pi_j \mathrm{r}_{ij}$ that are the values of the 20 X 20 exchangeability matrix Q. The total number of free parameters of such a model are 20 X 19 (replacement rates r_{ij}) + 19 (stationary frequencies π_j) – 1 (because only relative rates are considered) = 398. Empirical and mechanistic models are usually based on the assumption that the replacement process is reversible and thus assume that the substitution probability of one character to another is the same in both directions (the GTR assumption: $\pi_j \mathrm{r}_{ij} = \pi_i \mathrm{r}_{ji}$). This obviates the need for a rooted tree in the estimation of model parameters and makes the replacement matrix symmetrical - almost halving the number of free parameters in the model.

1.3.4 Among site heterogeneity of the replacement process

A possible problem, which is generally not taken into account in phylogenetic reconstructions, is the heterogeneity of the replacement process among sites. This characteristic is intrinsic to the structure of proteins, whose amino acids are fundamentally heterogeneous, due to the alternation, for example, of buried and exposed residues which posses different evolutionary dynamics. Rate heterogeneity is commonly accounted for with a distribution of among sites rate variation, for example the Gamma distribution of Yang (1996) or its CAT approximation implemented in RAxML (Stamatakis 2004), which does not have to be confounded with the CAT model described below. However, most models of protein evolution (JTT, WAG, LG, mtREV) assume homogeneity of the replacement process and treat all positions of the alignment the same (Jones et al. 1992, Whelan and Goldman. 2001, Le and Gascuel 2008). Use of these homogenous models may promote phylogenetic artefacts due to model violations, because the models assumes among site homogeneity where none exists. Problems from heterogeneity of the replacement process are exacerbated by using many unrelated genes as in the phylogenomic approach.

A significant improvement in accommodating site heterogeneity has been made by a complex model that assigns sites to 10 different structural classes using Hidden Markov models (Liò and Goldman 2002). This model, named MT126 and explicitly proposed for mitochondrial amino acid sequences, has been reported to perform better than MtREV over a large range of eukaryotes (Metazoa, Fungi and plants), but has been shown to be comparable to MtREV when analyzing a vertebrate dataset, suggesting that the great complexity of this model may not be justified by a modest increase in likelihood. Recently, the CAT model (Lartillot and Philippe 2004) and the empirical adaptation of it (Le et al. 2008) allowed the relaxation of the assumption of homogeneity among sites and has been shown to lessen problems of model violation, retrieving more reliable phylogenies and outperforming homogeneous models (Lartillot et al. 2007). The CAT model assumes the existence of distinct classes of amino acid profiles and sorts the sites into different classes on the basis of the equilibrium frequencies of the 20 amino acids (calculated at each site). More recently, principal component analysis has been used to define four classes of sites, which can be used in a class frequency (cF) model (Wang et al. 2008).

A further underestimated problem is the variation of the replacement rate over time, a characteristic known as heterotachy (Lopezet al. 2002). This problem is intrinsic in the heterogeneous nature of evolution. Some lineages evolve at a constant rate of evolution and similarly to their ancestor, a condition which has referred to as the molecular clock. However, it has become clear that the molecular clock is extremely local within a phylogenetic tree: some lineages may undergo an acceleration or a reduction of the replacement rate (or of the fixation rate). It is surprising that the vast majority of evolutionary models (and the programs in which models are implemented) assume the stationarity of the replacement rate and expect that all taxa in a dataset evolve clocklike. The problem with heterotachy is that it is difficult to address computationally to the extent that the number of free parameters in a phylogenetic reconstruction would become dramatically high. In other words it is impossible to assign a different replacement rate to each site of all taxa of the alignment. However, the covarion approach, although it is a simplification of the heterotachy process, has been shown to be a quick and effective estimator of this problem (Zhou et al. 2007).

A similar problem to heterotachy is the heterogeneity of the stationary frequencies over time (or among lineages). This problem is intimately correlated with Heterotachy, as in a GTR framework, the replacement probability q_{ij} is composed of both replacement rates r_{ij} and stationary frequencies π_j or π_i (depending on the direction of substitution). This problem has been particularly studied in mitochondrial sequences (see next section 1.3.4) which are extremely heterogeneous in their stationary frequencies and heterogeneous models of evolution, such as CAT-BP and the vector model implemented in P4 have been built (Blanquart and Lartillot 2008, Foster 2004). These models allow the stationary frequencies to vary in different parts of tree. The advantage of CAT-BP is that it accounts for both among sites and among lineages compositional heterogeneity.

1.3.5 Mitogenomics: ease and caveats

Despite an ongoing debate concerning their utility in phylogenetics, mitogenomic studies continue to abound in the scientific literature (Cameron et al. 2004). This can be explained both by conceptual advantages such as a conserved gene set, the unambiguous orthology of genes and the presence of rare genetic changes including gene rearrangements or differences in genetic code. Moreover, there are historical and methodological reasons that favor mitochondrial DNA such as the availability of primers for many lineages and the relative ease of generating new data. On the other hand, mitochondrial sequences are well known to suffer from a variety of problems that may be responsible for the dilution of the true phylogenetic signal and the generation of homoplasies.

One of the main problems of mitogenomics is lineage-specific compositional heterogeneity. This can be so extreme as to influence the amino acid content of the encoded proteins (Foster, Jermiin and Hickey 1997; Singer and Hickey 2000; Gibson et al 2005). The main source of compositional heterogeneity in mtDNA is mutational pressure correlated with a deficiency in the mitochondrial DNA repair system which, especially evident in some arthropods, is believed to be inefficient at replacing erroneous insertions of A nucleotides (and consequently of Ts on the opposite strand, Reyes et al. 1998). The consequence of this mutational pressure is that susceptible genomes are impoverished in G and C. Both nucleotide and amino acid based phylogenies may be misled by directional substitutions (Foster et al. 1999) resulting in the erroneous grouping of species that share a similar (but convergently evolved) mutational bias. G+C content varies significantly among different mtDNA metazoan groups, but is typically low in arthropods. Some Ecdysozoan lineages, such as some arthropods and the nematodes, are especially enriched in A and T and, in the absence of strong purifying selection, encoded proteins are enriched in amino acids encoded by A+T rich codons.

A second type of compositional heterogeneity, typical of mtDNA, is strand asymmetry correlated with the origin and direction of mtDNA replication. During replication, the lagging strand remains for a time in an unpaired state and is more susceptible to deamination (chemical conversion of As to Gs and Cs to Ts) than the leading strand (Reyes et al. 1998). This leads to the lagging strand being enriched in T and G while the leading strand is enriched in A and C. Strand bias is generally expressed in terms of GC and AT skew, expressed as a number between 1 and -1. A GC skew value of 0 indicate that the two strands have the same proportions of G and C, while a value close to 1 indicates that strand of interest is enriched in G. Variations in GC skew have been reported in all metazoan mitochondrial genomes (Saccone et al. 1999) and it has been shown in arthropods to represent a clear source of misleading phylogenetic signal (Jones et al. 2006, Hassanin et al. 2005). All genes in a mitochondrial genome usually have a similar G+C content, however, homologous genes from different organisms may have different GC (and AT) skew depending on the strand on which the gene is located (which depends on its direction of transcription) and its position relative to the origin of replication (Lavrov et al. 2000). It has been shown that both sources of compositional heterogeneity may play a key role in generating artefactual phylogenetic conclusions from the analyses of mtDNA sequences (Gibson et al 2005, Jones et al 2007, Masta Longhorne and Boore 2009).

Compositional heterogeneity is only one of the factors responsible for making mitochondrial-based deep phylogeny problematic: accelerated substitution rates may also play a role in masking and eroding the phylogenetic signal. These result in increased sequence divergence and a higher susceptibility to systematic biases (e.g. Felsenstein 1978, Brinkmann et al 2006). Mitochondrial genomes are also particularly prone to outgroup-effects, with different outgroups rooting in different parts of the ingroup tree (Cameron et al 2004, Rota Stabelli and Telford 2008). These characteristics, if shared by phylogenetically unrelated species, may be responsible for convergent evolution (homoplasy) and promote the dilution of the true phylogenetic signal. One effective approach to deal with these problems is to improve models of mitochondrial sequence evolution both at the nucleotide (Hassanin et al. 2005) and protein level (Abascal et al. 2007, Rota Stabelli, Yang and Telford 2009). More sophisticated evolutionary models such as the heterogeneous CAT model, which account for among site heterogeneity (Lartillot and Philippe 2004) and the derived CAT-BP model, (Blanquart and Lartillot 2008, Foster 2004) can be also useful to lessen the effects of various mitochondrial compositional biases. Another obvious approach is to enlarge the taxonomic sample: more taxa, in particular close to weakly supported nodes, may break problematic long branches and reduce the number of homoplasies responsible for long branch attraction type artifacts. This is particularly true for the ecdysozoans, which include some highly derived lineages, parasites for example, whose particular life style is responsible for bottle-neck events and therefore extreme acceleration of substituion rates.

1.4 Aims and objectives

Knowledge of ecdysozoan evolution is critical for comparative biological studies as these animals include the two most important invertebrate animal models (the fruitfly and the nematode worm) plus some emerging models such as the beetle *Tribolium castaneum*, the amphipod *Parhyale hawaiensis*, and the priapulid *Priapulus caudatus*. Furthermore, an international consortium is completing the genome sequence of five key ecdysozoan species (amphipod, horshoe crab, centipede, tardigrade and priapulid) and a tenable description of their relationships is fundamental to draw conclusions from the comparison of their genomes. Knowledge of ecdysozoan evolution also has relevant economic implications as they include some of the most important zooplankton (krill, copepods), parasites (lice, aphids, scales, filariasis), disease vectors (malaria, dengue, tse-tse), crop pests (weevils, fruitflies, thrips and lepidopterans) and in many cases consumers or biocontrollers of pests (ladybirds, parasitoid wasps, nematodes). These animals, in particular the arthopods, have been studied in detail for the last two centuries - Charles Darwin himself spent a whole decade classifying barnacle crustaceans - but many aspects of their affinities are still far from being resolved.

This thesis aims to resolve some of the problematic nodes within the ecdysozoans, in particular those of elusive myriapods (centipedes, millipedes and their kin), mysterious tardigrades (water bears) and bizarre onychophorans (velvet worms). I will use molecular approaches to study their relationships in particular the mitochondrial (chapter 3 and 4) and the EST (chapter 6) markers. I will address possible reconstruction problems such as stochastic and systematic errors - the first due to short datasets which do not contain enough phylogenetic information and the second correlated with model violations and consequent artefacts such as Long Branch Attraction. These problems will be tackled by increasing the taxon sampling at sensitive nodes and by assembling large datasets in order to reduce both stochastic and systematic problems (chapter 4 and 6). For the same reason I will employ sophisticated models of evolution and develop new ones in order to describe the evolutionary processes more accurately (chapter 2).

Chapter 2
Improving models of amino acid evolution for animal mitogenomic studies

2.1 Abstract

Existing empirical models of mitochondrial amino acid evolution have been derived from the comparison of taxonomically restricted datasets; they cover only two out of 30 metazoan phyla. Additionally, these models do not discriminate between structural or chemical characteristics such as highly hydrophobic transmembrane alpha-helices and hydrophilic loop regions. In this chapter I present two new, empirical amino acid substitution models for mitochondrial proteins based on a taxonomically diverse sample of metazoans and protein structural information. My aim is to generate models that better describe the evolutionary history of mitochondrial proteins of metazoans in order to overcome possible systematic biases and to generate more reliable phylogenies. I assembled a large alignment of mitochondrial-coded proteins from more than 100 metazoan species and estimated a reversible replacement matrix (MtZoa) using a Maximum likelihood approach. I also used secondary structure information to partition the alignment into two subsets, one containing hydrophobic and one hydrophilic sites. From the two partitions I estimated two corresponding substitution models, characterized by strikingly different amino acid frequencies and replacement rates and which are intended to be used simultaneously as a single model (MtHydro) when modelling correspondingly partitioned datasets. According to test of model fit, and in the absence of data partitions MtZoa is clearly preferable when diverse metazoan, lophotrochozoan and deuterostomes species are analyzed. Conversely, MtArt and MtREV are preferable for ecdysozoan and mammalian datasets respectively, suggesting that taxonomic representation may play a key role in the selection of the best model. Models that implement my partition strategy, either as empirical (MtHydro) or mechanistic (two distinct GTRs) fit all metazoan mitochondrial datasets better than any existing homogenous (non-partitioned) model, suggesting that my structural partitioning

strategy is a legitimate improvement. I also show that my models result in more reliable phylogenies. Finally, I show that when Likelihood scores of different models are penalized by the degree of parameterization (using BIC), all the datasets are fitted best by empirical models, suggesting that ultra-parameterization of mechanistic models may not be entirely justified by the increase in Likelihood.

2.2 MtZoa: a general metazoan empirical model

2.2.1 The need for taxa specific model.

In the past decade, some models of amino acid evolution have been explicitly designed for mitochondrial studies. A current problem with these models is that they are based on the comparison of restricted datasets, covering 2 of approximately 30 metazoan phyla: MtREV (Adachi and Hasegawa 1996) or MtMamm (Yang et al. 1998) are dominated by mammalian sequences and the recently released MtArt (Abascal et al. 2007) and MtPan (Carapelli et al. 2007) are both based on the analysis of arthropod-only datasets (Figure 2.1). These matrices reflect the substitution processes of either mammals or arthropods only and may be not appropriate for the analysis of other metazoan lineages, in particular lophotrochozoans and non-mammalian deuterostomes, for which many mitogenomic datasets are available, but few analyses have been conducted (Waeschenbach et al. 2006). Furthermore, the mtDNA genetic code varies to different degrees between different metazoan lineages. In the light of this, mitogenomic studies are in need of realistic models of evolution that best represent the evolutionary process and reduce systematic bias. In order to overcome systematic biases from restricted dataset sampling and to promote reliable metazoan phylogenies, I estimated MtZoa (figure 2.2), an empirical transition probability matrix based on the general reversible model (GTR, described in the chapter 1.3).

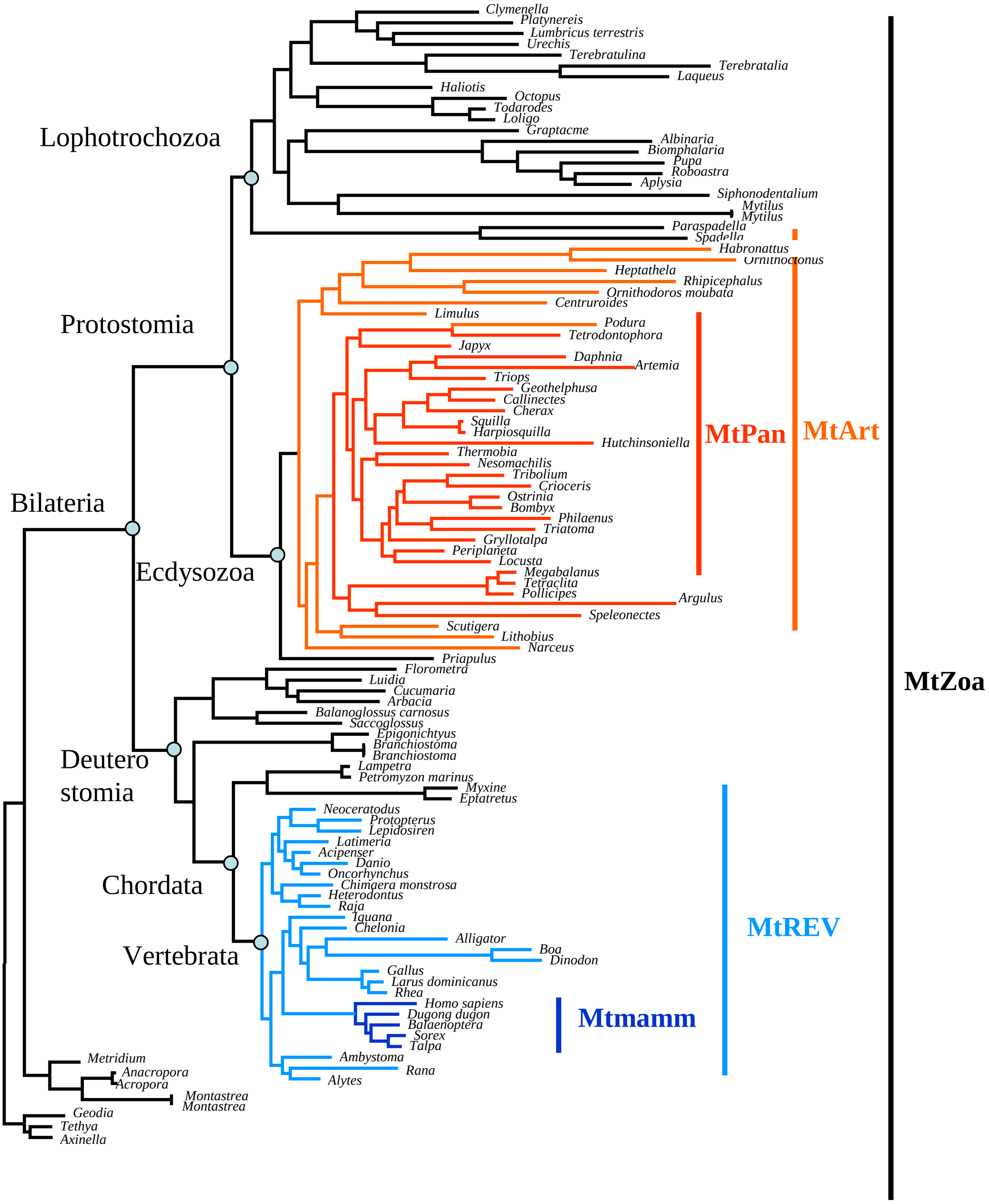


Figure 2.1. Phylogenetic tree of the 108 metazoan species used to infer the MtZoa model. Commonly used empirical models such as MtREV or Mtmamm (which are derived from vertebrates, in blue) and MtArt or MtPan (derived from arthropods in red/orange) are based on the comparison of restricted datasets. MtZoa is based on a larger and wider dataset, including lophotrochozoans, non-vertebrate deuterostomes and diploblastic metazoans. The topology was inferred using MrBayes under the MtREV model and some nodes have been constrained to reflect current knowledge of metazoan relationships; branch length was estimated using PAML (Yang 2004), during the inference of the model. Only the genus name is given.

2.2.2 Estimation of MtZoa

The accuracy of an empirically inferred replacement matrix depends on the accuracy of the tree topology and on the taxonomic sampling. The alignment should contain a phylogentically balanced sample of taxa avoiding overrepresentation of some of the metazoan phyla, which may result in the estimation of a biased replacement matrix. Bearing this in mind, I assembled a large 108 metazoan protein dataset from 13 phyla and the corresponding tree (figure 2.1) has been built in order to reflect current knowledge of metazoan relationships (Dunn et al. 2008 among others). In order to prevent the inference of a saturated replacement matrix, I excluded lineages characterized by accelerated substitution rate. I used the maximum likelihood approach implemented in PAML (Yang 2007) to estimate an empirical amino acid replacement model. The model assumes reversibility of the replacement process (GTR assumptions), so that the rate matrix Q={q_{ij}} satisfies the condition π_j r_{ij} = π_i r_{ji} for all the amino acid pairs, where π_j is the stationary frequency of amino acid j and r_{ij} is the replacement rate between amino acids i and j. More details on the inference of MtZoa are in chapter 7.2 Material and Methods.

2.2.3 Compositional and replacemental aspects of MtZoa.

The MtZoa model is characterized by replacement rates that differ considerably from those of MtREV (Fig. 2.2A) and of MtArt (Fig. 2.2B). Replacements involving cysteine, valine and serine are more common in MtZoa than in MtREV (white bars in Fig. 2.2A), while those involving histidine, asparagine and tyrosine are less frequent (grey bubbles). Stationary frequencies also differ: phenylalanine and valine are more frequent in MtZoa (white bars in Fig. 2.2A), while threonine is distinctly less frequent than in MtREV (grey bars). The diversity of the replacement information in mtREV and MtZoa can be explored in the figure 2.6 of page 52.

Compared to MtArt, MtZoa is impoverished in serine (grey bar in Fig. 2.2B), reflecting the differences between the invertebrate and the vertebrate mitochondrial genetic code (MtArt is based only on species with an invertebrate genetic code). Compared to MtArt, MtZoa is also enriched in alanine (whose corresponding codon GCN is GC rich) and impoverished in methionine and asparagine (corresponding codons, ATR and AAY are

AT rich; bars in Fig. 2.2B). Additionally, glycine, proline and arginine, whose codons are all enriched in G and C nucleotides, are slightly more frequent in MtZoa, while glutamate, isoleucine, tyrosine and phenylalanine (AT rich) are less frequent. Similarly and more importantly, most of the replacements involving AT rich amino acids (NKMIYF) are favoured in MtArt, while those involving GC rich amino acids (GARP) are favoured in MtZoa. This is a key difference, which seems to reflect the compositional properties of the arthropod mtDNA that is typically biased toward a high content of A and T nucleotides and suggest that MtZoa may be a more appropriate estimator than MtArt for the study of differently biased datasets such as lophotrochozoans and deuterostomes, which are less AT rich (Rota-Stabelli and Telford, 2008).

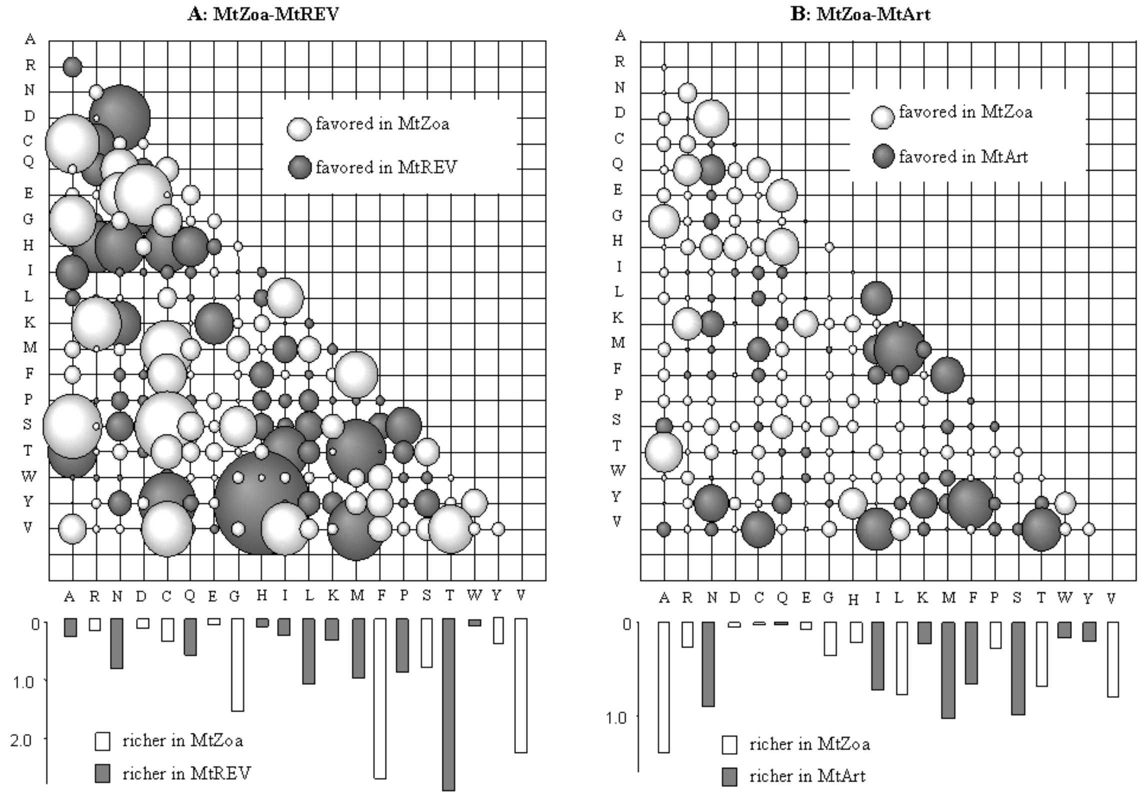


Figure 2.2. MtZoa differs to other models. Differences in replacement rates (bubbles in matrices) and stationary frequencies (bars) between (A) MtZoa and MtREV and (B) MtZoa and MtArt. Areas of bubbles are proportional to the absolute differences between replacement rates. The size of the bubbles in the legend correspond to a difference of 50. Length of bars corresponds to the absolute difference between stationary frequencies expressed as a percentage. White indicates a higher replacement rate or higher amino acid frequency in MtZoa and grey shows the reverse. Note that in B, amino acids whose codons are rich in A and T (NIKMFY) are enriched and more replaceable in MtArt than in MtZoa.

2.2.4 Test of MtZoa fit to various metazoan datasets

I used the **Akaike Information Criterion** (AIC) and **Bayesian Information Criterion** (BIC) methods to assess how MtZoa and other models fit diverse metazoan mitochondrial datasets. Both criteria penalize the model in a way that is proportional to the number of parameters and have been proved to be an appropriate tool for non-nested model selection (Posada and Buckley 2004). For the calculation of AIC and BIC, I used the harmonic mean of the log-likelihood of the trees sampled from the Bayesian analyses of 6 different mitochondrial dataset using MtREV, MtArt, MtZoa and the GTR model. Results are summarized in Table 2.1, which show for each dataset and model the mean log-likelihood, the AIC and the BIC values. According to this table, MtZoa is the preferred empirical model when diverse metazoan, lophotrochozoan and deuterostome species are analyzed. For these datasets, the differences in AIC or BIC values between MtZoa and MtArt or MtREV are high, in the range of, respectively hundreds and thousands. Conversely, MtArt and MtREV clearly better fit the ecdysozoan and the mammalian datasets respectively, reinforcing the view that the taxonomic level from which the matrices are estimated and different genetic codes (Abascal et al. 2006) may play a decisive role in the assessment of the model that best fit a certain dataset.

The log-likelihoods associated with the mechanistic GTR model (whose parameters have been deduced directly from the datasets) are clearly the highest for all the datasets. This is easily explained by the 208 free parameters of the GTR model (empirical models have none, because they are all pre-estimated), which are responsible for an inevitable increase in the log-Likelihood. Interestingly, at least one of the empirical models (MtZoa, MtArt or MtREV) shows a significantly better fit to the data for some (according to AIC) or all datasets (according to BIC). This can be explained by the reduced size of the alignments, whose amount of substitutional information is not enough to satisfactorily estimate a GTR replacement matrix. This result suggests that, in the cases of small datasets, the considerable computational time required for the estimation of all the parameters of a mechanistic GTR model is unlikely to be justified by a relatively moderate increase in the corresponding log-likelihood. It is remarkable that in some cases GTR required more than 100 times the computational time required by any of the empirical models, for the log-likelihoods of the sample tree to plateau.

Model	Statistic	Dataset					
		Metazoa	Lophotro chozoa	Ecdysozoa	Deutero stomia	Arthro poda	Mammalia
MtZoa	Δ lnl	-658	-293	-641	-62	-277	-2364
	AIC	900	170	866	BEST	462	4312
	BIC	BEST	BEST	751	BEST	463	3473
MtArt	Δ lnl	-1217	-706	- 266	-542	-46	-3094
	AIC	2018	996	116	960	BEST	5572
	BIC	1118	827	BEST	961	BEST	4933
MtREV	Δ lnl	-5607	- 3072	-3571	-1294	-2055	-628
	AIC	10798	5728	6618	2464	4018	840
	BIC	9898	5559	6503	2465	4019	BEST
GTR	Δ lnl	HIGHEST	HIGHEST	HIGHEST	HIGHEST	HIGHEST	HIGHEST
	AIC	BEST	BEST	BEST	292	324	BEST
	BIC	1977	2707	2761	3165	3208	2142

Table 2.1. Fit of different models to six metazoan mitochondrial datasets. For each of the datasets and models I show 3 statistics: the differences in log-likelihoods (Δ lnl), the AIC and the BIC (from top to bottom). The highest value of the log-likelihood is shown as HIGHEST and the highest value of AIC and BIC is shown as BEST. Other values are reported as the difference compared to these values.

2.2.5 Support for Mandibulata using MtZoa

I determined the consensus trees of the Bayesian analyses performed for the calculation of the harmonic mean for the AIC. For most of the datasets the tree topology using different models did not vary or only varied slightly. However, in the case of the Ecdysozoa dataset, different models support different topologies: while use of MtREV supports a group of paraphyletic Myriochelata (myriapods plus chelicerates, pp 1.00), and MtART does not resolve myriapod affinity, use of MtZoa or of GTR support a group of myriapods plus crustaceans/hexapods (Mandibulata hypothesis pp 0.90), in accordance with the morphological point of view (Telford et al. 2008). However, all models recover a group of unrelated long branched species (ticks, nematodes and tardigrades), suggesting that some aspects of the tree are subject to systematic errors.

2.3 MtHydro: a structural based partitioned model.

2.3.1 The structure of mitochondrial coded proteins.

The 13 mitochondrial genome encoded proteins are all subunits of four large trans-membrane protein complexes that lie in the inner membrane and participate in oxidative phosphorylation. The structure of these subunits consists of highly hydrophobic regions (mainly transmembrane alpha helices, as in the crystallographic structure of Complex IV shown in figure 2.3) alternating with hydrophilic regions (predominantly loops that lie in the mitochondrial matrix or in the inner membrane space). Transmembrane helices are characterized by a greater number of hydrophobic residues, while exposed loops show a higher frequency of hydrophilic residues (Goldman et al. 1996) leading transmembrane helices to be characterised by different amino acid frequencies and replacement patterns when compared to hydrophilic regions.

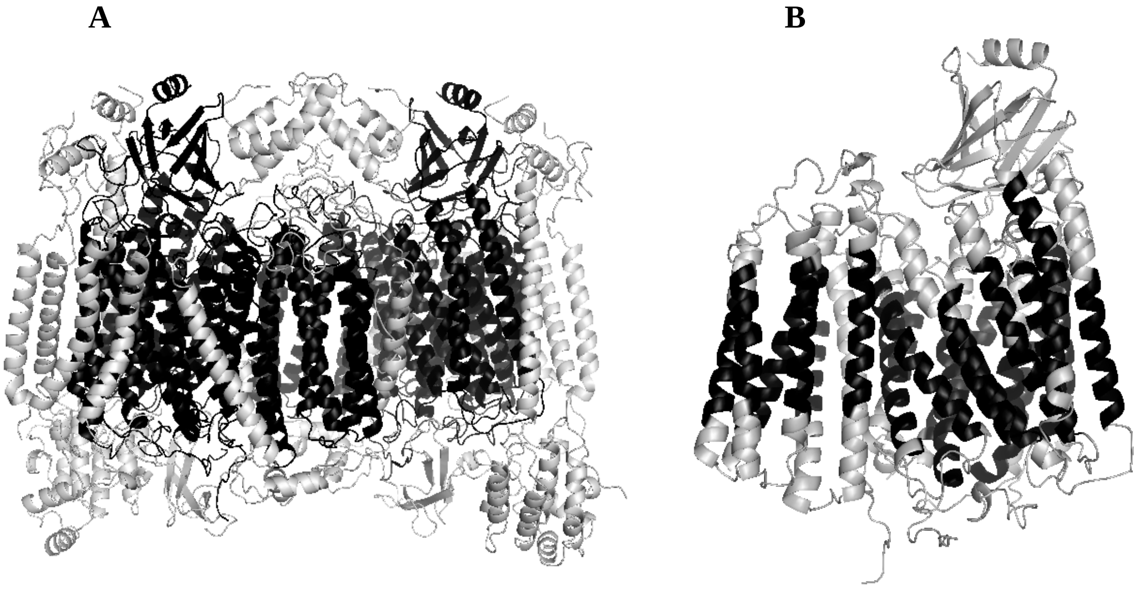


Fig 2.3. Crystal structure of the mitochondrial complex IV. (A) The complete complex, which is composed of two identical dimers; subunits coded by the mtDNA are in black and those coded by the nuclear DNA are in white. Note that the mitochondrial subunits reside in the internal part of the complex and therefore are expected to be highly hydrophobic. (B) A zoom on the three subunits coded by the mtDNA (Cox1, Cox2 and Cox3): my in-silico predictions of transmembrane residues are shown in black and clearly match the transmembrane helices. Structures have been drawn with PyMOLWin, using the secondary structures in databases (see chapter 7 for more details).

2.3.2 Room for improving existing partitioned models.

Neither empirical (MtREV, MtArt and the above described MtZoa) nor mechanistic (GTR) models account for likely heterogeneity of the substitution patterns among sites but rather make the assumption that all sites evolve under the same evolutionary process. However, two models of evolution described in the introduction, MT126 and CAT are clear improvements on conventional models as different parts of proteins are respectively described by different replacement rates (MT126) and different stationary frequencies (CAT models). MT126, however, has the disadvantage of being implemented in a likelihood framework, rather than a Bayesian one, and is not accessible to the popular Bayesian inference software MrBayes (Huelsenbeck and Ronquist, 2001). Furthermore, only the replacement matrix for the transmembrane class of MT126 has been generated from mitochondrial data, while other classes were from nuclear coded proteins, leaving space for the development of models more appropriate to the analyses of mitochondrial data. CAT models, which do not need a pre-specification of the distinct classes of sites, are useful when no structural information is available. However, this is not the case for mitochondrial proteins for which reliable transmembrane information can be obtained: four of the longest of 13 subunits (COX1, COX2, COX3 and CYTB) have been characterized with crystallographic studies (Tsukihara et al. 1996) and the combined use of different bioinformatic methods allows the confident deduction of secondary structure information (Liò 2005).

2.3.3 Pipeline and estimation of the structural model MtHydro.

In order to generate an empirical model that takes structural properties into account while being based on a large taxonomic sample, I assembled a large alignment of the whole mt-proteome from 100 diverse metazoan species and used structural information to partition the alignment into hydrophobic and hydrophilic subsets. I used information from crystallographic structures, where available, and in-silico predictions using three different methods, to split a metazoan alignment into hydrophobic (Figure 2.4D) and hydrophilic (2.4E) partitions. Interestingly, independent predictions carried out on the two distant metazoan species (the cow *Bos taurus* and the horshoe crab *Limulus Polyphemus*) overlapped for the majority of the sequences, suggesting a high degree of structural conservation within the metazoan mt-proteome (last rows of alignment in figure 2.4C). I also noticed that bioinformatic predictions were substantially similar to information from crystallographic structures: in figure 2.3B, I have highlighted in black the transmembrane helices predicted with bioinformatic methods on the crystallographic tertiary structure: predictions correspond in all cases with the transmembrane alpha-helices. This reassured me of the accuracy of the bioinformatic predictions that were the only available method for the analyses of 9 out of 13 proteins. According to in-silico predictions and 3D structure (and following personal communication with crystallographers form Birkbeck college), the central part of transmembrane helices lies in the membrane, while the tips of the helices may lie on the surface of each subunit where accessibility to the solvent or other proteins is higher. Consequently, helix tips are characterised by a higher frequency of hydrophilic residues (see first seven alignment positions of figure 2.4C). Corresponding sites were therefore considered part of the hydrophilic partition. For each of the two partitioned datasets, I estimated a distinct replacement model named MtPhobic (figure 2.4 F) and MtPhilic, (figure 2.4 G), using, as in the case of MtZoa, the maximum likelihood approach implemented in PAML (Yang 2007) and the GTR assumptions ($\pi_j\ r_{ij} = \pi_i\ r_{ji}$).

The two distinct matrices are intended to be used simultaneously in phylogenetic studies with a related hydrophobic/hydrophilic partition as a dual model named MtHydro. Notably, the partitions can be modelled by two distinct mechanistic GTR models, emancipating the MtHydro models from the (relatively) limited taxonomic range they have been estimated from. More details of secondary structure prediction and model inference can be found in chapter 7 materials and methods.

2.3.4 The two sub-matrices of MtHydro

The two sub-matrices of MtHydro show extremely different amino acid replacement rates and are characterized by striking differences in amino acid frequencies. According to values of π in figure 2.4 F and 2.4 G, MtPhobic is rich in hydrophobic amino acids (I, L, M, F, V) , while MtPhilic composition is more widely distributed and relatively enriched in charged hydrophilic amino acids (R, N, D, E). This is in accordance with the structural characteristics of the regions from which the two matrices have been estimated.

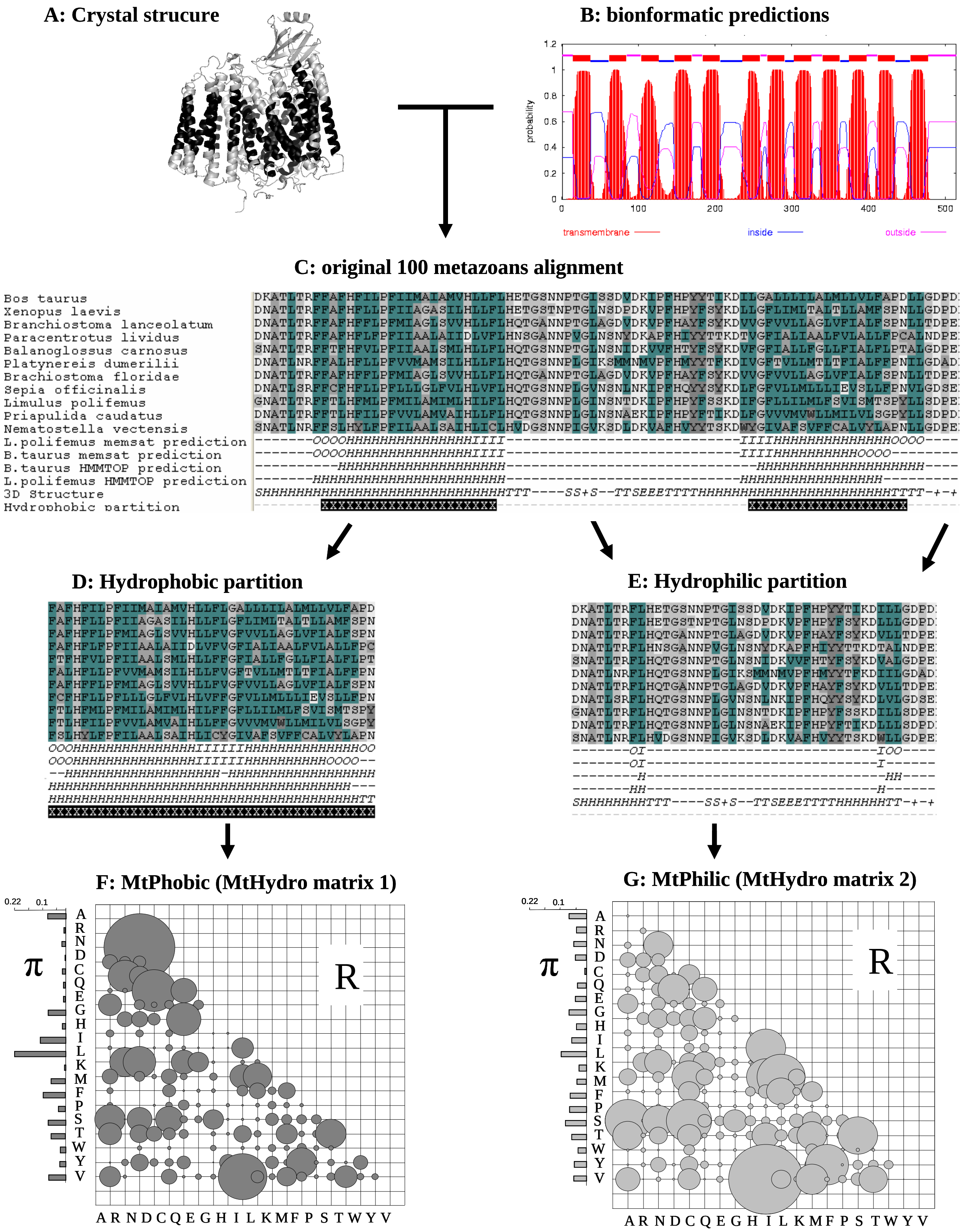


Figure 2.4. Pipeline for the estimation of the MtHydro empirical model. An alignment of concatenated mitochondrial coded proteins from 100 metazoans species (C) has been partitioned into a hydrophobic (D) and hydrophilic (E) sub-alignment, on the basis of crystallographic 3D structure (A) and bioinformatic predictions (B). Amino acids in the alignments have been coloured accordingly to their hydrophobicity, with hydrophobic residues in blue, hydrophilic in light grey and intermediate amino acid in dark grey. Note that the hydrophobic partition contains mostly hydrophobic blue residues. The two partitions have been used to estimate distinct empirical sub-models called MtPhobic and MtPhilic (respectively F and G), composed of a replacement matrix R and the stationary frequencies π. Areas of bubbles in matrices R are proportional to substitution rates. The two sub-models are intended to be use simultaneously on a pre partitioned dataset as a dual model called MtHydro.

The pattern of replacement rates is also very different: for example substitutions between hydrophilic amino acids are favored in MtPhobic (grey bubbles in figure 2.5), while substitutions involving C, S, H and the hydrophobic amino acids are favored in MtPhilic (white bubbles in figure 2.5). These differences are confirmed by comparing amino acid replacement groups (figure 2.6) using the AIS method (Kosiol et al. 2004): C and H, but also hydrophilic N and D show a different replacement behavior in MtPhilic than in MtPhobic.

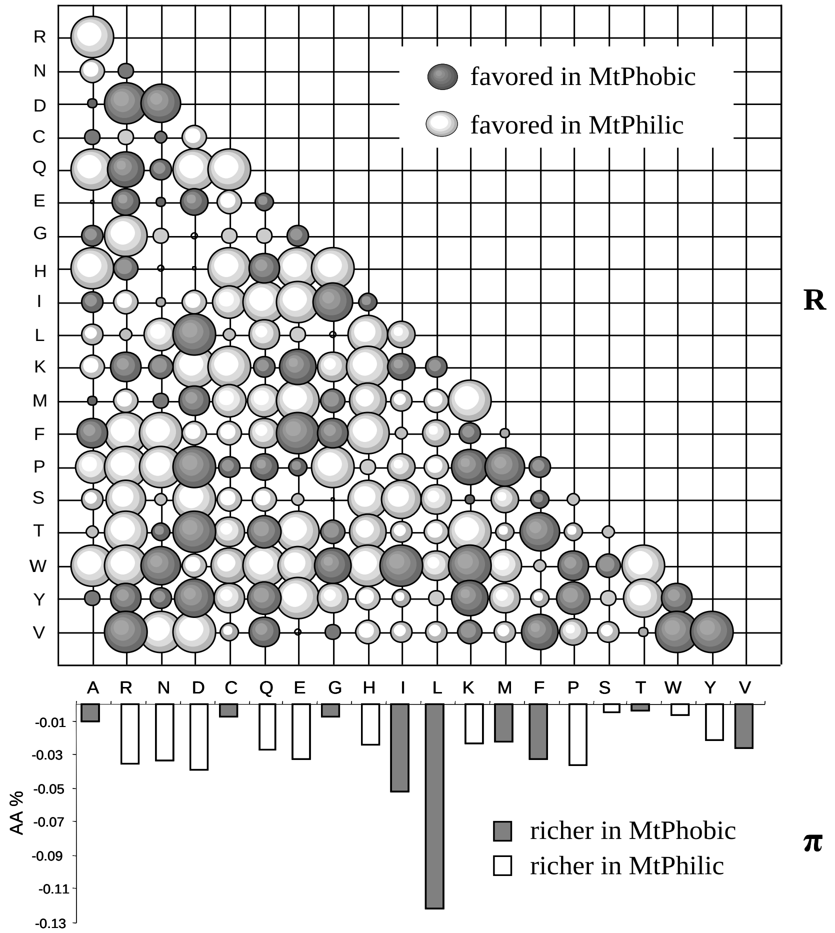


Figure 2.5. The two sub-matrices of MtHydro. Relative normalized differences between the replacement rates of the two MtHydro sub-matrices (R) and differences in their stationary frequencies (π). Grey indicates that a replacement is favoured in the hydrophobic sub-matrix MtPhobic and white the reverse.

Some of the differences in the replacement rates are not intuitively explainable: for example hydrophilic residues (the most polar amino acids: R, N, D, E, Q, K) might be expected both to occur more often and to replace one another more often in a hydrophilic context (the MtPhilic model) than in a hydrophobic one. However, while more frequent in the hydrophillic domains, these amino acids are more exchangeable in the MtPhobic matrix, which is derived from hydrophobic regions. A possible explanation comes from considerations of a structural nature: hydrophilic charged residues are scarce in hydrophobic alpha-helices, but they form stable polar bonds and are essential for helix-helix interactions and tertiary structure stabilization; consequently they should be uniquely replaced by other hydrophilic residues to preserve inter/intra-helices bonds and correct protein folding.

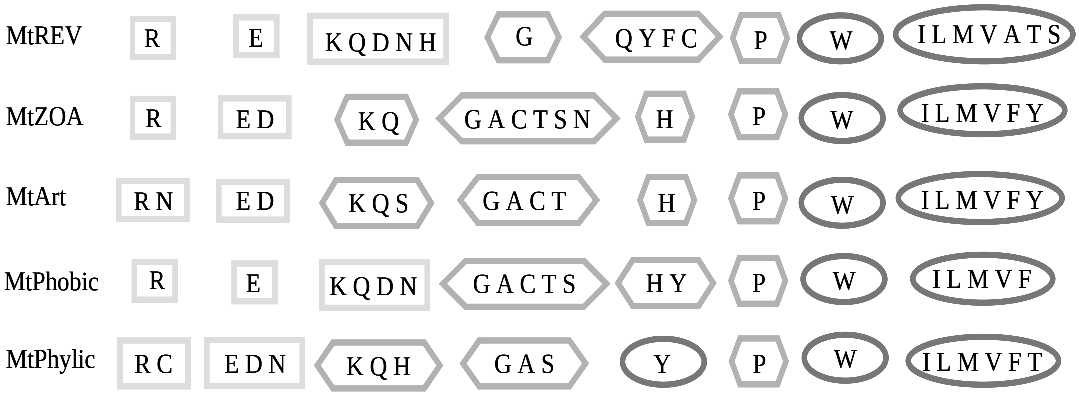

MtREV R E KQDNH G QYFC P W ILMVATS
MtZOA R ED KQ GACTSN H P W ILMVFY
MtArt RN ED KQS GACT H P W ILMVFY
MtPhobic R E KQDN GACTS HY P W ILMVF
MtPhylic RC EDN KQH GAS Y P W ILMVFT

Figure 2.6. Replacemental properties of various matrices. Amino acids have been grouped in eight classes according to their probability of change among those of the same group, using the program AIS (Kosiol et al. 2004). Squares indicate groups of mostly hydrophilic amino acids, circles indicate groups containing only or prevalently hydrophobic amino acids and hexagons indicate moderately hydrophilic or a mixed state of amino acids. Note that MtREV is very different from MtZoa or MtArt and that the two sub-models of MtHydro are also different, with MtPhobic sharing more similarities with MtREV.

As mentioned, I used the AIS method (Kosiol et al. 2004), which identifies groups of amino acids with a high interchange probability, to compare the two MtHydro sub-matrices with other empirical models commonly used in phylogenetic analyses, including MtZoa. All the models share a similar exchangeability behavior with respect to hydrophobic residues (right part of figure 2.6), while other residues are differently grouped in different models. MtZoa, the model described in the first section of this chapter, and MtArt are quite similar, but both dissimilar to MtREV. Interestingly, MtPhobic shares more similarities with MtREV, while MtPhilic with MtZoa or MtArt. This may be partially explained, by the fact that MtREV is estimated from a vertebrate dataset, whose mitoproteome is overall more hydrophobic than these of protostomes, and in particular for the ND5 subunit (Liò 2005); consequently MtREV should be considered more “hydrophobic” than more general models such as MtZoa, which are estimated from a wider metazoan sample.

2.3.5 Test of models fit to various metazoan datasets

I compared MtHydro and other evolutionary models using the AIC method, which penalizes the log-likelihood (lnl) values proportionally to the number of parameters in the model (table 2.2). In table 2.1, I compared the fit of MtZoa and other homogenous models using BIC and the difference in LnL; the two latter values only confirmed results of the AIC for MtHydro comparisons and have been removed for sake of clarity form table 2.2. Similarly to table 1, AIC has been estimated using the harmonic mean of the LnLs of trees sampled during the Bayesian analyses of six metazoan mitochondrial protein datasets. In addition to table 2.1, I have analysed the datasets both as un-partitioned and as hydrophobic/hydrophilic partitioned and used seven different models.

In all cases, models which implement my hydrophobic/hydrophilic partition strategy (empirically using MtHydro or mechanistically using two distinct GTRs) fit all datasets better than the existing MtREV, MtArt, MtZoa and the unpartitioned GTR models. I also modelled the hydrophobic/hydrophilic partitioned datasets with “nuclear” empirical matrices: I assigned the hydrophobic partition to the JTT “transmembrane” matrix and the hydrophilic partition to the WAG “globular” matrix (Jones et al. 1992, Whelan and Goldman 2001). Interestingly the corresponding AIC values are the highest (lowest fit) among the models tested, suggesting that mitochondrial amino-acid substitution dynamics are highly specific. Unfortunately, it was impossible to test my model against mixture models such as CAT or against MT126, because of the nature of their implementations and the differences in how the likelihood is calculated in different programs.

As previously observed in table 2.1 (comparison of homogenous models only) in some cases empirical models fit datasets better than mechanistic GTR models (table 2.2): the partitioned MtHydro model fits the deuterostome and arthropod datasets better than two GTRs and the lophotrochozoan dataset better than a single GTR model. As also previously shown, MtArt and MtZoa fit the ecdysozoan and the deuterostome dataset respectively better than does the single GTR (previous section and Rota Stabelli et al. 2009). This suggests that in some cases the ultra-parameterization of mechanistic models is not justified by a relatively modest increase in LnL. Moreover, the LnL of trees sampled during Bayesian analyses using GTR models, took up to 2 million generations to plateau, while empirical models reached a plateau in a few thousand generations. This is easily explained by the high number of parameters of the replacement matrix that the GTR models have to estimate. For the same reasons GTR analyses are much slower per generation than those using empirical models. These considerations make the GTR analyses of short mitochondrial datasets extremely time consuming and in some cases of little value if, as according to AIC, they are in some cases comparable with empirical models in their fit to the data.

Model		Partioned	DATASET						Parameters
			Metazoa (N=44)	Lophotrocozoa (N=24)	Ecdysozoa (N=30)	Deuterostomia (N=30)	Arthropoda (N=23)	Mammalia (N=41)	
Empirical	MtHydro	YES	1102	60	790	**BEST**	**BEST**	3568	4
	JTT/WAG	YES	15912	8122	12712	7698	8708	8830	4
	MtArt	NO	3018	1548	710	1946	154	6112	2
	MtZoa	NO	1900	722	1460	986	616	4652	2
	MtREV	NO	11798	6280	7212	3450	4172	1180	2
Mechanistic	GTR/GTR	YES	**BEST**	**BEST**	**BEST**	852	268	**BEST**	416
	GTR	NO	1000	552	594	1278	478	340	208

Table 2.2. Fit of different models (AIC values) to six metazoan datasets. Models which implement the hydrophobic/hydrophilic partition (either empirical MtHydro or two mechanistic GTRs) fit different mitochondrial datasets better than non partitioned dataset. The lowest value of AIC is set as the BEST and corresponds to the model which best fits the corresponding dataset; other values are set as the difference from the lowest AIC.

2.3.6 MtHydro lessens LBA artefacts: applications to deuterostomes

In some of the tree searches, models implementing the hydrophobic/hydrophilic partition support different tree topologies than other models, in particular MtREV, which has been the model of choice for mitogenomic studies.

The most clear example is from the mammalian dataset: non partitioned models MtREV and mechanistic GTR support a sister relationship between the enigmatic scaly-tailed flying squirrel *Anomalurus sp.* and the Hystricognathi (the infraorder including the guinea pig), in accordance with the source from which the dataset has been taken (Horner et al. 2007). Conversely, MtHydro (and the corresponding two mechanistic GTRs model) support an alternative position of *Anomalurus*, as sister to a group composed of *Jaculus jaculus* and the Muroidea (mice and rats) a position consistent with analyses of nuclear encoded genes and concatenated nuclear/mitochondrial genes (Adkins et al. 2003, Douzery et al. 2003, Montelard et al. 2008) and one favoured when the fastest evolving amino acid sites were removed from the dataset in the original publication (Horner et al. 2007). While encouraging, it is somewhat questionable to assess which model has more credibility: accordingly to table 2.2, MtREV supports the dataset better than MtHydro, but the partitioned GTR model outperforms all other models. In any case, this is a clear example of how our simple partitioning strategy results in a different and probably more accurate tree topology. Another example comes from the deuterostome dataset: MtREV supports a sister relationship between the sea urchin and the holothurians, while MtHydro (and all the other models which fit the dataset better than MtREV) support a sister relationship between the sea urchin and the sea stars, with the holothurians as sister to this group (trees not shown).

Furthermore, an analysIs of a deuterostome mitochondrial dataset (Bourlat et al. 2009) suggests that use of MtHydro slightly weakens the long branch attraction (LBA) between urochordates and basal non bilateral metazoans, which is strongly supported by MtREV. However, use of the CAT heterogeneous model and in particular of the related CAT-BP model overcomes the LBA, suggesting the superiority of CAT over MtHydro in this case (see discussion, chapter 8.1 for more details).

2.4 Conclusions

In order to better describe the evolutionary history of mitochondrial proteins and to promote more reliable metazoan phylogeny estimation, I have estimated MtZoa, which is a general transition probability matrix. Tests of model fit suggest that MtZoa should be used for datasets containing diverse or basal metazoan groups and for the analysis of deuterostome and lophotrochozoan datasets . Conversely, MtArt and MtREV should be used respectively for ecdysozoan and mammalian datasets. As a general rule, my results advocate that the taxonomic set from which models are estimated plays a decisive role in the assessment of the best fit to datasets and that, in the case of poor phylogenetic signal or problematic nodes, the use of a more appropriate model which reflects the evolutionary pattern of the given taxonomic sample, results in a much higher likelihood, a better fit to the dataset and may consequently help lessen possible systematic errors.

As mitochondrial coded proteins are characterized by a clear alternation of transmembrane helices and hydrophilic regions, I used this information to estimate two additional replacement matrices which are intended to be used simultaneously as a single model called MtHydro in a pre-partitioned dataset. An interesting point of my partitions is that the two sub-alignments can also be modelled by two distinct mechanistic GTR models, emancipating the MtHydro model from the taxonomic range it has been estimated from. I have used the AIC approach to compare the fit of MtHydro and other models to different metazoan mitochondrial protein datasets and found that models which implement my partitions, either as empirical (MtHydro) or mechanistic (two distinct GTRs) fit all metazoan mitochondrial datasets better than existing mitochondrial models. I also show that the use of my partitioned models, in contrast with non-partitioned ones, recovers topologies which are more in accordance with nuclear encoded genes. Results suggest that my structural partition is a simple and legitimate improvement that may help in reducing possible systematic biases in mitogenomics and promote the generation of a more reliable phylogeny of metazoans.

Tests of fit to the model also suggest that empirical models may be preferable to the mechanistic GTR models (table 2.1 and 2.2). My interpretation is that a moderate increase in the log-likelihood of GTR trees, may not justify the much larger amount of time needed for computation. This is particularly true for taxonomically small datasets (such as the ones I used for the test of model fit) which may not contain sufficient substitutional information for a correct estimation of the replacement rates of the GTR model.

Chapter 3
The effect of outgroup choice and the affinity of myriapods

3.1 Abstract

The choice of an appropriate outgroup is a fundamental prerequisite when the difference between two conflicting phylogenetic hypotheses depends on the position of the root. This is the case for the myriapods that may group either with Pancrustacea, forming a clade called Mandibulata in accordance with morphological characters, or with chelicerates to form Myriochelata as has recently been proposed by molecular phylogenies. The importance of a suitable outgroup is highlighted by the possibility that the node describing myriapod affinity may be subject to stochastic and/or systematic error related artefacts. In order to understand the impact that outgroup choice may have on phylogenetic reconstruction, I have investigated compositional heterogeneity and genetic distance in mtDNA sequences of several different outgroups to the arthropods, selected from deuterostomes, lophotrochozoans and non-arthropod ecdysozoans, and have used them to root a phylogenetically balanced and compositionarily homogeneous arthropod dataset. Results indicate that some outgroups, in particular from lophotrochozoans, nematodes and an onychophoran, have G+C content and strand specific biases which are very different from those of arthropods, suggesting that the use of such outgroups may interfere with the stationarity of the model and might create a random outgroup effect. I propose a new metric (called the skew index) which can be used for comparative mitogenomic studies and have defined a set of a priori criteria for the identification of optimal outgroups (and ingroups). Inference of phylogeny shows that use of phylogenetically distant and compositionally distinct lophotrochozoans as outgroups supports Myriochelata and use of more closely related, while fast evolving nematodes provide contrasting signal. Optimal outgroups selected according to our multi-criteria selection supports Mandibulata. In conclusion, support for the Myriochelata hypothesis from mitochondrial sequences may depend on the nature of the

outgroup sequences rather than a true phylogenetic signal. I advocate a careful analysis and an objective choice of outgroup when dealing with highly derived sequences, such as mitochondrial genomes.

3.2 A matter of outgroup position

As the sister group relationship between crustaceans and hexapods is well accepted (with myriapods and chelicerates lying outside this 'pancrustacean' group) the question of the affinities of myriapods depends entirely in the position of the outgroup (see figure 1.4 C). It is clear that the choice of a correct outgroup may have a significant impact on phylogeny estimation around this node.

When inferring arthropod phylogeny from mtDNA genes, researchers to date have rooted the tree using annelids or molluscs as outgroups, possibly because annelids, at least, were thought to be the true sister of arthropods according to the now discredited Articulata hypothesis, but more obviously because no suitable ecdysozoan sequences have been available. Among the ecdysozoans, several fully sequenced nematode mtDNAs exist, but have not been selected as outgroups because of their high substitution rate: a divergent outgroup may generate difficulties in the aligning process, loss of signal, random outgroup effects and will tend artifactually to attract fast evolving (Philippe et al. 1998) and/or compositionally similar (Foster et al. 1999) ingroup species towards the base of the tree. As a consequence, many authors have been forced to choose between different sorts of inadequate outgroups: phylogenetically close, but genetically distant nematodes or phylogenetically distant lophotrochozoan annelids and molluscs. Use of lophotrochozoan outgroups consistently supports the Myriochelata hypothesis. While it has been reported that nematodes may perform better than lophotrochozoans as outgroups in mtDNA based arthropod phylogeny (Cameron et al. 2004), more recent comparative analysis suggest than nematodes are characterized by both fast evolving nuclear and mitochondrial genes, discouraging their use as arthropod outgroups (Webster et al. 2006).

Accelerated substitution rates and composition heterogeneity may have also diluted the natural phylogenetic signal and left the actual sequences prone to systematic and stochastic errors, the latter due to the relatively small size of mitogenomes. In the light of this, and bearing in mind that the affinity of myriapods depends entirely on the position of the outgroup, a careful analysis and an accurate selection of available ingroup and outgroup sequences may help to lessen the problems involved with mitogenomics and ultimately help to clarify arthropod relationships.

3.3 Compositional aspects of outgroups

3.3.1 Different outgroups to the Arthropoda have different compositional characters.

In order to understand the impact that outgroup choice may have on phylogenetic reconstruction, I investigated compositional heterogeneity and genetic distance in mtDNA sequences of several different outgroups to the arthropods. I chose outgroups with the invertebrate mitochondrial genetic code (code 5 in NCBI) from the phyla Annelida, Mollusca, Brachiopoda, Chaetognatha, Nematoda and Cephalochordata (in table 3.1, for more details see chapter 7.3).

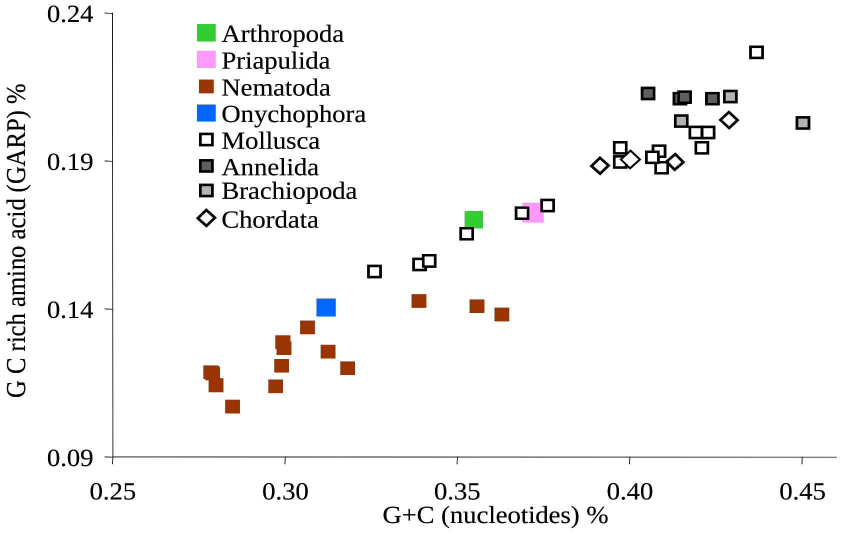


Figure 3.1. Compositional properties of metazoan species considered in this chapter. Nucleotide (G+C %) and amino acid content (GC rich amino acids %) are highly correlated in metazoan mtDNA (R=0.96). G+C % is calculated on 1st and 2nd codon positions of conserved sites and GC rich amino acids are calculated on the frequency of G, A, R, P amino acids (codons with G and/or C at both first two positions). The arthropod value is calculated on the average of 21 selected arthropods species. Please note that Priapulida, and not Onychophora have compositional characters similar to the arthropods. Compare these values with figure 4.2, which is based on a similar, but enlarged dataset and uses a similar colour code.

For each of the outgroup sequences, I analysed the nucleotide content in terms of G+C% and the amino acid content in terms of the percentage of amino acids that are coded by codons that have G or C in both the first and second codon position (amino acids G, A, R and P in figure 3.1). The two values are strongly correlated (Figure 3.1) showing that amino acid content is influenced by the nucleotide content which therefore has to be considered. According to Figure 3.1 (but see also Table 3.1), nematodes and the onychophoran are characterized by high A+T % (low G+C %) while most of the lophotrochozoans (Annelida, Brachiopoda, some of the Mollusca, Echiura) have higher G+C % (between 0.4 and 0.43) compared with arthropods (0.36). Some of the molluscs and the priapulid show very similar G+C % to that of the arthropods. The plot of figure 4.1 in the next chapter gives a better view of the arthropod composition than the plot of figure 3.1, which shows an averaged value calculated over all the sampled arthropods. The information from the two plots are consistent, however: the non-ecdysozoan outgroups are GC enriched, the nematodes and the onychophorans are AT enriched and the priapulids show an intermediate state.

3.3.1 Strand asymmetry.

As discussed in chapter 1.3, mitochondrial genomes can be extremely heterogeneous in the distribution of nucleotides between the two stands, a characteristic known as strand asymmetry; accordingly I calculated strand asymmetry for each of the outgroups in terms of GC skew. This value was calculated for each gene independently (right part of Table 3.1) and for the concatenated genes, as usually done in comparative genomic studies. To highlight the similarity with arthropod strand bias I plotted, for each of the coding genes, the difference between the GC skew value of a given outgroup and the corresponding mean GC skew in arthropods. Some of the outgroups are characterized by a skew that differs strongly from that of arthropods (Fig 3.2 A); these taxa were excluded from subsequent analyses.

The majority of outgroups I selected for phylogenetic inference have a similar strand asymmetry to that of arthropods (e.g. figure 3.2 B). Priapulid strand asymmetry is the most similar to that of arthropods (Figure 3.2 C), while the onychophoran is dissimilar in part due to numerous genomic rearrangements.

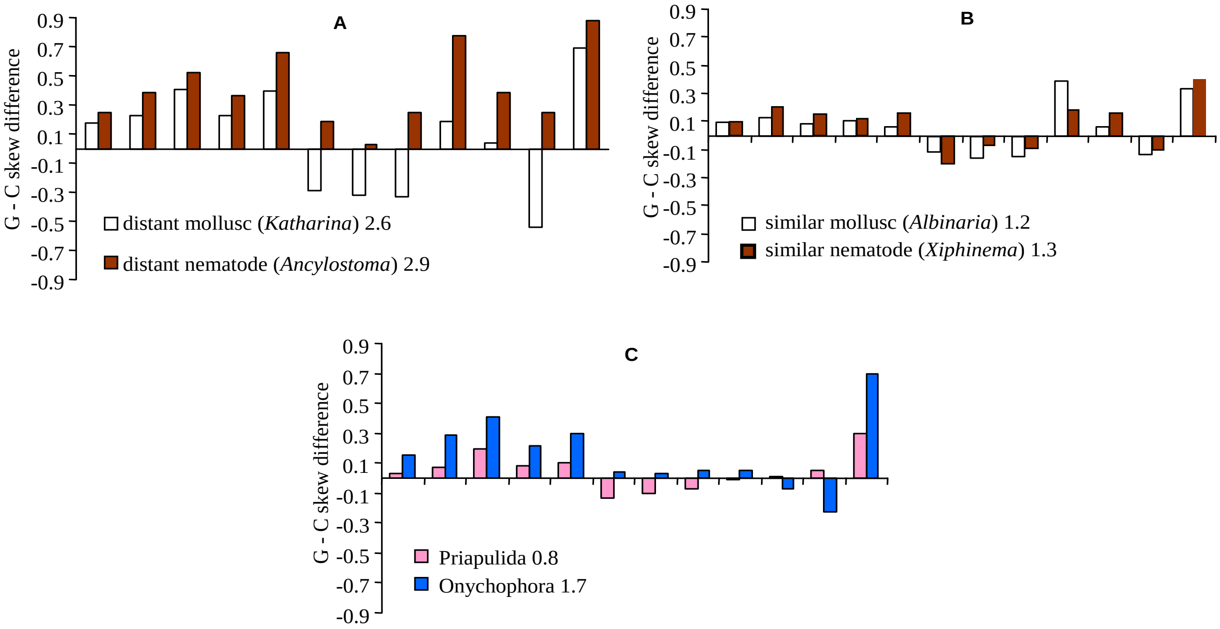


Figure 3.2. Different strand asymmetry in different outgroups. Difference between GC skew of selected outgroups and the mean of arthropods GC skew calculated for each gene of the mtDNA. The order of genes is as in the Ancient Arthropod Gene Order (AAGO, see chapter 4) starting with COX1. A: Most of the lophotrochozoan and nematode outgroup species have a skew profile very different to the arthropod one. B: Selected lophotrochozoan outgroups (mollusc) and some nematodes have GC skew values more similar to the arthropods ones. C: Priapulida is the outgroup showing the smallest difference with respect to arthropod skew values for all of the genes. Color code as in figure 3.1 and 3.3.

3.3.2 The new metric " skew index" and its utility in mitogenomic studies

The skew summed over all genes may be meaningless as two taxa may have an opposite skew in every gene yet end up with the same mean skew (compare the mollusc *Katharina tunicata* in table 3.1 and figure 3.2 A)**.** As the graphical representation of skew along many genes is problematic, I captured the similarity of skew values in a single measure I called the "skew index". This value gives a direct description of how much the overall strand asymmetry of one species differs from an hypothetical genome without strand asymmetry. The skew index is defined as the absolute sum of the GC skew value (GC) for each of the genes (j), normalized for the length of the gene (length j) and the number of genes (n):

SI (Skew Index) $= [\sum_{j} (\text{abs } GC_j) / (\text{length } j / \text{total length} *100)] / n$.

One application of the Skew Index is the comparison of different mitogenomes and indeed this was my aim when generating this index. The skew index can be calculated relatively to a reference mitogenome, and, in the case of this study, I used the mean calculated over various arthropod species (see materials and methods 7.4.3 for a justification of the averaging). This “relative skew index” is then defined as the absolute sum of the differences between GC skew value of the considered species (GCs) and the mean of arthropods (GCa) for each of the genes (j), normalized for the length of the gene and the number of genes:

RSI (relative SI)= [$\sum_j$ (abs (GCs_j-GCa_j)) / (length j / total length *100)] / nj.

This value gives a direct description of how much the overall strand asymmetry of one species differs from the mean of arthropods and may be considered as a concise description of a skew plot. A low skew index indicates a species with a skew profile similar to that of arthropods (e.g. *Priapulus* RSI=0.8), while a high value corresponds to species with a very different strand asymmetry to that of the arthropods, such as some nematodes (RSI=3.00).

Interestingly, I found that Skew index is positively correlated with ML genetic distances (Figure 3.3A), suggesting, not entirely surprisingly, that species with a greater skew difference from arthropods are also more genetically distant and that skew index may be considered a useful predictor of outgroup adequacy. On the other hand the skew index is not correlated with the G+C content of the mitochondrial genome, similarly calculated as the difference between the G+C content of each outgroup and the arthropod mean (Figure 3.3B). This suggests the importance of accounting for strand asymmetry, in addition to G+C%, in the selection of adequate taxa for phylogenetic purposes.

3.3.3 Best Putative outgroup

Based on my *a priori* assumptions that short genetic distance and similar G+C content and GC skew are indicative of optimal outgroups, my results show that *Priapulus caudatus* is the best available outgroup. First, its mtDNA has the lowest genetic distance to *Limulus* (0.284 in table 3.1), suggesting that it is characterised by a slow mutational rate and it will be least prone to long branch attraction. Second, its compositional characters are very similar to those of arthropods, both for G+C% (in

Figure 3.1) and GC skew (Skew index 0.8 and see Figure 3.2 C). These aspects should reduce the possibility of non-stationarity of the nucleotide/amino acid sequences during inference of phylogeny. Last but not least, the priapulid is an ecdysozoan and consequently an ideal root for the arthropod that are themselves ecdysozoans.

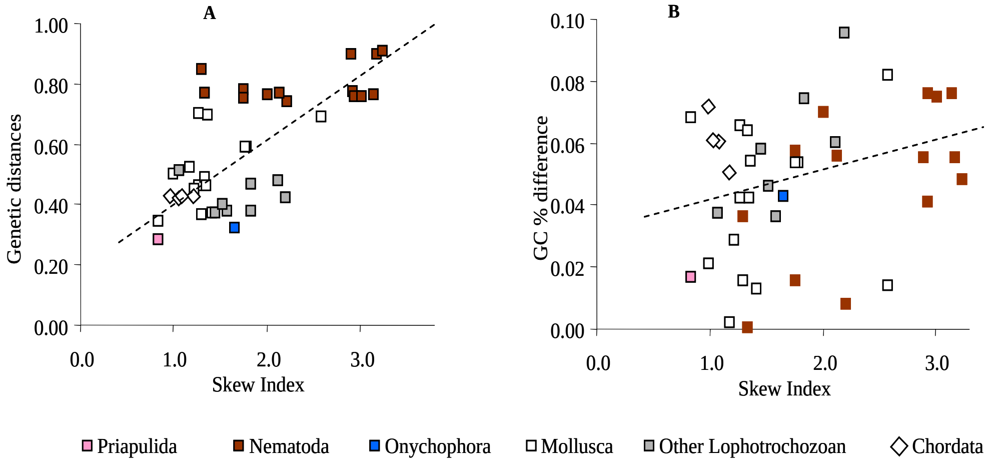


Figure 3.3. Utility of Skew index in mitogenomics. (A): Skew index is correlated with genetic distances in metazoan mtDNA (R=0.743). Correlation could suggest that species with strand asymmetry compositional bias have a higher probability of accumulating more substitutions or more simply that a difference in skew leads to a difference in sequence. In either case the skew index may be considered a good predictor of outgroup adequacy. (B): Skew index is not (or very poorly) correlated with nucleotide content (R=0.298), calculated as the difference between each of the outgroups and the mean of arthropods. This suggests that strand asymmetry of genes is not correlated with the mutational pressure that is responsible for GC content, consequently both sources of bias ought to be taken into account. Colours code is given and is the same as in figure 4.1, 3.2 and 4.6.

Onychophorans with tardigrades and arthropods are likely to form a clade called Panarthropoda (Nielsen 2001) and consequently could be considered a closer and more valuable outgroup than the priapulid. However, the onychophoran genome is AT enriched (G+C % 0.31) in a way that resembles derived nematode species (see Figure 3.1, but also figure 4.2, pag. 75). This fact, together with the tendency of onychophoran nucleotide sequences to branch within paraphyletic pancrustaceans (tree not shown) and a fairly high value of skew index (1.7), suggest they may not be the ideal outgroup to root the arthropod tree. The outgroup study of this chapter, has been carried out prior to the analyses of the next chapter, which accessorily encompass tardigrades. Inspection of tardigrades compositional characters (figure 4.2 and 4.3 in next chapter), as well as their hypothetical panarthropod affinity, suggests that tardigrades may be a good candidate to root the arthropod tree. However, tardigrades, are clearly characterised by an accelerated rate of evolution (table 4.1, pag. 77), a fact which may promote systematic type reconstruction errors, in particular long branch attraction (LBA) one. This possibility is clearly manifested by the inference of phylogeny in the presence of tardigrades (throughout the next chapter), which shows that tardigrades are prone to reiterated LBA artefact.

Phylum	Species	ML	G+C	Amino	GC	Skew	Gene GC skew												Unsp
		dist.	%	%	Skew	index	C1	C2	A6	C3	N3	N5	N4	NL	N6	CB	N1	N2	bran
	Mean of Arthropoda	0.00	0.36	0.17	0.02		0.01	-0.1	-0.2	-0	-0.1	0.18	0.2	0.32	-0.2	-0.1	0.17	-0.2	
	Mean Lenght of gene						512	229	276	261	115	572	446	100	153	377	310	339	
Nematoda	*Xiphinema americanum*	0.77	**0.36**	**0.14**	**0.08**	**1.3**	0.11	0.14	-0	0.1	0.1	-0	0.13	0.23	-0	0.07	0.07	0.21	yes
	Thaumamermis_cosgrovei	0.85	0.32	0.12	**0.08**	**1.3**	0.1	0.1	0.1	0.2	0.1	0.0	0.0	-0.1	0.0	0.0	0.2	0.0	
	Trichinella spiralis	**0.74**	**0.36**	**0.14**	**0.05**	2.2	-0	-0.1	-0.4	-0.2	-0.2	0.42	0.39	0.71	-0.4	-0.1	-0.48	0.8	yes
	Anisakis_simplex	0.78	**0.34**	**0.14**	0.20	1.8	0.16	0.23	0.25	0.28	0.51	0.21	0.11	0.37	0.29	0.19	0.19	0.17	
	Ascaris suum	0.77	0.31	0.13	0.31	2.9	0.24	0.31	0.33	0.34	0.62	0.35	0.2	0.66	0.38	0.34	0.45	0.74	
	Ancylostoma duodenale	**0.76**	0.28	0.12	0.32	2.9	0.26	0.32	0.34	0.34	0.6	0.37	0.22	0.56	0.54	0.3	0.41	0.69	
	Caenorhabditis elegans	0.77	0.30	0.12	0.23	2.1	0.19	0.27	0.17	0.26	0.62	0.29	0.15	0.52	0.33	0.23	0.04	0.33	
	Cooperia oncophora	0.76	0.28	0.11	0.33	3.0	0.26	0.32	0.38	0.35	0.74	0.4	0.2	0.67	0.42	0.32	0.37	0.7	
	Necator americanus	0.77	0.28	0.12	0.34	3.1	0.25	0.31	0.35	0.37	0.59	0.42	0.24	0.61	0.48	0.34	0.45	0.78	
	Steinernema carpocapsae	**0.75**	0.30	0.11	0.20	1.8	0.18	0.22	0.27	0.23	0.47	0.22	0.09	0.52	0.05	0.19	0.15	0.2	yes
	Strongyloides stercoralis	0.76	0.28	0.11	0.21	2.0	0.21	0.27	0.3	0.23	0.36	0.17	0.14	0.58	0.09	0.22	0.3	0.41	
	Brugia malayi	0.90	0.30	0.13	0.32	2.9	0.18	0.3	0.3	0.37	0.52	0.37	0.36	0.59	0.68	0.27	0.27	0.72	
	Dirofilaria immitis	0.90	0.30	0.13	0.34	3.2	0.17	0.34	0.31	0.39	0.64	0.37	0.34	0.68	0.73	0.27	0.49	0.76	
	Onchocerca volvulus	0.91	0.31	0.13	0.35	3.2	0.15	0.34	0.34	0.44	0.56	0.42	0.41	0.6	0.58	0.25	0.5	0.77	
Priapulida	*Priapulus caudatus*	0.28	0.37	0.17	0.05	0.8	0.04	0.01	0.01	0.06	0.04	0.04	0.09	0.25	-0.2	-0.1	0.22	0.11	**no**
Onycophora	*Epiperipatus biolleyi*	0.33	0.31	0.14	0.15	1.7	0.16	0.23	0.22	0.19	0.24	0.22	0.23	0.37	-0.2	-0.2	-0.06	0.51	yes
Chordata	*Branchiostoma lanceolatum*	0.42	**0.41**	0.21	0.10	1.0	0.04	0.14	-0	0.1	0.2	0.2	0.17	0.08	-0.2	0.03	0.02	0.27	
	Branchiostoma floridae	0.42	0.42	0.21	0.10	1.0	0.04	0.14	-0	0.1	0.2	0.2	0.17	0.08	-0.2	0.02	0	0.26	
	Branchiostoma belcheri	0.41	**0.41**	0.21	0.12	1.2	0.05	0.17	0.02	0.12	0.23	0.17	0.21	0.15	-0.2	0.04	0.05	0.45	
	Epigonichthys lucayanus	**0.41**	0.42	0.21	**-0.02**	**1.0**	0.01	0.04	-0.2	0.04	-0.1	0.13	-0	-0.1	-0.2	-0.1	-0.36	-0.1	**no**
Chaetognath	*Spadella cephaloptera*	0.51	0.39	0.18	0.03	**1.1**	0.07	-0		0.08	-0.2	-0	0.14	0	-0	0.09	0.17	0	yes
Annelida	*Urechis caupo*	0.38	0.43	0.20	-0.18	1.8	-0	-0.1	-0.3	-0	-0.1	-0.2	-0.3	-0.2	-0.4	-0.1	-0.37	-0.3	yes
	Lumbricus terrestris	**0.37**	0.41	0.19	-0.14	1.4	-0	-0.1	-0.3	0	-0.1	-0.2	-0.2	-0.1	-0.4	-0.1	-0.25	-0.2	
	Clymenella torquata	**0.38**	**0.39**	0.19	-0.14	1.6	-0	-0.1	-0.2	0.06	-0.3	-0.2	-0.2	-0.2	-0.3	-0.1	-0.29	-0.2	yes
	Platynereis dumerilii	0.40	0.40	0.19	**-0.08**	**1.5**	0	0.04	-0	0.09	-0.1	-0.1	-0.2	-0.1	-0.2	-0	-0.3	-0.2	
Branchiopod	*Laqueus rubellus*	0.47	0.43	0.21	0.19	1.8	0.06	0.16	0.21	0.23	0.18	0.25	0.08	0.3	0.37	0.12	0.27	0.54	
	Terebratalia transversa	0.48	**0.42**	**0.20**	0.24	2.1	0.1	0.1	0.21	0.19	0.37	0.34	0.29	0.47	0.47	0.09	0.36	0.6	
	Terebratulina retusa	**0.42**	0.45	**0.20**	-0.20	2.2	-0.1	-0.1	-0.3	-0	-0.2	-0.3	-0.2	-0.1	-0.2	-0.2	-0.54	-0.4	yes
Mollusca	*Crassostrea gigas*	0.70	0.41	0.19	0.15	1.4	0.06	0.18	0.12	0.19	0.15	0.13	0.21	0.3	0.09	0.12	0.15	0.38	
	Crassostrea virginica	0.70	0.42	0.19	0.13	1.3	0.05	0.16	0.04	0.05	0.26	0.19	0.15	0.08	0.07	0.13	0.1	0.35	
	Mytilus edulis	0.59	0.41	0.19	0.23	1.8	0.2	0.18	0.25	0.21	0.4	0.18	0.27	0.31	0.47	0.1	0.15	0.38	
	Mytilus galloprovincialis	0.59	0.41	0.19	0.23	1.8	0.2	0.17	0.25	0.18	0.42	0.18	0.27	0.33	0.48	0.1	0.15	0.36	
	Placopecten magellanicus	0.69	0.44	0.23	0.31	2.6	0.13	0.27	0.38	0.32	0.32	0.35	0.35	0.5	0.38	0.24	0.47	0.5	
	Todarodes pacificus	**0.37**	**0.34**	**0.15**	0.13	1.3	-0	-0.1	-0.2	0.02	-0.1	0.32	0.4	0.57	0.51	0.22	0.41	-0.3	yes
	Loligo bleekeri	**0.37**	**0.34**	**0.16**	0.14	1.4	-0	-0.1	-0.2	0	-0.1	0.35	0.45	0.61	0.55	0.24	0.35	-0.3	
	Albinaria coerulea	0.52	**0.35**	**0.17**	**0.06**	**1.2**	0.1	0.06	-0.1	0.08	0	0.07	0.04	0.17	0.16	-0	0.03	0.15	**no**
	Aplysia californica	0.47	0.40	0.19	0.09	1.3	0.11	0.05	0.01	0.12	0.03	0.14	0.07	-0	0.19	0.01	0.08	0.23	
	Biomphalaria glabrata	0.45	0.33	**0.15**	0.09	**1.2**	0.11	0.05	-0.1	0.06	0	0.11	0.11	0.14	0.26	0.04	0.01	0.28	
	Haliotis rubra	**0.35**	0.42	0.20	**0.07**	**0.8**	-0	-0.1	-0.2	0.01	-0.1	0.26	0.3	0.44	0.33	0.05	0.32	-0.3	yes
	Pupa strigosa	0.49	0.42	0.20	**0.08**	1.3	0.07	0.02	-0.1	0.14	0.13	0.02	0.13	0.12	0.24	-0	0	0.25	
	Roboastra europaea	0.46	0.40	0.19	0.10	1.3	0.07	0.16	-0	0.09	-0.1	0.16	0.09	0.11	0.38	0.03	-0.02	0.3	
	Siphonodentalium lobatum	0.50	0.38	**0.17**	0.10	**1.0**	0.12	-0.1	-0.1	0.02	-0.1	0.35	0.23	-0.2	-0.3	-0.1	0.16	0.37	yes
	Katharina tunicata	**0.38**	**0.37**	**0.17**	**0.08**	2.6	0.18	0.16	0.22	0.21	0.34	-0.1	-0.1	-0	-0	-0	-0.37	0.51	

Table 3.1. “decision maker table” used to select an optimal set of arthropod outgroups. In phyla with more than one species I highlighted best values of distance or compositional character with a border and problematic values in red. From left to right various evolutionary characters: (1) classification; (2) the ML corrected distance from *Limulus polyphemus*; (3) the percentage of G+C nucleotides counted at the 1st + 2nd position of the whole supergene used in phylogenetic reconstruction; (4) the percentage of amino acids that are coded by G/C rich codons; (5) GC skew values calculated on the 1st +2nd codon position of the whole supergene; (6) skew index as measure of how much the overall strand asymmetry differs from that of arthropods ;(7) GC skew values calculated for each gene at 1st + 2nd position. Genes with a negative GC skew (high in G and low in C) are highlighted in grey in order to make a comparison with the average skew values of arthropods easier (first row in the table); (8) selected outgroups and their ability to avoid unspecific branching.

3.4 The effect of outgroup selection

3.4.1 Selection of optimal outgroups

The metrics discussed in the previous paragraphs, and summarised in table 3.1, were used to choose a set of 14 appropriate outgroups representative of the available phyla. Table 3.1 has been used as a decision maker and the criteria I have used in order of given precedence are: (1) low substitution rate, (2) ingroup-like G+C%, (3) low relative skew index, (4) phylogenetic proximity to arthropods and (5) the ability of the outgroup to avoid a "random branching effect". The latter character has been based on preliminary tree searches, using one outgroup at time. I have noted that all Lophotrochozoan outgroups I have tested, apart from the mollusc *Albinaria cerulea*, resulted in trees with a diphyletic Pancrustacea, generally attracting the problematic taxa *Speleonectes*, *Pollicipes* and *Gomphiocephalus*. Additionally all the longer branched ecdysozoans outgroups, (but not *Priapulus*) have shown unspecific branching, in some cases with the problematic taxa mentioned above, and in other cases with fast evolving species such as ticks. Interestingly, some outgroups attracted ingroup species with similar composition to the base of the tree: G+C rich species such as annelids or the mollusc *Haliotis* branch with G+C rich crustaceans while the A+T rich onychophoran and nematodes respectively attracted more A+T rich insects or chelicerates.

These results suggest that genetically and compositionally distant outgroups have incorporated misleading signal in their sequences and they may be considered inappropriate outgroups and a likely source of systematic error. The ability of an outgroup to root a tree without attracting long branch ingroups (as seen using nematodes) or compositionally biased species (as with lophotrochozoans) is a minimal condition for the adequacy of the outgroup itself. The mollusc *Albinaria*, the cephalochordate and the priapulid seem least affected by this problem of nonspecific rooting.

I used the set of 14 selected outgroups to build different datasets. I decided to use four sets of selected outgroups as follow: a group of 6 Lophotrochozoans, a group of 3 nematodes, a group of 4 ecdysozoans and a fourth group that consist on the 3 species that were not prone to unspecific branching with some of the ingroups. This last group contains the priapulid *Priapulus caudatus* that, according to my criteria, has been shown to be the best possible outgroup, the mollusc *Albinaria coerulea*, that, among all other Lophotrochozoans, is characterised by the most arthropod-like compositional characters and the cephalochordate.

3.4.2 Different outgroups promote different tree topologies.

The four sets of outgroups were used to root my well-balanced 18 arthropod dataset, chosen to represent major arthropod clades equally and to limit the effects of over-sampled groups (see materials and methods for details). For each dataset I used three approaches for inferring the phylogeny: Bayesian tree searches both at nucleotide (GTR model) and amino acid level (MtZoa model) and Likelihood bootstrapping at nucleotide level only (GTR model, chapter 7.3, page 129 for more details). Results are summarised in figure 3.4, where I show, for each of the datasets, the most resolved tree among the three I inferred with support values at nodes of interest. However the majority of other nodes were supported with values close to 1.00/100% for posterior probailities and bootstrap supports respectively.

Lophotrochozoans used in this study root the arthropod tree at the base of the Pancrustacea, making Myriapods and Chelicerates a monophyletic clade: the Myriochelata (figure 3.4 A). The signal is strong and independent of the method and data used. This result is in accordance with all the previous published mtDNA analyses that, until now, used only Lophotrochozoan species as outgroup sequences (Nardi et al 2003, Pisani et al 2004 among others). These previous studies generally included *Katharina tunicata* that has a very derived strand asymmetry (see table 3.1 and figure 3.2 A). I have repeated tree searches using only annelids or only molluscs and got similar results (trees not shown).

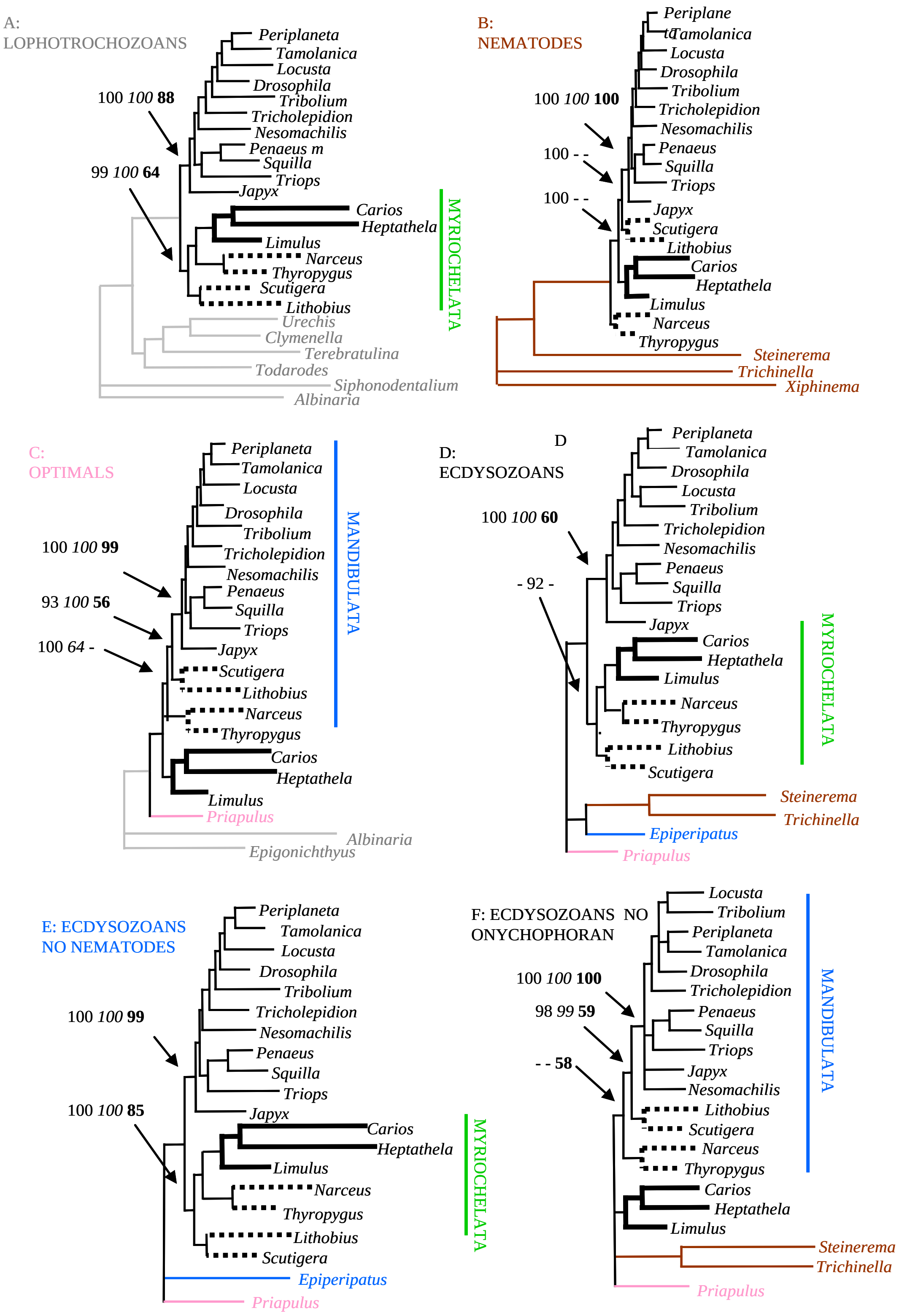


Figure 3.4. Different outgroups give different tree topologies. Values at nodes of interest correspond to Bayesian posterior probabilities for nucleotides (plain text), amino acids (italic) and bootstrap probabilities from Maximum likelihood analysis of nucleotides (bold) A: distant lophotrochozoans supports the Myriochelata hypothesis, in accordance with previous mtDNA phylogenies. B: The use of more phylogenetically related, but genetically distant Nematoda, gave contrasting phylogenetic signal, with some evidence for paraphyletic myriapods. C: my set of optimal outgroups support Mandibulata. D: Ecdysozoan outgroups give weak support for Myriochelata, while exclusion of nematodes does not affect the tree topology (E), and exclusion of the onychophoran recovered Mandibulata (F).

The use of nematodes gives contrasting results depending on character state (see posterior probabilities, PP for nucleotides versus amino acids) suggesting that the phylogenetic signal from these species is weak. In general the use of nematodes tends to make myriapods paraphyletic (figure 3.4 B), placing Diplopoda (millipedes) at the base of the arthropod tree and Chilopoda (centipedes) at the base of Pancrustacea. However no support is given to a monophyletic group of myriapods and chelicerates, as deduced from inspection of the partition probabilities calculated form the trees sampled during MCMC.

I have also used a set of 4 ecdysozoan species containing the priapulid, the onychophoran and 2 slowly evolving nematodes. Results are mixed: while in the Bayesian nucleotide tree, the outgroups do not form a monophyletic group (nematodes are placed within paraphyletic myriapods while onychophoran and priapulid are at the base of Pancrustacea), the corresponding amino acid tree supports Myriochelata (figure 3.4 D). The ML nucleotide tree is unresolved. I have repeated the analyses in the absence of the nematode sequences and I recover strong support for Myriochelata (figure 3.4 E), whereas exclusion of the onychophoran leads to support for Mandibulata (figure 3.4 F).

3.4.3 The use of optimal outgroups supports Mandibulata.

I have used as an outgroup clade the set of 3 species that, in accordance with my multi-criterion approach, were considered optimal. Use of these three taxa recovered monophyletic Mandibulata (Figure 3.4 C). While *Albinaria* and *Priapulus* have moderate and “arthropod like” composition patterns (see G+C content and GC skew), it is arguable that the GC rich genome of the cephalochordate may be a misleading factor. However, if this was the case I should expect this outgroup to branch specifically with G+C rich crustacean species, as in the case of the mollusc *Haliotis*, which has a similar G+C content, but this does not happen. In any case, exclusion of the cephalochordate still results in monophyletic Mandibulata (tree not shown). Another possible argument against the reliability of this outgroup set is that *Albinaria* is not characterised by a very reduced substitution rate (table 3.1). For this reason I have repeated the analyses in the presence of the slower evolving mollusc *Todarodes*, in place of *Albinaria* and results show no substantial differences.

In order to validate results obtained using a compositionally balanced, but numerically restricted dataset of 18 arthropods species, I repeated the analysis with a larger dataset of 41 arthropods, more representative of the main arthropods lineages, in particular of chelicerates and crustaceans. I have repeated Bayesian tree searches using the 4 different set of outgroups: with 4 lophotrochozoans, 3 nematodes, 5 ecdysozoans and the 3 best putative outgroups. Results show that both nematodes and lophotrochozoans are prone to long branch artifacts, as they tend to branch with the fast evolving maxillopod crustaceans (trees not shown). Interestingly, when using amino acids, lophotrochozoans do not show any support for the Myriochelata hypothesis, suggesting that the use of an enlarged sample may contribute to decrease the signal responsible for their supporting Myriochelata. The use of ecdysozoans including the onychophoran again gives support for Myriochelata, while the use of the 3 best selected outgroups (Priapulida, *Albinaria*, Cephalochordata) support Mandibulata at both nucleotide (72%) and amino acid level (95%).

3.5 Conclusions: the importance of outgroup selection and some support for Mandibulata

My results show that different outgroup taxa may root the ingroup tree (arthropods in the case I studied) in different positions. This may have critical consequences when the answer to a certain phylogenetic question (myriapods affinity in this case) relies on the position of a distant root and suggests a careful analysis of the rooting process. While the main criterion of outgroup adequacy is phylogenetic proximity, my results show that some of the closer related outgroups may have accelerated substitution rates (as in the case of nematodes) and/or extremely derived compositional characters (onychophoran). It may also be hazardous, in the outgroup selection process, to rely entirely on a reduced substitution rate. For example, according to their substitution rates, I should choose *Ancylostoma duodenale* and *Katharina tunicata* as one of the best nematode and mollusc outgroups respectively for rooting the arthropod tree (Table 3.1). However, inspection of their skew indices revealed that those species show a reverse strand bias (figure 3.2 A) for almost all the genes and alert us to possible systematic biases.

For this reason I advocate the use of a multi-criterion approach in order to compare diverse evolutionary characters of the outgroup sequences. A similar approach can be used, and indeed was used in this chapter, for the selection of ingroup sequences. For this purpose I have compiled a "decision maker" table with taxonomic, genetic and compositional characteristics. In particular I have introduced a new "relative skew index" of composition heterogeneity that is particularly effective in describing short genomes such as mtDNA and I suggest a possible use in enlarged phylogenomic analyses, from nuclear genes and especially chloroplasts.

In the presence of strong phylogenetic signal one shouldn't expect ingroup relationships to change significantly when different outgroups are used. On the other hand, if the phylogenetic signal is weak even minor non phylogenetic signal may annihilate the true phylogenetic signal. In these cases, it is possible to apply methods that help in discerning between phylogenetic and nonphylogenetic signal. Useful methods are removing fast evolving sites, functionally recoding sequences and improving the evolutionary model. I suggest that critical analyses of the outgroup characteristics and comparisons of the effect that a diverse set of outgroups may have on the ingroup phylogeny could be a good indicator of signal stability and may be adopted more widely as a congruence test for phylogeny. I indeed carried out a careful exploration of outgroup effects using a phylogenomic dataset in chapter 6.

Because the difference between the Mandibulata and the Myriochelata hypotheses relies on the position of the arthropod outgroup, the qualities of different outgroups and their effects on the internal arthropod phylogeny should be scrutinised carefully. I have considered a number of *a priori* criteria for choosing superior outgroups for rooting the arthropod phylogeny: phylogenetic proximity (arthropod sister groups or members of the Ecdysozoa being preferred) genetic proximity (short branch lengths preferred) and two measures of compositional similarity, GC content and GC skew (outgroups more similar to the arthropods being preferred *a priori*).

I have shown that different outgroup taxa give us different positions for the root within the Arthropoda with some outgroups (the most derived in my study) supporting

Myriochelata and others (closer related) supporting Mandibulata. This lack of consistency seems to suggest that this particular phylogenetic problem is one that is hard to resolve and that the significant internal node is likely to be short. In conclusion, the Myriochelata hypothesis, supported by all previously published mitochondrial phylogenies (which usually incorporate lophotrochozoan outgroups only), may depend on the presence of phylogenetically and/or compositionally distant outgroups and is likely to be an artefact due to a systematic bias rather than a true phylogenetic signal. This impression is reinforced by the analyses, in the next chapter, of a larger mitochondrial dataset, which shows that taxon sampling and character choice plays a crucial role for the affinity of myriapods using these sequences.

Chapter 4
Exploring subtle signal: a mitogenomic analysis of the Ecdysozoa

4.1 Abstract

Evolutionary relationships within the ecdysozoans are unresolved, impairing the correct interpretation of comparative genomic studies. In particular, the affiliation of the three Panarthropoda phyla (Arthropoda, Onychophora, and Tardigrada) and the position of Myriapoda within Arthropoda (Mandibulata vs Myriochelata hypothesis) are among the most contentious issues in animal phylogenetics. To elucidate these relationships, I have analyzed complete mitochondrial genome sequences of two Tardigrada, Hypsibius dujardini and Thulinia sp. (the first genomes to date for this phylum), one Priapulida, Halicryptus spinulosus, and two Onychophora, Peripatoides sp. and Epiperipatus biolleyi, and a partial mitochondrial genome sequence of the Onychophora Euperipatoides kanagrensis. Tardigrada mitochondrial genomes resemble those of the arthropods in term of the gene order and strand asymmetry, while Onychophora genomes are characterised by numerous gene order rearrangements and strand asymmetry variations. In addition, Onychophora genomes are extremely enriched in A and T nucleotides, while Priapulida and Tardigrada are more balanced. Phylogenetic analyses based on concatenated amino-acid coding sequences support a monophyletic origin of the Ecdysozoa and the position of Priapulida as the sister group of a monophyletic Panarthropoda (Tardigrada plus Onychophora plus Arthropoda). The position of Tardigrada is more problematic, most likely because of long branch attraction (LBA). However, experiments designed to reduce LBA suggest that the most likely placement of Tardigrada is as a sister group of Onychophora. The same analysis also recovers monophyly of traditionally recognized arthropod lineages such as Arachnida and Mandibulata, reconciling morphology and molecules.

4.2 Mitogenomic characters of the Ecdysozoa

According to the previous chapter (figures 3.2 and 3.3), some ecdysozoan mitochondrial genomes are characterized by a considerable heterogeneity of genomic characters. In this section I enlarge this compositional analysis to the newly sequenced genomes and carefully explore the relationships of these characters in these and other ecdysozoans lineages.

4.2.1 Gene order analyses

In collaboration with the Dennis Lavrov lab, I compared the mitochondrial gene orders in species sampled to those of other representatives from Priapulida and Onychophora (Webster et al. 2006, Podsiadlowski et al. 2008) and the putative *A*rthropod *A*ncestral *G*ene *O*rder (AAGO) identical to that in *Limulus polyphemus* (Lavrov et al. 2000).

The mitochondrial gene arrangement in the priapulid *Halicryptus spinulosus* is exactly the same as that in the other priapulid *Priapulus caudatus* (Webster et al. 2006). Both gene arrangements differ from the AAGO by a single inversion of the *rns-trnS1* cluster (Fig 4.1). The mitochondrial gene order of the tardigrade *Thulinia* sp. differs from the AAGO only in the position of *trnI*, which is located between *trnL1* and *trnL2* and has an opposite transcriptional polarity. *trnI* is found in the same location in the mitochondrial genome of *Hypsibius dujardini.* The latter genome displays several additional rearrangements not present in *Thulinia*. These autapomorphies include the interchange in the positions of the *trnT-nad6-cob-trnS2* and the *nad1-trnL2* clusters, and transpositions of *nad2* and two clusters of tRNAs (*trnW-trnC-trnY* and *trnK-trnD*). Furthermore, *trnR* is inverted in *Hypsibius* mtDNA. Finally, the gene order in the onychophoran *Peripatoides* sp. is identical to the AAGO with the exception of a single inversion of *trnQ*. This inversion is found in two out of the three onychophorans sampled so far and is a good candidate for a synapomorphy of the group. In contrast to *Peripatoides* mtDNA, mitochondrial genomes from the two other representatives of Onychophora (Podsiadlowski et al. 2008) display multiple additional gene rearrangements, which appear to be autapomorphic for each species. As a result, the three onychophoran mitochondrial genomes share very few gene boundaries, namely *atp6-atp8, nad1-trnL2* and *cob-trnS2*.

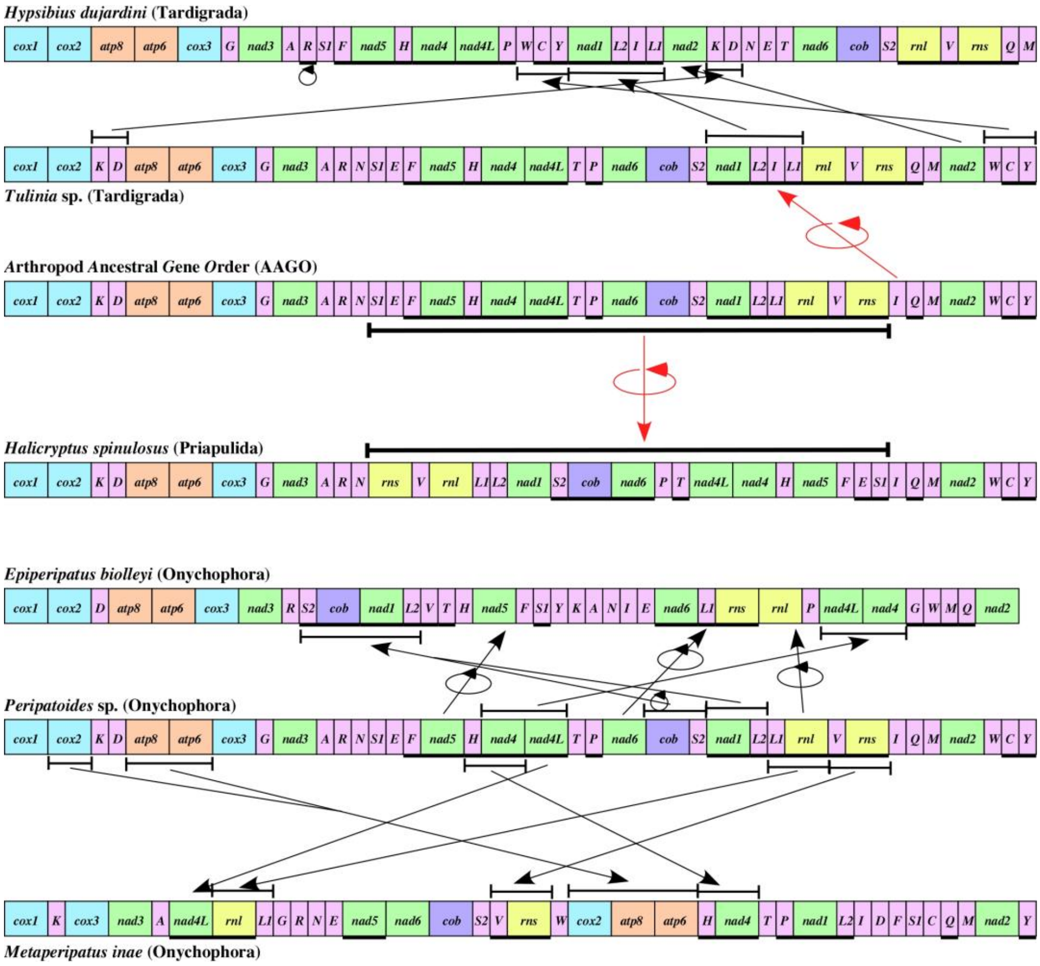

Hypsibius dujardini (Tardigrada)
cox1 cox2 atp8 atp6 cox3 G nad3 A R S1 F nad5 H nad4 nad4L P W C Y nad1 L2 I L1 nad2 K D N E T nad6 cob S2 rnl V rns Q M
cox1 cox2 K D atp8 atp6 cox3 G nad3 A R N S1 E F nad5 H nad4 nad4L T P nad6 cob S2 nad1 L2 I L1 rnl V rns Q M nad2 W C Y
Tulinia sp. (Tardigrada)
Arthropod Ancestral Gene Order (AAGO)
cox1 cox2 K D atp8 atp6 cox3 G nad3 A R N S1 E F nad5 H nad4 nad4L T P nad6 cob S2 nad1 L2 L1 rnl V rns I Q M nad2 W C Y
Halicryptus spinulosus (Priapulida)
cox1 cox2 K D atp8 atp6 cox3 G nad3 A R N rns V rnl L1 L2 nad1 S2 cob nad6 P T nad4L nad4 H nad5 F E S1 I Q M nad2 W C Y
Epiperipatus biolleyi (Onychophora)
cox1 cox2 D atp8 atp6 cox3 nad3 R S2 cob nad1 L2 V T H nad5 F S1 Y K A N I E nad6 L1 rns rnl P nad4L nad4 G W M Q nad2
Peripatoides sp. (Onychophora)
cox1 cox2 K D atp8 atp6 cox3 G nad3 A R N S1 E F nad5 H nad4 nad4L T P nad6 cob S2 nad1 L2 L1 rnl V rns I Q M nad2 W C Y
cox1 K cox3 nad3 A nad4L rnl L1 G R N E nad5 nad6 cob S2 V rns W cox2 atp8 atp6 H nad4 T P nad1 L2 I D F S1 C Q M nad2 Y
Metaperipatus inae (Onychophora)

Figure 4.1. Gene order in arthopods, tardigrades, onychophorans and priapulids. Mitochondrial gene order comparisons and proposed gene rearrangements for the onychophorans, the priapulid *Halicryptus spinulosus*, the tardigrades and the *A*rthropod *A*ncestral *G*ene *O*rder (AAGO). tRNAs are labeled by the one-letter code for their corresponding amino acids. Genes are transcribed from left to right unless underlined, which indicates an opposite transcriptional polarity. Black arrows indicate inferred gene rearrangements. Red arrows show inferred synapomorphies of the two phyla Priapulida and Tardigrada. Multiple tRNA gene rearrangements found between *Peripatoides* sp. and the two other onychophoran species have been omitted for clarity.

The lack of unambiguous shared derived gene rearrangements (synapomorphies) among arthropods, tardigrades, onychophorans, and priapulids proves that no resolution can be achieved for their interrelationships using the current mitochondrial gene order data. This conclusion rejects some previous claims based on mitochondrial gene order data of close relationships between arthropods and tardigrades (Ryu et al. 2007), as the ancestral arthropod gene arrangement has also been inferred as the putative Protostome Ancestral Gene Order (Lavrov and Lang, 2005).

4.2.2 High degree of compositional heterogeneity

Figure 4.2 shows, for each of the main group of Ecdysozoa and various outgroups, the average compositional properties of the coding sequences expressed in percentage of G+C (guanine + cytosine) plotted versus the percentage of amino acids whose codons are enriched in G and C. As expected, and in accordance with the similar plot of figure 3.1, the nucleotide and the amino acid contents are highly correlated (R^2=0.76).

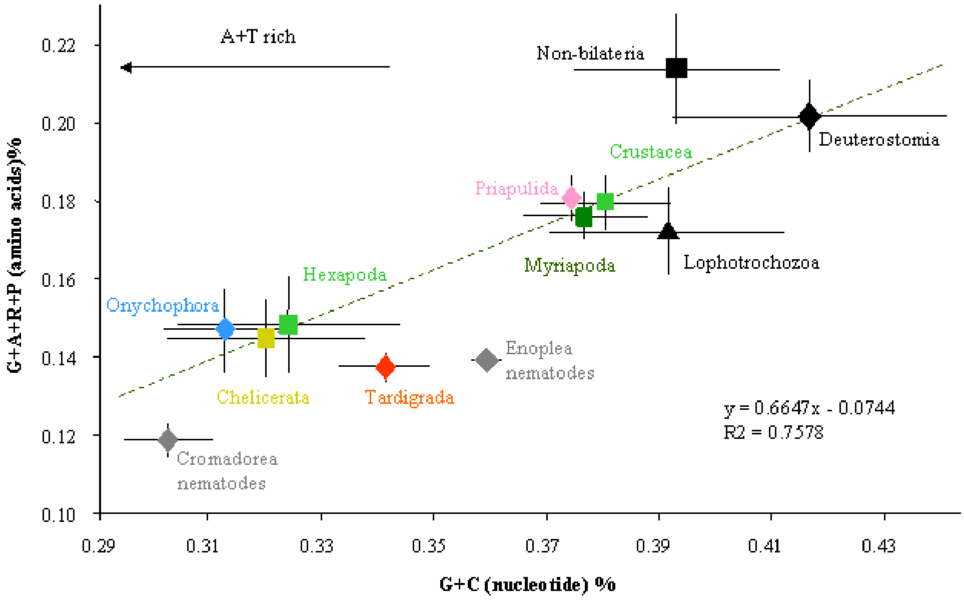


Figure 4.2. Compositional properties of ecdysozoan mitochondrial coding sequences. Nucleotide frequencies are plotted against amino acid frequencies. Values are averaged for some major groups, with standard deviations indicated. All Ecdysozoa are A+T rich if compared to outgroup sequences. Onychophora are extremely A+T rich, while Priapulida and Tardigrada are more balanced. Amino acid frequencies are more homogenous within groups than are the corresponding nucleotide frequencies . Compare this figure with figure 3.1 which uses the same colour code.

However, for onychophorans, tardigrades and hexapods, the amino acid composition is evidently less biased than the nucleotide one, suggesting that inference of phylogeny based on amino acids may be less prone to compositionally driven systematic errors.

Compared to the outgroups, all the ecdysozoans are characterised by coding sequences impoverished in G and C; however, the degree of heterogeneity among the main ecdysozoan groups is remarkable. Onychophora are extremely A+T rich, to a degree which is comparable only to those of the well known compositionally problematic ticks and nematodes (table 4.1). On the other hand, tardigrades and especially priapulids are characterised by a more balanced nucleotide composition. Notably, the four "subphyla" of the arthopods are rather heterogeneous, with hexapods and chelicerates being A+T rich and myriapods and crustaceans less biased. Interestingly the Cromadorea nematodes are extremely A+T rich, but the relatively slower evolving Enoplea (data not shown), are characterised by more balanced values of nucleotide (and amino acid) composition.

4.2.3 Strand asymmetrical properties

Genes on different strands, or in different positions of the same strand, may accumulate more mutations toward C than G. This property, known as strand asymmetry has been shown to vary extensively within the Ecdysozoa to the extent that phylogenetic reconstruction may be misled (Hassanin 2005, Hassanin et al. 2005, Jones et al. 2007, Rota Stabelli and Telford 2008, Masta et al. 2009).

I have explored strand asymmetry in the tardigrades, onychophorans, priapulids and arthropods by calculating the GC skew for each gene independently (using separately 1st + 2nd and 3rd codon position) and plotting corresponding values in the skew profiles of figure 4.3, using the ancestral arthropod gene order (AAGO) to order genes on the abscissa. For comparative reasons, I calculated the average for the arthopods, using for the calculation only species which share a similar skew profile, conserved for example from the basal *Limulus polyphemus* to the dipterans, suggesting that it may be intimately related to the AAGO. (see materials and methods for more details). Most of the coding genes in the "AAGO skew profile" are found on the C rich strand and are therefore characterised by a negative GC skew (green values in figure 4.3).

PHYLUM	SPECIES	ML distance	GC% 123	GC% 12	GC skew	abs skew index	aa GC rich	aa G/C skew
Echinodermata	Asterina pectinifera	0.49	0.40	0.41	-0.17	2.6	0.19	-0.23
	Paracentrotus lividus	0.48	0.40	0.43	-0.15	1.6	0.20	-0.21
	Florometra serratissima	0.46	0.27	0.35	0.13	1.6	0.17	-0.13
	Ophiura lutkeni	0.52	0.35	0.38	-0.10	1.2	0.17	-0.19
Hemichordata	Balanoglossus carnosus	0.49	0.50	0.48	-0.30	2.8	0.23	-0.29
Mollusca	Haliotis	0.41	0.42	0.42	0.07	2.7	0.20	-0.04
	Aplysia californica	0.50	0.33	0.40	0.10	0.8	0.19	-0.09
	Biomphalaria glabrata	0.48	0.26	0.33	0.10	1.0	0.15	-0.13
Annelida	Clymenella torquata	0.44	0.34	0.39	-0.22	1.9	0.19	-0.30
	Lumbricus terrestris	0.44	0.39	0.41	-0.22	1.7	0.19	-0.25
Priapulida	Priapulus caudatus	0.36	0.32	0.37	0.00	0.9	0.17	-0.14
	Halicryptus spinulosus	0.32	0.32	0.38	0.19	2.5	0.19	-0.09
Tardigrada	Hypsibius dujardini	0.49	0.33	0.36	-0.07	1.9	0.14	-0.17
	Tulinia	0.47	0.30	0.33	-0.10	2.4	0.13	-0.24
Onychophora	Metaperipatus inae	0.38	0.23	0.28	0.09	1.1	0.14	-0.20
	Peripatoides sp	0.36	0.20	0.32	0.09	1.1	0.13	-0.23
	Epiperipatus biolley	0.37	0.27	0.32	0.21	3.1	0.14	-0.10
	Euperipatoides kanagrensis	0.25	0.25	0.34	0.10	1.4	0.18	-0.14
Myriapoda	Narceus annularus	0.39	0.38	0.41	-0.06	3.8	0.19	-0.14
	Thyropygus sp.1	0.35	0.34	0.37	-0.09	2.8	0.18	-0.20
	Antrokoreana	0.40	0.32	0.36	0.01	0.7	0.19	-0.11
	Scutigerella causeyae	0.36	0.39	0.41	-0.05	2.8	0.16	-0.13
	Bothropolys sp-2004	0.37	0.31	0.37	-0.06	2.8	0.17	-0.16
	Lithobius forficatus	0.38	0.34	0.38	-0.03	2.4	0.18	-0.14
	Scutigera coleoptrata	0.34	0.29	0.35	0.00	2.9	0.16	-0.15
Chelicerata	Limulus polyphemus	0.34	0.34	0.37	-0.09	3.8	0.17	-0.15
	Nymphon gracilis PYC	0.44	0.23	0.30	0.08	2.3	0.14	-0.15
	Achelia	0.39	0.24	0.29	0.06	1.0	0.13	-0.17
	Mastigoproctus	0.40	0.31	0.35	-0.04	2.8	0.17	-0.15
	Nothopuga	0.34	0.32	0.35	0.01	3.4	0.16	-0.13
	Heptathela hangzhouensis	0.39	0.29	0.34	-0.02	2.4	0.15	-0.18
	Nephila clavata Araneae	0.44	0.25	0.31	0.09	2.4	0.14	-0.11
	Habronattus oregonensis	0.46	0.26	0.32	0.07	2.9	0.15	-0.10
	Ornithodoros moubata	0.39	0.22	0.32	-0.01	3.8	0.14	-0.16
	Ixodes holocyclus	0.38	0.21	0.29	0.02	2.6	0.13	-0.16
	Carios capensis	0.38	0.24	0.31	-0.05	3.6	0.14	-0.19
	Ornithoctonus huwena	0.47	0.22	0.34	0.08	3.3	0.15	-0.06
	Aphonopelma	0.50	0.37	0.38	0.12	2.5	0.17	0.04
	Damon	0.38	0.38	0.40	0.01	2.0	0.19	-0.07
	Eremobates	0.33	0.31	0.36	0.00	3.2	0.18	-0.10
	Hypochiuls	0.50	0.31	0.33	0.06	2.4	0.15	-0.07
	Phalangium	0.35	0.30	0.35	-0.01	3.0	0.18	-0.13
	Phrynus	0.37	0.34	0.37	-0.01	3.4	0.18	-0.12
	Pseudocellus	0.42	0.32	0.34	-0.03	2.9	0.15	-0.13
Crustacea	Lepeophtheirus salmonis	0.55	0.36	0.39	0.08	1.8	0.18	-0.08
	Penaeus monodon	0.33	0.31	0.38	-0.03	1.2	0.18	-0.16
	Daphnia pulex	0.38	0.40	0.41	-0.02	1.0	0.19	-0.14
	Artemia franciscana	0.43	0.36	0.39	-0.05	0.6	0.17	-0.14
	Geothelphusa dehaani	0.34	0.28	0.35	-0.02	2.6	0.16	-0.18
	Cherax destructor	0.36	0.40	0.40	-0.05	2.3	0.19	-0.13
	Portunus trituberculatus	0.34	0.31	0.38	-0.03	2.1	0.17	-0.14
	Squilla empusa	0.35	0.33	0.38	0.01	1.3	0.19	-0.13
	Lysiosquillina maculata	0.34	0.38	0.40	-0.01	1.4	0.19	-0.11
	Triops cancriformis	0.35	0.32	0.37	-0.03	1.3	0.17	-0.20
Insecta	Anopheles gambiae	0.31	0.24	0.32	0.04	1.3	0.16	-0.21
	Drosophila melanogaster	0.31	0.23	0.31	0.05	1.3	0.15	-0.22
	Locusta migratoria	0.33	0.26	0.33	-0.04	1.6	0.15	-0.18
	Nesomachilis australica	0.34	0.33	0.38	-0.02	2.5	0.17	-0.15
	Ostrinia nubilalis	0.32	0.21	0.28	0.04	1.8	0.14	-0.22
	Periplaneta fuliginosa	0.31	0.25	0.33	0.03	1.4	0.16	-0.17
	Petrobius brevistylis	0.34	0.30	0.38	-0.02	1.9	0.18	-0.17
	Pyrocoelia rufa	0.34	0.15	0.30	-0.03	1.9	0.14	-0.20
	Sclerophasma paresisensis	0.32	0.24	0.33	0.00	1.2	0.15	-0.18
	Thermobia domestica	0.34	0.32	0.39	-0.09	2.4	0.18	-0.20
	Tribolium castaneum	0.32	0.31	0.34	-0.07	2.7	0.15	-0.19
	Tricholepidion gertschi	0.34	0.32	0.37	-0.03	2.1	0.17	-0.17
	media arthropoda*	0.41	0.29	0.34	0.02	2.1	0.16	-0.17

Table 4.1: Compositional statistics of the 66 taxa used in the phylogentic analyses of this chapter. From left to right: (1) the classification, (2) species name, (3) the ML distance to the arthropods, (4) the nucleotide content in term og G+C% calculated on all codon positions and (5) on first and second positions only. (6) Strand asymmetry calculated on first plus second position and (7) the absolute skew index. In (8) is the percentage of amino acid whose codons are enriched in G and C (G,A,R,P) and in (9) the skew between aminoacid, whose codons are G rich and those C rich.

Different genes are differently affected by strand asymmetry: for example the conserved genes of complex IV (Cox1, Cox2 and Cox3, Nardi et al. 2003) are slightly affected by strand bias while the faster evolving genes of complex I (Nadh subunits) are clearly positively or negatively skewed (green bars in figure 4.3 A, B and C). The 3rd codon position is less constrained then the 1st and 2nd positions and accumulates directional mutations more quickly; it is more likely to be at equilibrium and therefore a better estimator of the strand asymmetrical tendency of genes.

The GC skew values calculated both at 1st plus 2nd and at 3rd codon positions are extremely similar in the two tardigrades and resemble the AAGO strand profile (red and orange bars in figure 4.3 A and B). However, while the gene order in the tardigrade *Thulinia sp*. is nearly identical to that of the arthropods, in *H. dujardini* I observe at least six rearrangements including various coding genes. I therefore expect the strand profile to be slightly different in the two tardigrades yet GC skew values calculated at all codon positions are extremely similar in the two tardigrades and resemble the arthropod AAGO profile (red and orange bars in figure 4.3 A and B), suggesting a similar replicatory system and/or the same direction of replication in the two phyla. Intriguingly, the high number of rearrangements observed in *H. dujardini* does not seem to affect the strand asymmetry, perhaps suggesting recent rearrangement events in *H. dujardini*. On the other hand, the strand profile in the onychophorans (especially in *Peripatoides* sp.) differs significantly from that of the arthropods, in particular at the 3rd codon position (figure 4.3B), with the *trnI*-*trnN* fragment (highlighted by an arrow) showing a complete reversal of strand asymmetry compared to arthopods. This inversion of strand profile also characterises the priapulid *Halicriptus spinulosus* (figure 4.3 E and F).

According to figure 4.1 the genomes of the two priapulids posses the same gene order as the arthropods except for a single inversion of the *trnI*-*trnN* cluster. As GC skew is principally related to gene orientation, we should expect the inverted fragment in the priapulids to possess different values of GC skew compared to the arthropods. Accordingly, genes in the *trnI*-*trnN* cluster (arrowed genes ND2-ND3) have different values compared to the arthopods (Fig 4.3 E).

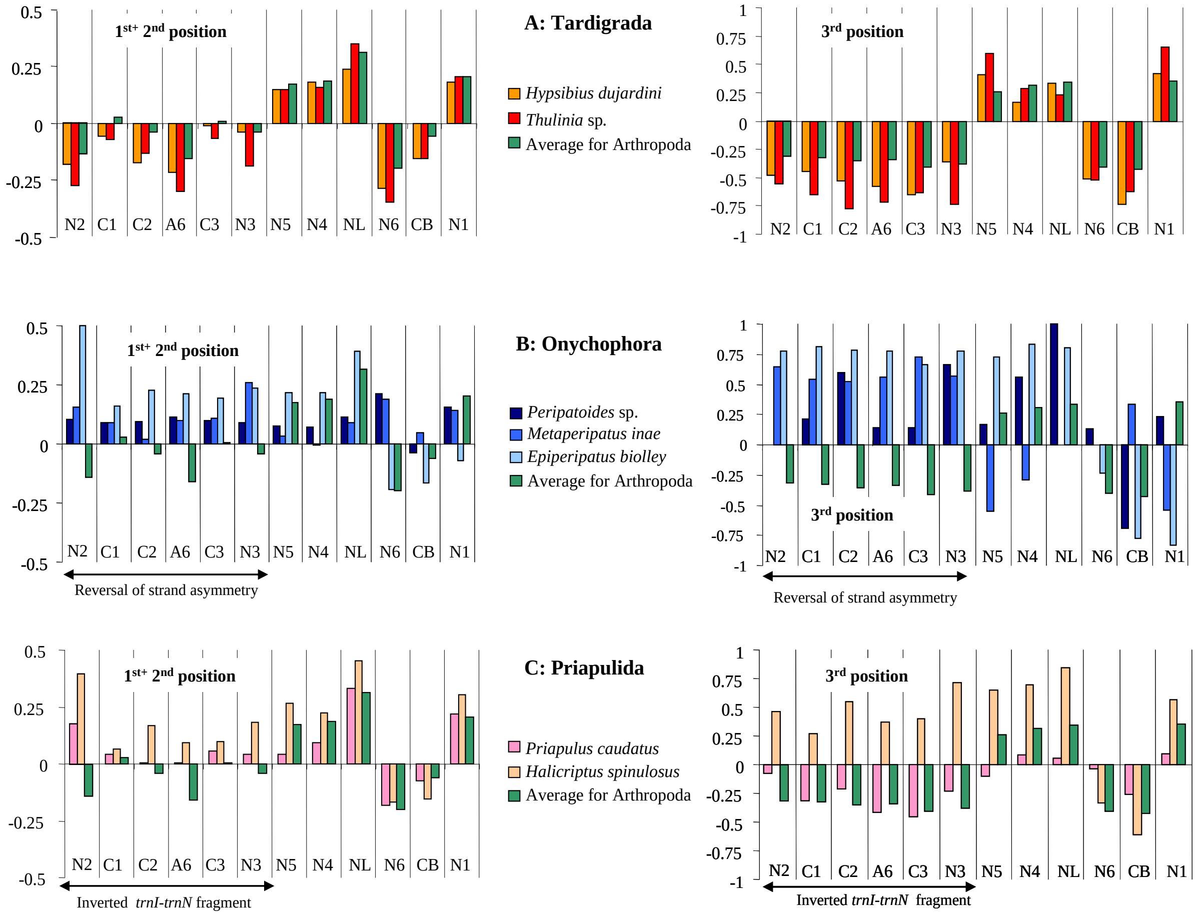


Figure 4.3. Strand asymmetry in priapulids, tardigrades and onychophorans. GC Sskew calculated on 1st plus 2nd (on the left) and 3rd (on the right) codon positions for the 13 coding genes of Tardigrada (A), Onychophora (B) and Priapulida (C) lineages. Genes are ordered as in the Ancestral Arthropod Gene Order (AAGO) and for each plot the average calculated over all arthropods with AAGO and a similar strand profile is given. A: Tardigrada and Arthropoda share a similar strand profile, suggesting a conserved replicatory system. B: Onychophora display a reversal of strand asymmetry for genes in the fragment N2-N1 (arrow underlined) compared to the arthropods. C: The same fragment has complete reversal of strand asymmetry in the priapulid *Halicryptus*, but not in *Priapulus*, although the two priapulids share an identical gene order.

In *Halicryptus* all the genes in the *trnI-trnN* fragment show a clear positive GC skew, (a complete reversal of the arthropods values which are negative), but this tendency is less evident in the other priapulid *Priapulus caudatus*, whose GC skew has intermediate values between arthropods and the *Halicryptus*. This has been interpreted (Webster et al. 2006) as a recent inversion of the *trnI-trnN* cluster, which hadn't left enough time to allow mutational pressure to invert completely strand asymmetry at positons 1st and 2nd, as happened in the other priapulid *Halicryptus*. If this is true we should expect the 3rd codon position, which evolves faster than the two first codon positions, to have a similar direction of asymmetrical mutation in the two priapulids. Unexpectedly, it is extremely different in the two priapulids (figure 4.3 F): in *H. spinulosus* it mirrors the skew at 1st and 2nd position, suggesting that the mutational pressure in its genome is at equilibrium, but in *Priapulus* resembles those of the arthropods and not those of the other priapulid (at least for the genes in the shared inverted *trnI-trnN* cluster. This can not be explained by two distinct convergent inversions in the two priapulids, otherwise the skew at the 3rd codon position in Priapulus should mirror (and exacerbate) that at 1st and 2nd positions. One explanation for this discrepancy is that there has been a recent inversion of the control region in *P. caudatus*, and that skew values have not yet reached equilibrium in their new mutational pressure regime.

4.3 Phylogenetic analyses

Using the newly sequenced genomes, the partial sequences of *E. kanagrensis* plus other sequences present in NCBI datasets, I assembled a large dataset of the 13 protein coding sequences, which resulted in a final dataset of 66 taxa and 2307 amino acid residues. I carried out phylogenetic analyses using a variety of methods and models, in particular those developed in chapter 2. I selected species to obtain a balanced representation of the main arthropod lineages as well as to choose deuterostome and lophotrochozoan outgroups characterised by moderate rates of evolution and with compositional characters that resemble those in the ecdysozoans (in accordance with a decision table similar to that of pag. 64 in the previous chapter).

The 66 taxon dataset does not contain nematodes, although they are of key importance for resolving the affinities of the ecdysozoans, in particular of the tardigrades which have been linked to nematodes by phylogenomics studies (Lartillot and Philippe 2008, Dunn et al 2008). Mitochondrial sequences of nematodes are very fast evolving and previous attempts to use them in phylogenetic reconstruction had led to dubious assemblages of nematodes and other fast evolving lineages (Mwinyi et al 2009, Podsiadlowski, Braband and Mayer 2008). I tested the feasibility of using nematode sequences by assembling preliminary datasets and sampling in particular slow evolving enoplean nematodes. Results using two different datasets show that nematodes tend to branch with fast evolving lophotrochozoan outgroups (trees not shown). In order to avoid the misleading effect of the fast evolving nematode lineage I have excluded them from further analyses.

4.3.1 An unlikely chelicerate affinity of the tardigrades

Figure 4.4 shows the Bayesian and Maximum lihlehood (ML) analyses of 1st and 2nd codon positions using the GTR model. The tree supports monophyly of Ecdysozoa, with priapulids basal to a group of Onychophora plus arthropods, although the latter clade is very weakly supported (Posterior Probability (PP) 0.55). Arthropods are paraphyletic in this tree, as tardigrades are grouped with fast evolving pycnogonids and symphylans. Bootstrap support (BS, support in bold in figure 4.4) from the ML analyses are low, suggesting that, given the nature of the bootstrap test, the phylogenetic signal is weak in this dataset. Furthermore, inspection of branch lengths indicates that tardigrades, pycnogonids and symphylans are all fast evolving lineages, suggesting that their grouping may be the result of a Long Branch Attraction artefact.

Table 1 and figure 4.3 show that mitochondrial genomes of ecdysozoans are characterised by different patterns of strand asymmetry. For this reason, I have further analysed the nucleotide dataset using the NTE recoding strategy which has been proven to lessen strand bias artefacts (Hassanin et al 2005, Jones et al. 2007). The NTE tree, however, is extremely similar to that using unrecoded positions, grouping tardigrades and pycnogonids with the symphylan (see PPs underlined in figure 4.4). Morover, tardigrades are characterised by a “typical” strand asymmetrical pattern (figure 4.3 A) which is shared by the majority of the arthropods.

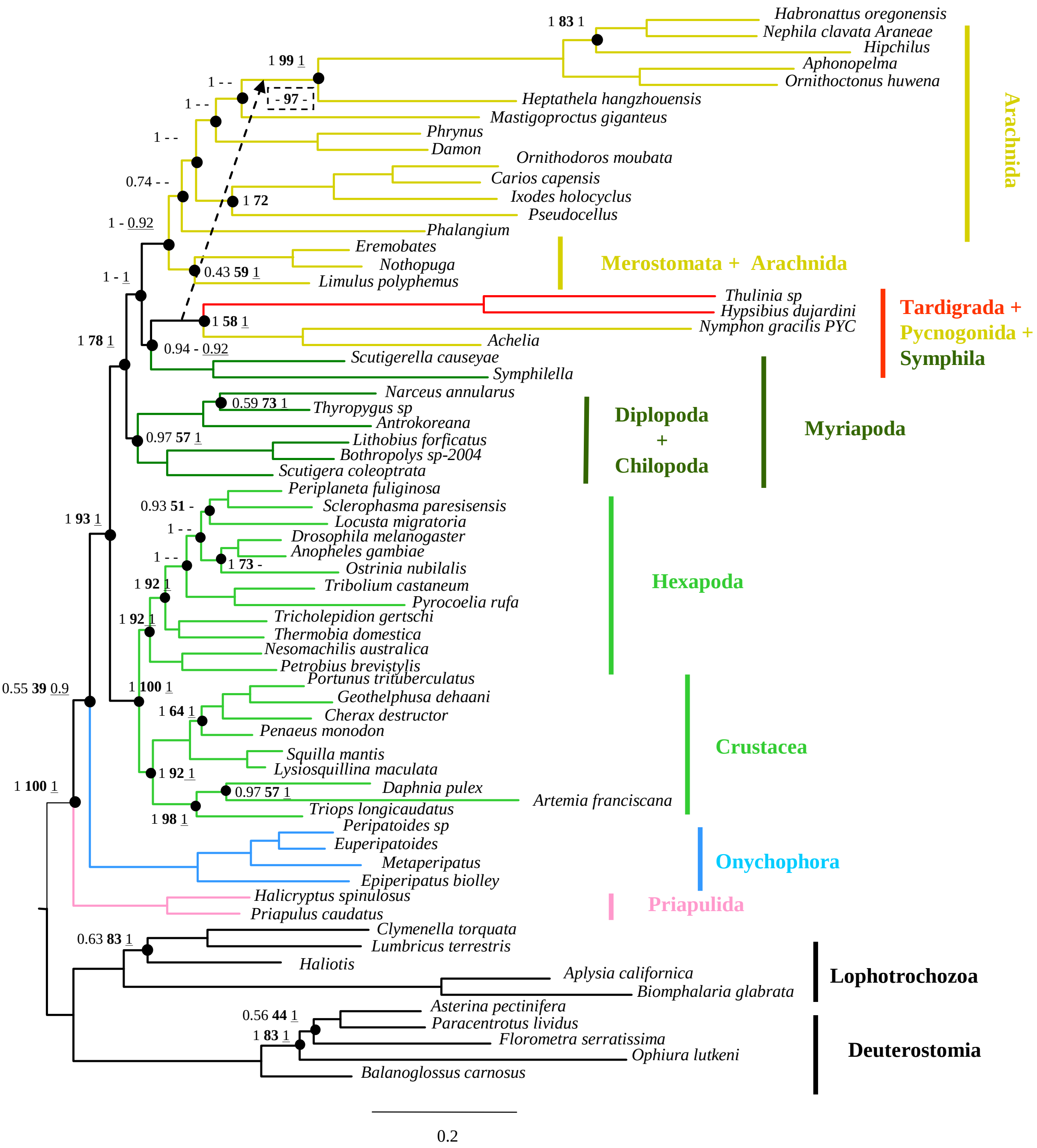


Figure 4.4. Bayesian and Maximum likelihood analyses using nucleotide sequences. Consensus tree from the Bayesian analysis using the GTR model. Support at nodes are from left to right the posterior probability (PP) from the GTR Bayesian analysis, the bootstrap supports (BS) from the Maximum likelihood analysis using GTR and the PP from the Bayesian analysis using the NTE model. Tardigrades are consistently recovered as closer related to chelicerates and myriapods. An alternative position for the tardigrades using Maximum likelihood is shown by the dotted arrow.

According to figure 4.2, the amino acid content is markedly more homogeneous among different lineages than nucleotide content, suggesting that structural constraints acting at the protein level may reduce the effects of mutational pressure acting at the nucleotide level. Amino acid sequences seem better markers for inference of phylogeny, as homogeneity of the stationary frequency is an assumption of the majority of evolutionary models. According to a crossvalidation test of model fit to the dataset, I chose my homogeneous MtZoa and the heterogeneous CAT models as the best fitting models and used them for the majority of my amino acid analyses (more details in chapter 7.4). In figure 4.5A I show a schematic representation of the consensus tree from the Bayesian and Maximum likelihood analyses using the MtZoa model. The tree partially resembles the nucleotide tree of figure 4, with the exception that the group of tardigrades plus pycnogonids is nested within paraphyletic arachnids and not as sister of the symphylan myriapods. I have also analysed the amino acid dataset using other models of evolution based on empirical replacement matrices (MtArt, MtREV and MtHydro) as well as after exclusions of amino acids (Proline and Glycine) which are mostly affected by strand asymmetrical bias. Corresponding trees resulted in extremely similar topology to the MtZoa model tree, thus supporting a chelicerate affinity for the tardigrade (trees not shown).

4.3.2 The LBA nature of the tardigrades-chelicerates group and support for Panarthropoda using the CAT model.

The grouping of tardigrades and pycnogonids/symphylan does not find any support from morphological evidence and challenges two commonly accepted notions, monophyly of chelicerates (supported for example by the presence of chelicerae) and that of the arthropods (which posses articulated appendages). A possible LBA artifact is suggested by the extreme branch length and the consequent accelerated rate of evolution in tardigrades, pycnogonids and symphylans. Furtermore, the NTE model and the exclusion of asymmetrical biased amino acids fail to recover a different topology than the unrecoded nucleotide models, suggesting that a pycnogonid/symphylan affinity of the tardigrades is not due to strand asymmetry problems, but rather to a more general LBA artefact.

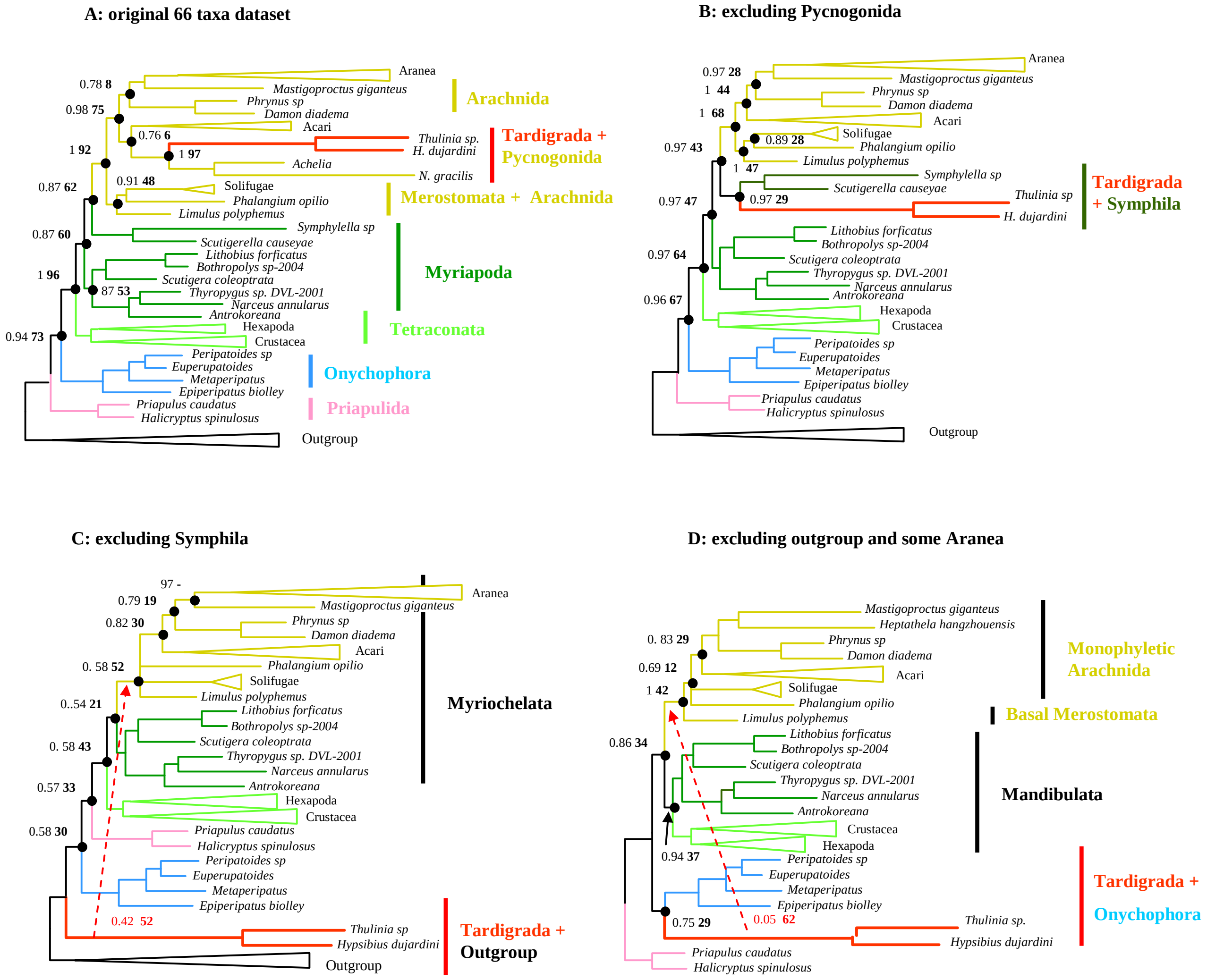


Figure 4.5. Unstable position of Tardigrada using the MtZoa model. Consensus trees from the Bayesian analysis of the amino acid dataset. Values at nodes are the PP from the Bayesian analysis and the BS from the Maximum likelihood analysis (in bold). Fast evolving lineages have been sequentially removed from the original dataset (tree A) by removing Pycnogonida (B), Symphyla (C) and outgroups plus some fast evolving chelicerates (D). I show a schematic version of the Bayesian trees with some lineages collapsed for clarity. The position of Tardigrada changes as the taxon sampling is reduced, suggesting a reiterated LBA artefact. When all fast evolving lineages are excluded and only close, slow evolving Priapulida are used as outgroups (tree D), support for a group of Tardigrada plus Onychophora is recovered.

In order to test the possible effect of systematic errors such as LBA, I sequentially removed from the 66 taxa dataset the fast evolving pycnogonids, symphylans and all the outgroups plus some fast evolving chelicerates (which are characterised by accelerated rate of evolution and/or inversion of strand asymmetry, see table 4.1). Using the homogeneous MtZoa model, the position of tardigrades varies extensively, from sister to the pycnogonids (figure 4.5 A, using the full 66 taxa dataset) to sister of symphylan myriapods (figure 4.5 B) to a basal ecdysozoan position (sister of the outgroups in figure 4.5 C). Eventually, when all fast evolving lineages are removed and only slow evolving/closely related priapulids are used as an outgroup, tardigrades are weakly recovered as basal panarthropods, sister to the onychophorans (figure 4.5 D). Notably, as fast evolving lineages are removed, support for a grouping of myriapods plus chelicerates decays and in the last dataset (characterised by more homogeneity of among lineages rate of evolution) myriapods are grouped with the Tetraconata, according with Mandibulata hypothesis. Similar analyses, using MtREV, MtArt and GTR, gave extremely similar results (trees not shown).

The grouping of tardigrades and onychophorans is consistently recovered using the CAT model, regardless of presence of fast evolving lineages (figure 4.6). The CAT heterogeneous model has been shown to overcome the effects of LBA (Bourlat et al. 2009, Lartillot et al. 2007, Lartillot and Philippe 2008). Accordingly, an onychophoran affinity of the tardigrades, as suggested by the CAT model, should be regarded as a more likely topology than that obtained using homogenous models and the full set of taxa. I have also analysed my dataset using the CAT-BP model, which models heterogeneity among different branches. Surprisingly, two independent runs supported a sister group relationship between tardigrades and pycnogonids, but analysis in the absence of the pycnogonids recovers a sister relationship between tardigrades and onychophorans as basal panarthropods with modest support (trees not shown)

4.3.3 More support for Panarthropoda using site stripping

Fast evolving sites are more likely to be saturated because of successive multiple substitutions and for this reason they are considered a possible source of misleading signal. Furthermore, Castoe and colleagues (2009) have recently shown that slow evolving sites are the more likely to bring signal from adaptive evolution in unrelated lineages. Accordingly, I have explored the distribution of signal in the alignment by separating fast and slow evolving sites from the moderately evolving ones, which are believed to be the most reliable source of correct signal.

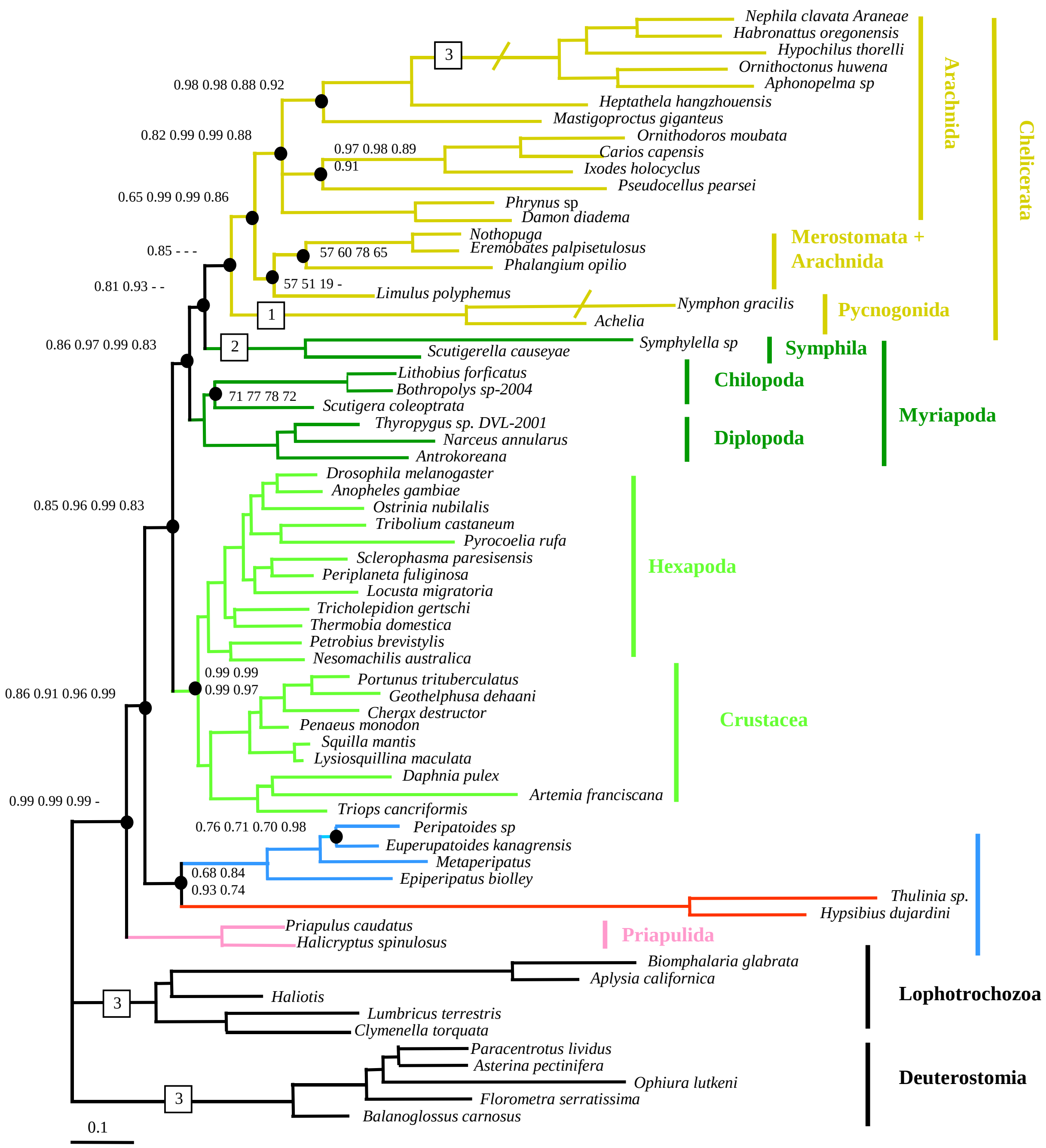


Figure 4.6. Consistent support for Tardigrada plus Onychophora following sequential taxa removal using CAT model. Consensus tree from the Bayesian analysis of the amino acid dataset using the heterogeneous CAT model. Fast evolving lineages have been sequentially removed from the original dataset as in figure 5. The four resulting analyses resulted in similar topologies and consistently supported a group of Tardigrada plus Onychophora. Values at nodes are posterior probabilities using (from left to right) the original 66 taxa dataset and the sequential removal of Pycnogonida (branch square labelled 1), Symphyla (labelled 2) and all the outgroups plus fast evolving Chelicerata (labelled 3). Where not indicated pp are 1.

The CAT tree using fast and slow evolving sites, which fall in the external quartiles of a distribution of classes of rates (figure 4.7A), supports a tardigrades affinity for the chelicerates, in contrast with the same analyses using all classes of sites (figure 4.6). The same tree supports also a group of onychophorans plus Aranea as well as paraphyletic Tetraconata, hypotheses which are clearly dubious and never observed in other analyses, suggesting that the signal associated with fast/slow evolving sites carries a high amount of non-phylogenetic signal. On the other hand, the tree from the analysis of medium evolving sites (figure 4.7 B) support monophyly of all the commonly accepted groups such as tetraconates, chelicerates, and arachnids, suggesting that signal in these sites is more geniune. Interesting, this set of sites weakly supports a sister relationship between Tardigrada and Onychophora; the resolution at this node is probably reduced because of lack of some sites which carry phylogenetic signal, but fall in the external quartiles. Notably, use of medium evolving sites supports Mandibulata, as discussed below.

Overall, conditions which reduce the effects of LBA – use of closely related outgroups (figure 4.5D), more effective CAT model of evolution (figure 4.6) and exclusion of sites which are possible source of errors (figure 4.7B) result in support for a close relationship between onychophorans tardigrades in a monophyletic (pan)arthropod clade.

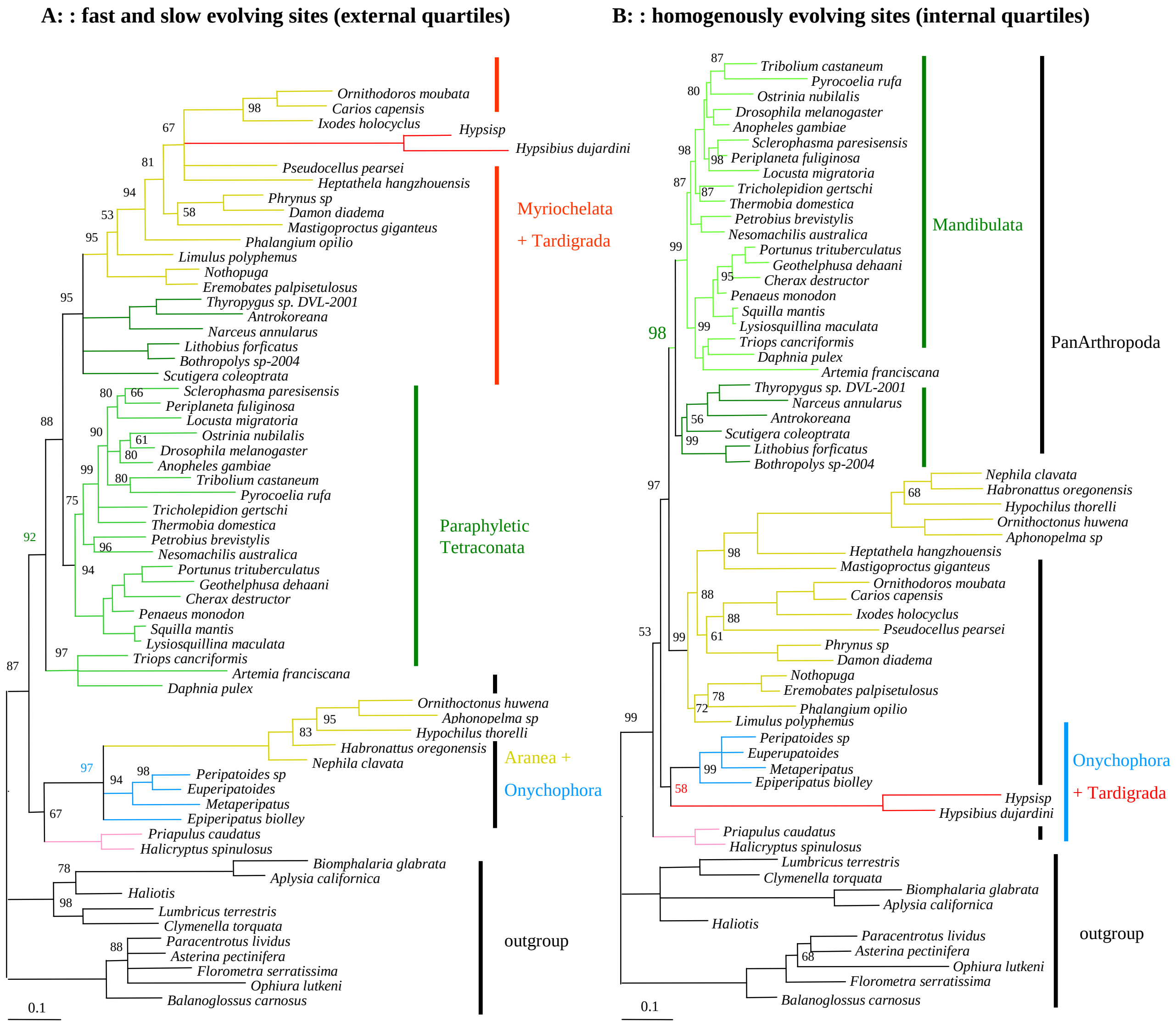
4.3.4 An arthropod affinities for the onychophorans

Ecdysozoa are strongly recovered as monophyletic and in the vast majority of the analyses priapulids are sister to a group of onychophorans plus arthropods. This topology reflects the accepted view of a common origin of the limbs of onychophorans and arthropods and the basal position of the Cycloneuralia, represented in my sample by the priapulids (Telford et al 2008, Dunn et al. 2008). The only exception is the tree of figure 4.5 C, which supports the priapulids as sister to the euarthropods, a topology which can be interpreted as an artifact due to the mutual attraction of tardigrades and outgroups, which may have pulled the onychophorans, which show a tardigrade affinity in other analyses (figure 4.5 D and 4.6). This view is reinforced by the analysis of the same dataset in the absence of the tardigrades which recovers onychophorans as sister group of the arthopods (tree not shown).

Compared to the work of Podsialowski and colleagues (2008) which used only one onychophoran and did not recover a monophyletic origin of panarthropods, the addition of new sequences from *Peripatoides* and *Euperipatoides* species, increased the phylogenetic signal and the resolution of the onychophoran lineage. For the internal onychophoran relationships, the Peripatopsidae (austral onychophorans) are clearly monophyletic with the Australian species (*Euperipatoides kanagrensis*) closer related to the Guinea species (*Peripatoides sp*.) than to the New Zealand species (*Metaperipatus inae*), reflecting the geographical distribution of the three islands and/or the breakup of the ancient Australasian continent.

4.3.5 Some evidence in support for the Mandibulata

Most of my nucleotide and amino acid analyses (figure 4.4, 4.5 A, 4.5 B and 4.6) support Myriochelata (myriapods plus chelicerates) in accordance with the majority of molecular studies, but in disagreement with the parsimonious distribution of morphological characters (Hwang et al. 2001, Pisani et al. 2004, Rota-Stabelli and Telford 2008, Telford et al. 2008). Interestingly, when fast evolving species are excluded from the dataset, support for Myriochelata decreases using the CAT model (PP 0.97 to 0.82 in figure 4.6) and support for Mandibulata is recovered using the MtZoa model (PP 96 in figure 4.5 D). Finally, when only moderately evolving sites are analysed, the CAT model supports Mandibulata (PP 98 in figure 4.7 B), while the remaining fast and slow evolving sites support Myriochelata (PP 95 in figure 4.7 A). The impression is that the signal supporting Myriochelata is found in fast evolving sites or is associated with datasets containing fast evolving species, in particular symphylan myriapods, which tend to group with chelicerates, making Myriapoda paraphyletic in most of my analyses (for example in figure 4.4 and 4.5). My interpretation is that when sources of systematic errors are reduced (excluding fast evolving sites and/or fast evolving lineages) datasets tend to support Mandibulata.



A: : fast and slow evolving sites (external quartiles)

B: : homogenously evolving sites (internal quartiles)

Figure 4.7. Signal decomposition supports Mandibulata and a basal Tardigrade position. Consensus tree from the CAT Bayesian analysis of (A) all the slow/fast evolving sites (corresponding to 1st and 4th quartiles) and (B) the moderately evolving sites (corresponding to 2nd and 3rd quartiles). Note that the 1st and 4th quartile tree supports Myriochelata, paraphyletic Tetraconata and a group of Tardigrada plus Chelicerata (as do most of my homogenous model analyses), while internal quartiles support Tardigrada basal to monophyletic Arthropoda, with weak support for a group of Tardigrada plus Onychophora.

4.3.6 Relationships of other arthropod groups

Hexapods and crustaceans are consistently grouped in my analyses in accordance with the Tetraconata hypothesis. Crustaceans (although I sampled only Malacostraca and Branchiopoda) are strongly supported as a monophyletic group in disagreement with the common accepted notion of paraphyletic crustaceans as supported for example by phylogenomic studies (Lartillot and Philippe 2007, Dunn et al. 2008, Roeding et al. unpublished, but see also chapter 6).

Finally, in some of my phylogenies (figure 4.4 and 4.5A), chelicerates are paraphyletic due to the tardigrades being sister of the pycnogonids an affinity which I have previously interpreted as a long branch attraction artifact. When pycnogonids are excluded from the analysis (figure 4.5 B, C and D), tardigrades branch in other parts of the tree leaving chelicerates monophyletic. On the other hand, using the CAT model, chelicerates is recovered as a monophyletic group with the Pycnogonida being sister to remaining (eu)chelicerates (figure 4.6) in accordance with a unique origin of the chelicerae in Chelicerata. Unexpectedly, in most of the analyses, the horshoe crab (*Limulus polyphemus*) is grouped with the harvestman *P. opilio* and/or the Solifugae, making the arachnids paraphyletic. However, when long branch species are excluded from the alignments, both CAT (figure 4.6) and MtZoa (figure 4.5 C and D) models recover *Limulus* as basal Chelicerata, while Opilionidae and Solifugae join Acari, in a monophyletic Arachnida.

4.4 Conclusions

Phylogenetic signal in ecdysozoan mitochondrial sequences is anticipated to be subtle and complicated by heterogeneity of among lineage rates of evolution and nucleotide composition (figure 4.2 and 4.3). This high level of heterogeneity is reflected by low bootstrap support and conflicting phylogenetic reconstructions using homogeneous models of evolution. It is clear that the amount of phylogenetic information in mitochondrial sequences at some nodes is fairly small. A possible explanation is that some of the lineages of interest, in particular tardigrades, are fast evolving thus promoting possible artifacts such as LBA, which my analyses suggest is responsible for grouping tardigrades with fast evolving arthropods (figure 4.4 and 4.5A). This problem is probably exacerbated by the mitochondrial datasets being relatively short, thus statistically carrying too little true phylogenetic signal.

I have, however, shown that experiments designed to reduce the effect of LBA – use of the heterogeneous CAT model (figure 4.6), exclusion of fast evolving lineages (figure

4.5 D) and exclusion of fast and/or slow evolving sites (figure 4.7B) – tend to recover a group of tardigrades plus onychophorans sister to (eu)arthopods in agreement with morphological predictions of a common origin of the panarthropods. Similar experiments also recover monophyly of other morphologically recognised groups such as Mandibulata and Arachnida.

Chapter 5
The longer the dataset, the more consistent the phylogeny: a need for a phylogenomic approach to study the arthropods.

5.1 Abstract

While the monophyletic origin of the four main groups of arthropods – chelicerates, myriapods, crustaceans and insects – is well established, the relative position of myriapods and chelicerates is ambiguous to the extent that different molecular sources have produced conflicting results. Conflict is even present between different phylogenies using the same type of marker. One possible explanation for these contradictory results is that the conflict stems from systematic and/or stochastic errors in phylogenetic reconstruction. To test this idea, in this short chapter I reanalyse five different phylogenetic datasets and show that short datasets are self inconsistent when at least one analytical parameter is allowed to vary and yield trees which are in conflict with each other. Conversely, longer datasets, in particular the phylogenomic one, are more congruent over parameter variations suggesting less susceptibility to stochastic problems. Results advocate the use of larger datasets, in particular phylogenomic ones, for addressing the affinity of myriapods.

5.2 Reanalysis of five arthropod phylogenetic datasets.

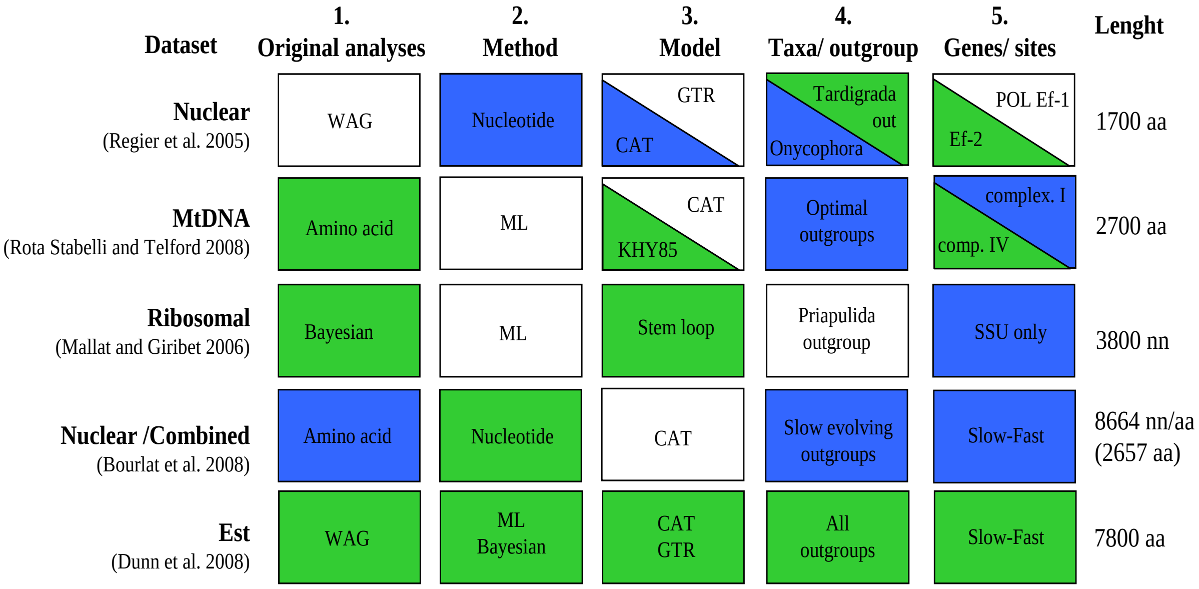
As addressed in the introduction, various molecular markers (Figure 1.3 A and 1.3 C), have linked myriapods with chelicerates in a group called Myriochelata, despite the strong morphological resemblance between myriapods, crustaceans and insects. I have shown in chapter 3, however, that support for Myriochelata from the analyses of mitochondrial sequences may be related to the use of fast evolving or distant outgroups. In chapter 4, I also show that the position of myriapods changes when fast evolving lineages and sites are removed. One strong possibility, in particular considering the contradictory results of some published molecular analyses (Regier et al. 2008, Rota-Stabelli and Telford 2008, Pisani 2004) and the results from the two previous chapters, is that the conflict stems from systematic and/or stochastic errors in molecular phylogenetic reconstruction (Philippe et al. 2005). This is probably exacerbated by the subtlety of the signal describing the affinity of myriapods as predicted by the short length of branches leading either to Myriochelata or Mandibulata (compare trees of figure 1.3 and figure 4.5 for example).

To test the possibility that stochastic and/or systematic errors are affecting myriapod affinity, I reanalysed the mitogenomic dataset I have used in chapter 3 plus the four published datasets of figure 1.3, pag 22 . More in detail, these five datasets comprise ribosomal (Mallat and Giribet 2006), nuclear (Regier et al 2005), mitochondrial (chapter 3 and Rota Stabelli and Telford 2008), combined (Bourlat et al. 2008) and phylogenomic (Dunn et al. 2008) markers and are effective representatives of currently available datasets for the study of arthropod relationships.

For each of the five datasets I allowed various phylogenetic and analytical parameters to vary (see figure 5.1). I inferred phylogeny using (1) different methods of phylogenetic inference and (2) evolutionary models than the ones used in the original analysis (for example using bootstrapped maximum likelihoods instead of Bayesian inference, stem-loop model for ribosomal subunits and CAT model instead of homogenous models). I also analysed different sub-datasets by (3) using different taxonomic sampling (for example using different sets of outgroups or excluding fast evolving or over-sampled lineages) and (4) varying the nature of the dataset (for example using nucleotides instead of corresponding amino acids, using a subset of the original dataset or exploring the effects of removing fast evolving sites). More details are given in the material and methods. As I analysed the variation of four classes of parameters in five different datasets, I generated more than one hundred different trees, which were far too many to be included in this thesis: for example, the analyses describing the effect of outgroup choice in the mitochondrial dataset (corresponding in figure 5.1 to one single box) took the whole chapter 3 to be satisfactorily addressed. I consequently opted for a brief summary of the results as in figure 5.1, specifying which key parameters (if any) resulted in a topology inconsistent with the original or basic analysis of the dataset (first box column of figure 5.1).

5.2 Short datasets are inconsistent over parameter variation

I found that all datasets, except for the phylogenomic one, are self inconsistent when at least one of the parameter is allowed to vary, yielding trees which are in conflict with each other.



Dataset	1. Original analyses	2. Method	3. Model	4. Taxa/ outgroup	5. Genes/ sites	Lenght
Nuclear (Regier et al. 2005)	⬜ WAG	🟦 Nucleotide	⬜ GTR 🟦 CAT	🟩 Tardigrada out 🟦 Onycophora	⬜ POL Ef-1 🟩 Ef-2	1700 aa
MtDNA (Rota Stabelli and Telford 2008)	🟩 Amino acid	⬜ ML	⬜ CAT 🟩 KHY85	🟦 Optimal outgroups	🟦 complex. I 🟩 comp. IV	2700 aa
Ribosomal (Mallat and Giribet 2006)	🟩 Bayesian	⬜ ML	🟩 Stem loop	⬜ Priapulida outgroup	🟦 SSU only	3800 nn
Nuclear /Combined (Bourlat et al. 2008)	🟦 Amino acid	🟩 Nucleotide	⬜ CAT	🟦 Slow evolving outgroups	🟦 Slow-Fast	8664 nn/aa (2657 aa)
Est (Dunn et al. 2008)	🟩 WAG	🟩 ML Bayesian	🟩 CAT GTR	🟩 All outgroups	🟩 Slow-Fast	7800 aa

Figure 5.1. Instability of phylogenetic signal using short datasets. Five phylogenetic datasets (indicated on the left) have been analysed using a variety of analytical settings. Green boxes correspond to support for the Myriochelata, blue for the Mandibulata and white for another tree topology or a lack of resolution. Inside boxes there is a brief description of (some of) the parameters modified. The first column of boxes (1) indicates the support from the original analysis and/or publication. Following columns indicate the results from one or more re-analyses using (2) different methods of inference, (3) different models of evolution, (4) alternative taxon sampling, in particular for the outgroups and (5) different character choice and slow-fast analysis.

The ribosomal dataset of Mallat and Giribet (2006) for example supports Myriochelata (green boxe in third row) when the whole dataset is analysed, but recovers Mandibulata when only the small ribosomal subunit (SSU, blue box) is analysed; the use of Scalidophora as outgroups recovers myriapods as basal arthropods (white box). The latter topology is the one favoured by the nuclear dataset of Regier et al (2005), which, however turns to support Mandibulata when the dataset is analysed at the nucleotide level or Myriochelata when only tardigrades are used as outgroup sequences. The parameter dependency is even more evident using the same mitochondrial dataset I used in chapter 3: while the full dataset support Myriochelata, use of optimal outgroups (see chapter 3 for a detailed analysis of outgroup choice using this dataset) supports Mandibulata, a result which is, however, model dependent: using CAT model myriapods are recovered as paraphyletic supporting neither Mandibulata or Myriochelata (white triangle of column 2). The larger combined dataset of Bourlat et al. (2008) is more consistent in supporting Mandibulata (blue boxes), although occasional support for Myriochelata is recovered when the dataset is analysed at nucleotide level (green box). Finally the phylogenomic dataset of Dunn et al. (2008) which contains 7800 characters for their sampled myriapod, consistently support Myriochelata (green boxes of last row).

I tested each of the datasets for four competing hypothesis using the SH and the AU test and testing Myriochelata, Mandibulata and, as a matter of internal comparison, two strongly unlikely topologies: the Atelocerata hypothesis (myriapods plus hexapods) and a topology as supported in Regier et al. (2005), which groups chelicerates and Tetraconata and which I have called “crazypoda”.

Test	Hypothesis	Nuclear	mtDNA	Ribosomial	Combined	EST
AU test	Mandibulata	0.843	0.229	0.596	0.977	0.048
	Myriochelata	0.016	0.811	0.304	0.027	0.949
	Crazypoda	0.192	0.016	0.512	0.001	0.002
	Atelocerata	0	0	0.0003	0.0002	0
SH test	Mandibulata	0.954	0.498	0.842	1	0.247
	Myriochelata	0.449	0.941	0.614	0.138	0.989
	Crazypoda	0.552	0.193	0.777	0.042	0.132
	Atelocerata	0	0	0	0.001	0

Table 5.1: AU and SH tests fail unambiguously to reject competing hypotheses in the five analysed datasets, in particualr in short nuclear, mitochondrial and ribosomal datasets. Rejected hypotheses are underlined in red.

According to the SH test, none of the datasets is able to reject at a significant level either of the major competing hypotheses (Mandibulata or Myriochelata), while other hypotheses such as Atelocerata or Crazypoda are rejected. Notably, short datasets, such as the ribosomal and the nuclear ones (compare their alignment lengths in figure 5.1), tend to be unable to reject the majority of competing hypotheses, while longer datasets when tested using AU test, in particular the combined dataset of Bourlat et al. (2008) reject them more easily (underlined red values).

5.3 Conclusions: a need for phylogenomic approach

These results, taken together, suggest that short datasets do not carry enough phylogenetic signal for resolving the affinity of myriapods. A likely explanation is that these datasets are too short and do not contain enough phylogenetic information to support one topology unambiguously over the other. On the other hand, the larger Bourlat et al. (2008) combined dataset is clearly more consistent over parameter variation, with a competing tree topology recovered only when the dataset is analysed at the nucleotide level. Finally, the larger phylogenomic dataset of Dunn et al. (2008) consistently recovers Myriochelata, suggesting that the longer the dataset, the more stable is the phylogenetic reconstruction. The stability of the signal is not a sufficient condition to assess the quality of a dataset and the veracity of supported topology. However, the indication is that the level of uncertainty is reduced in long datasets, possibly because they tend to overcome stochastic type errors. For this reason the affinity of myriapods should *a priori* be tested with a dataset, such as the phylogenomic one, which can rely on many positions and a higher likelihood of containing enough phylogenetic information. It is clear that while phylogenomic datasets may reduce the stochastic errors seen in short datasets, systematic errors remain (Lartillot and Philippe 2008) and should be taken in to account, as I do in the next chapter.

Chapter 6

A phylogenomic survey into (pan)arthropod relationships

6.1 Abstract

Although myriapods strongly resemble hexapods and crustaceans (Mandibulata hypothesis), the majority of molecular studies support a group of myriapods plus chelicerates, a clade which has been named Myriochelata or Paradoxopoda. Some molecular phylogenetic analyses also link tardigrades to nematodes, rather than to arthropods and onychophorans (Lartillot and Philippe 2008). Furthermore, for the Tetraconata, there is no consensus between various molecular markers regarding how crustacean classes are related to each other or how the insects fit within the Tetraconata assemblage. In the case of both myriapod and tardigrade affinities, molecules are in conflict with the parsimonious interpretation of ostensibly homologous morphological characters, suggesting that phylogenetic analyses of these lineages may be prone to systematic and/or stochastic errors. In the previous chapter I showed that short datasets are susceptible to parameters variations, most likely as a result of stochastic errors due to too few positions in the datasets. These datasets, therefore, may not be indicated to study the affinities within the arthropods (and ecdysozoan) main lineages. Following these indications, I assembled three distinct phylogenomic datasets of up to 201 genes and 59 taxa, centred respectively on basal arthropods, crustaceans and hexapods. Notably, I included new data from the myriapod Strigamia maritima. Analyses of these datasets gave support for (i) Mandibulata, suggesting that support for Myriochelata is due to the effects of systematic errors and reconciling the molecules with the known distribution of morphological characters, (ii) a paraphyletic origin of Cycloneuralia, (iii) gave some evidence for a group of tardigrades plus onychophorans in accordance with monophyletic Panarthropoda, (iv) support a monophyletic group of hemimetabolan (Exopterigota) insects and (v) gave some evidence for a closer affinity to the branchiopods than the malacostracans for the hexapods.

6.2 Support for Mandibulata and some evidence for Panarthropoda.

6.2.1 A phylogenomic dataset of 198 genes and 59 taxa centred on basal arthropods.

With the aim of elucidating myriapod relationships, a phylogenomic dataset of 59 taxa has been assembled, including novel data from a member of the previously poorly-represented myriapods, the geophilomorph centipede *Strigamia maritima*. *Strigamia*, diverged from the scutigeromorph *Scutigera* (the other myriapod for which ESTs were available at August 2009) more than 418 million years ago according to the presence of stem-group scutigeromorph fossils in the latest Silurian (Edgecombe & Giribet 2007). The addition of a second, phylogentically-distant myriapod is expected to increase the phylogenetic signal and to reduce the effects of autapomorphies resulting from the use of only one myriapod.

ESTs of *S. maritima* have been sequenced by Macrogen using a cDNA library provided by Michael Akam and Ariel Chipman (Chipman, Arthur and Akam 2004). I carefully screened the library, prior to EST sequencing, in order to check its quality: the library turned out to contain long fragments and not to be significantly redundant (see chapter 7.6 for more details). Contig assembly and ortholog selection have been carried out by my collaborators Hervè Philippe and Henner Brinkmann, following my suggestions for taxon sampling. We followed a procedure previously described in the literature (see chapter 7.6) and assembled a concatenated alignment of 40,100 reliably aligned amino acid positions derived from 198 genes. In order to reduce the effects of missing data, we only included genes sampled in at least two-thirds of the species. For the same reason 26 out of the 59 taxa were composed of chimeric sequences produced by merging two or more species belonging to non-controversial clades (see table for details). In most cases we merged species of the same genus and in a few cases of the same (super)family or (infra)order. Only in the case of Onychophora we merged two distantly related species as a consequence of the Peripatidae species being poorly sampled.

The large number of positions of this dataset is expected to reduce the possibility of stochastic error, while the dense taxonomic sampling, in particular the addition of an extra myriapod, is expected to reduce the effect of systematic errors. In an effort to reconcile molecular and morphological estimates of panarthropod phylogeny, I analysed in detail this large alignment as well as the gene set used by Dunn and colleagues (Dunn et al. 2008). However, in order to make computationally demanding bootstrapped analyses feasible, the majority of the analyses have been carried out on a reduced, although taxonomically balanced, 30 taxa dataset.

6.2.2 Support for Mandibulata and Panarthropoda using CAT

In Figure 6.1 I show the result of a Bayesian analysis of my 30 taxa data set using the CAT model and a non-parametric bootstrap approach. My analyses support the monophyly of Mandibulata and of Panarthropoda with a posterior probability (PP) of 1.0 and a high bootstrap support (BS: Mandibulata 79%; Panarthropoda 100%), two results that are in agreement with the conclusions derived from morphological considerations of these groups. The tardigrades are grouped with the Onychophora (PP 1.00; BS 79%).

I also conducted a bootstrap analysis on the 30 taxon dataset (tree not shown) using the CAT+GTR model, which is a mixed model in which equilibrium frequency profiles (those of CAT) are associated with heterogeneous substitution rates inferred from the dataset (R, the replacement matrix). The consensus tree from the analysis of 100 pseudo-replicates supports Mandibulata (BS 68; PP 1.00). Panarthropoda are also recovered (BS 64), but with Tardigrada as basal Panarthropoda instead of sister to the Onychophora.

I also analysed the full 59 taxon dataset using Bayesian inference and the CAT model (trees not shown). Results consistently support Mandibulata (PP 1.00), whereas the position of Tardigrada depends on the taxonomic sampling. Using all taxa, Tardigrada are grouped with Nematoda with weak support (PP 0.86), but exclusion of Nematoda recovers a group of Tardigrada plus Onychophora (PP 0.98). In the presence of nematodes, Acari (mites) are paraphyletic, due to fast evolving *Suidasia medanensis* being more basal, suggesting that fast evolving nematodes may promote artefactual reconstruction. This supports the possibility that the grouping of Nematoda plus Tardigrada is an artifact.

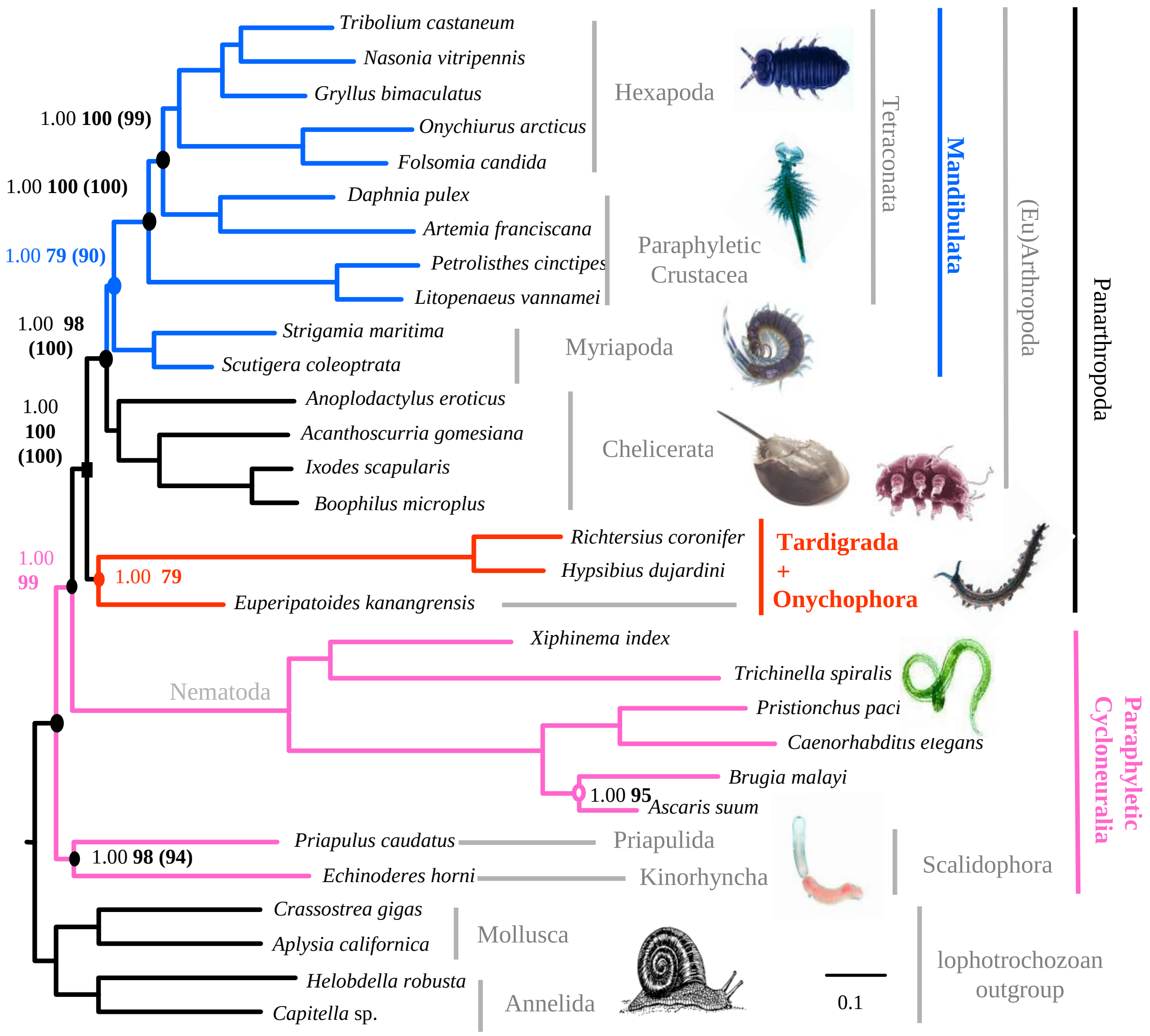


Figure 6.1. Bayesian analyses using the CAT model. Analyses support a monophyletic group of Mandibulata (Myriapoda, Hexapoda and Crustacea: black circle) and a monophyletic group of Panarthropoda (Arthropoda, Tardigrada and Onychophora: black square). Values at nodes correspond to posterior probabilities (plain text) from two independent runs and bootstrap support from 100 pseudo-replicates (in bold). Values in brackets are the bootstrap supports for the same dataset reanalysed without the long branched Nematoda and Tardigrada lineages. Where not shown, support corresponds to a posterior probability of 1.00 and bootstrap support 100%.

6.2.3 *Controversial signal using homogeneous models: the effect of taxonomic sampling and LBA.*

To understand the discrepancies between my analyses and previous work (Dunn et al. 2008, which using the CAT model recovers Myriochelata), I considered evidence for the effects of long-branch attraction (LBA) artefact, which arises from unequal rates of evolution among taxa (Felsenstein 1978 Lartillot and Philippe 2008). In this context, one notable aspect of the tree in figure 6.1 (and other trees describing myriapod relationships throughout this thesis) is the very different branch lengths seen in various taxonomic groups; tardigrades and nematodes have particularly long branches, and within the Euarthropoda, branches within Tetraconata are longer than those amongst myriapods and chelicerates. This distribution of branch lengths suggests that a systematic error could create the discord between previous molecular analyses and morphology: fast evolving (long-branch) tardigrades and nematodes may be artefactually associated with each other and the fast evolving Tetraconata could have been attracted towards the distant outgroup, resulting in an artefactual grouping of the more slowly evolving myriapods and chelicerates.

If support for Myriochelata is due to the Tetraconata being attracted to the distant outgroup species, one can predict that this artefact would be exacerbated by the use of outgroups with the longest branches and ameliorated when more slowly evolving outgroups are used. I have reanalysed my data set using several different outgroups with differing branch lengths (Fig 6.2). Analyses were performed using WAG and GTR models, which, contrary to the CAT model, assume homogeneity of the substitution process across sites following the procedures of the majority of previous studies (Dunn et al. 2008). Using my full complement of outgroup taxa I get a low level of support for Mandibulata over Myriochelata (Figure 6.2A). To see the effects of exaggerating potential LBA I used either the most phylogenetically distant outgroup (Lophotrochozoa) or the fastest ecdysozoan outgroup (nematodes). Under these conditions (figure 6.2B and 6.2C) I get support for Myriochelata rather than Mandibulata. In contrast, when I removed the fast evolving nematodes and tardigrades and the phylogenetically distant Lophotrochozoa (Fig 6.2D), support for Mandibulata increases. Notably, the GTR model, which fits the data better (accordingly to AIC test of model fit: data not shown) and therefore is a better estimator than WAG, consistently supports Mandibulata more robustly (compare values in plain and bold text in Fig. 6.2). Interestingly, when fast evolving nematodes and tardigrades are excluded, the CAT bootstrap support for Mandibulata increases from 79% to 90% (values in brackets in figure 6.1).

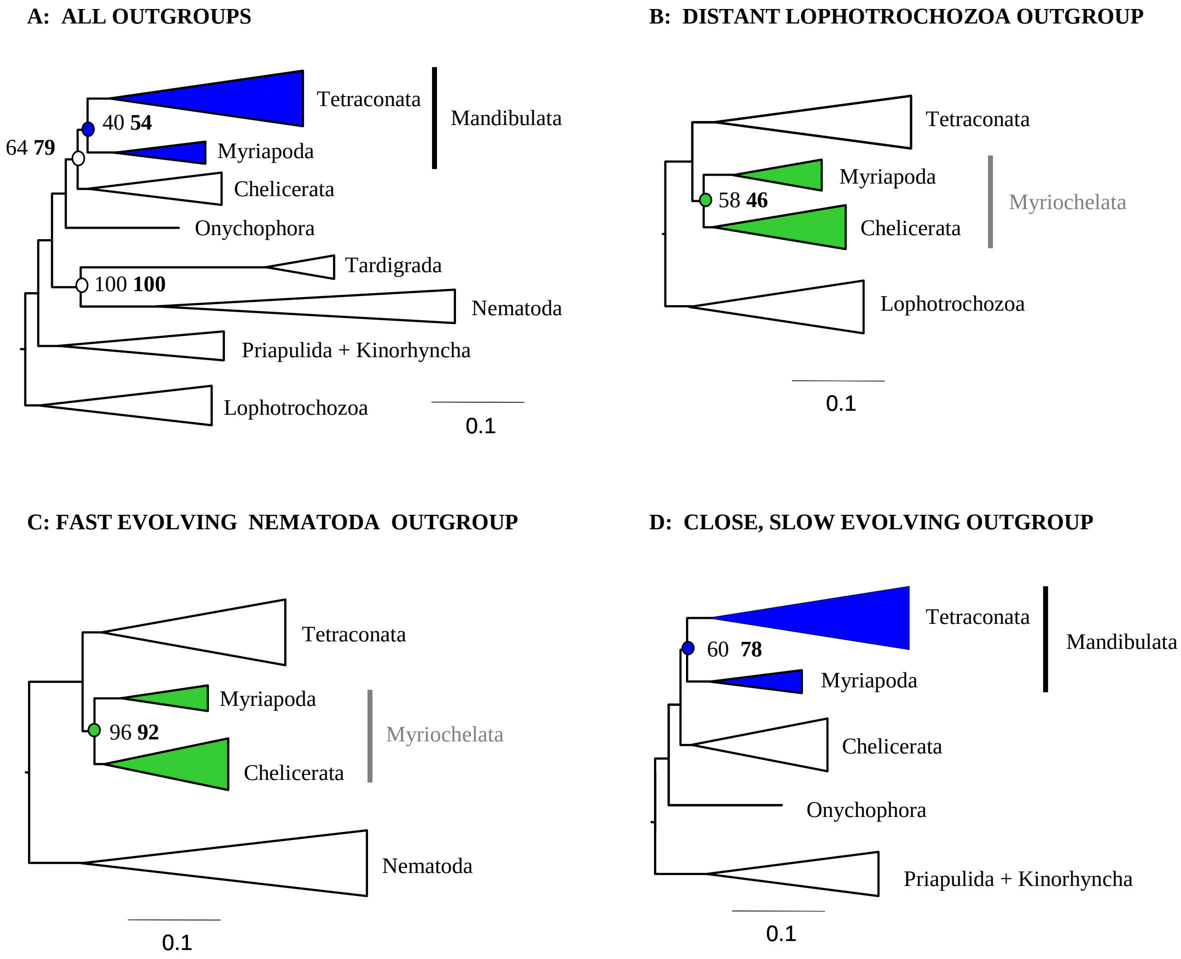


Figure 6.2. Taxon sampling and the effects of LBA. Phylogenetic analyses of my 30 taxa dataset using different taxon samples and maximum likelihood inference. (A) Support for Mandibulata (blue node and lineages) is low using the full dataset. Phylogenetically distant Lophotrochozoa (B) and fast evolving Nematoda (C) outgroups promote a possible LBA with the fast evolving Tetraconata lineage, leaving the slow evolving Myriapoda and Chelicerata together (Myriochelata, green nodes and lineages). Using slowly evolving and phylogenetically close ecdysozoans outgroups (D) increases support for Mandibulata. A similar analysis using the CAT model consistently recovers Mandibulata with PP 1.00. Tree topologies correspond to the whole dataset Maximum likelihood WAG trees and values at nodes are bootstrap supports from 100 replicates using the WAG (plain text) and GTR (bold text) models. Lineages have been collapsed for clarity with the length of triangles equal to the longest terminal branch in the collapsed lineage and stem branches equal to the originals.

The site-heterogeneous CAT model, which has been shown to fit real data better than site-homogeneous models (e.g. WAG and GTR) according with a crossvalidation test (Lartillot and Philippe 2004, data not shown), appears to be much less sensitive to the variations of taxon sampling performed above, since it always recovers Mandibulata with high PP (higher than 0.73) regardless of which outgroup set is used. Overall, conditions that reduce LBA show the highest support for Mandibulata, whereas conditions that increase LBA result in more support for Myriochelata.

6.2.4 *Support for Mandibulata from the reanalysis of Dunn dataset*.

In contrast to my analyses, the phylogenomic study of Dunn et al. (2008), hereafter only Dunn, supports a chelicerate affinity for the myriapods (Myriochelata). I have reanalysed a subset of Dunn's original dataset centred on Ecdysozoa, which resulted in a tree similar to that obtained by Dunn using their complete 77 sprcies dataset (green node of tree in figure 6.3A). To test if the difference between my phylogeny (which supports Mandibulata) and that of Dunn (which favours Myriochelata) is due to taxonomic representation, I expanded the Dunn dataset to include all 30 of my taxa and found that the support for Myriochelata decreased using both CAT and the WAG models (green node in figure 6.3B). The WAG model groups Myriapoda with Chelicerata – an improbable addition to this clade are the Onychophora (BS 80%) - suggesting a possible additional LBA effect with these data. Interestingly, the use of slowly evolving outgroups partially recovers Mandibulata while also supporting monophyletic Euarthropoda (figure 6.3.C).

Two other experiments designed to reduce systematic errors dissolved the spurious grouping of Onychophora and Myriapoda while also giving clear support for Mandibulata. First, with the removal of fast evolving sites, which are the most likely cause of systematic errors, support for monophyletic Euarthropoda increases (to a maximum of BS 90%), and under these conditions Mandibulata is the favoured topology (figure 6.4A). Second, the CAT+covarion model (Zhou et al. 2007), which tackles model violations by allowing rates of evolution of sites to change across the tree, recovers both monophyletic Euarthropoda and Mandibulata (PP of 1.0 for both clades, tree not shown). My interpretation of these results is that support for Myriochelata and for the grouping of myriapods/onychophoran may be attributed to a similar misleading signal due to systematic errors.

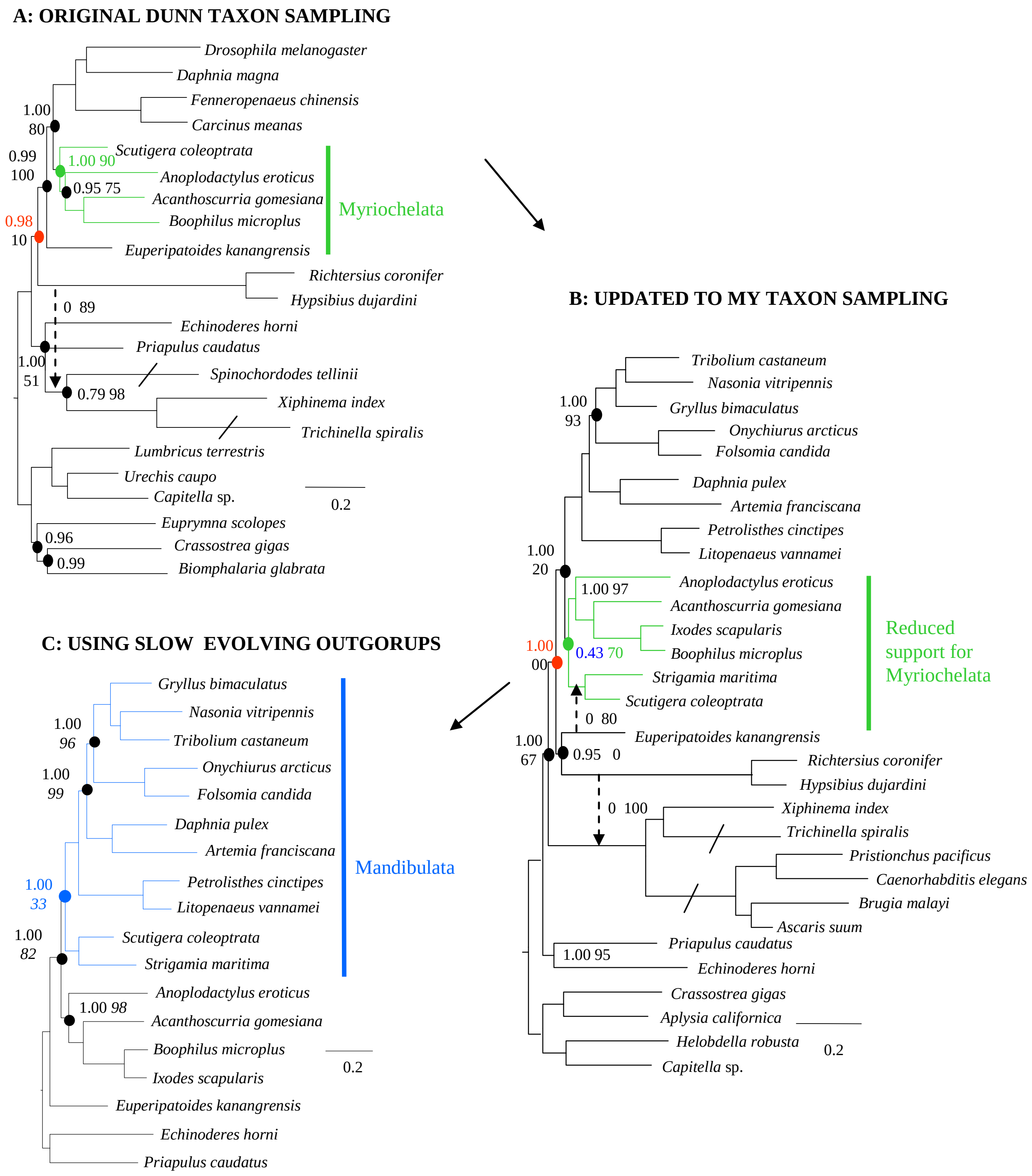


Figure 6.3. Bayesian and Maximum likelihood analyses of Dunn et al. (2008) dataset. Consensus trees from the Bayesian analyses using CAT. Values at nodes are the posterior probabilities using CAT and the bootstrap support using WAG. Where not indicated PP are 1.00 and BS 100%. The trees depict (A) the Dunn et al. original Ecdysozoan taxon sampling and (B) their gene set updated to my taxon sampling and (C) their gene set on my taxon sampling when slow evolving outgroups are used. Arrows indicate a different topology using WAG. Note that in tree B, WAG bootstrapped analyses support a group of Myriapoda plus Onychophora (see figure 6.4 A). Some nematodes branches have been halved to fit the figure.

Additional support for Mandibulata comes from a detailed exploration of the effect of taxon sampling on the Dunn et al. gene set. Similar to my dataset, outgroups that exaggerate the effect of LBA result in an increased support for Myriochelata, while outgroups that lessen LBA recover monophyletic Euarthropoda and Mandibulata (trees not shown). A similar outgroup analysis using the CAT model gave greater support for Mandibulata. In all trees, support for Myriochelata and for the myriapod/onychophoran grouping is much lower using GTR, a model which is a better estimator than WAG (using the AIC the difference between the two models is in the range of thousands in favour of GTR, data not shown).

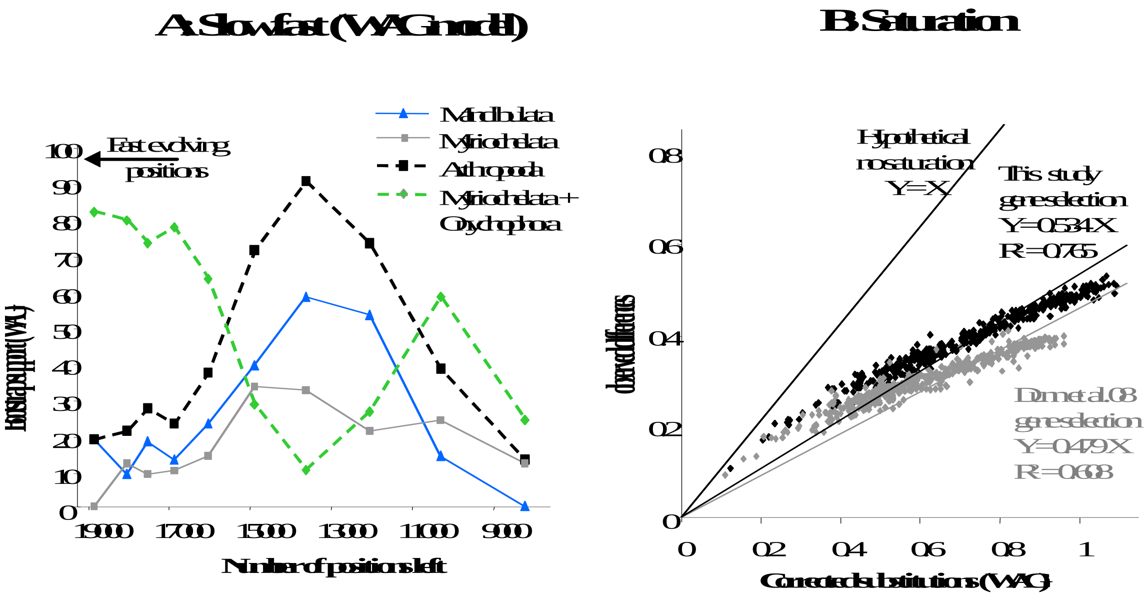


Figure 6.4: Signal exploration in the dataset of Dunn et al (2008) and mine. In (A) is the slow fast analysis conducted on the dataset of Dunne et al (2008) updated to my taxon sampling: classes of sites with different rates of evolution have been sequentially removed from the original alignments beginning with the fastest and sub-datasets analysed using the WAG model. An unlikely group of Myriapoda plus Onychophora (green dashed line) is associated with fast evolving positions; when Arthropoda (black dashed line) are recovered as monophyletic, Mandibulata is most supproted (blue line). The same analysis on my gene selection using WAG gave similar results. (B) Saturation analysis of the dataset of Dunn et al. (2008) using my taxon sampling (in grey) compared with my dataset (in black). Observed pairwise differences are plotted against the substitutions corrected by the WAG model to check for the level of saturation. Pearson's coefficient of regression (R^2) is higher for my dataset (data better fit a line) and the slope of the regression line is higher, suggesting that my gene set is less saturated than that of Dunn et al (2008).

Overall, my gene set always provides more support for Mandibulata than do the 150 genes of Dunn (compare figure 6.1 with 6.3B). The difference between Dunn and my analyses does not seem to be solely due either to taxon sampling or to tree reconstruction methods but may be partly explained by (i) the larger gene set of my dataset (approximately 40,000 versus 19,000 amino acid positions) (ii) the smaller amount of missing positions in my dataset (31% versus 39%) and (iii) the lower substitutional saturation seen in my genes (figure 6.4B). Less saturated data are preferred *a priori* as they have fewer homoplastic changes and are less susceptible to systematic error; it follows that clades (e.g. Mandibulata) supported by my less saturated data are more likely to be correct.

6.2.5 Evidence for monophyletic Panarthropoda, paraphyletic Cycloneuralia and monophyletic Chelicerata.

While less consistently supported than Mandibulata, several of my phylognemic analyses group the tardigrades with the other panarthropods. Analyses using the CAT model (figure 6.1) support a monophyletic group of Panarthropoda and a sister relationship between onychophorans and tardigrades. This finding is reinforced by the CAT analysis of the Dunn gene set (figure 6.3B). Analyses using CAT+GTR also support Panarthropoda, but with tardigrades as sister to a group composed of onychophorans plus Euarthropoda, in accordance with one of the analyses of Dunn et al (2008). These results are in contradiction with the bootstrapped analyses using WAG and GTR, which robustly support a group of tardigrades plus nematodes, both using my set of genes (figure 6.2) and the set of Dunn et al. (figure 6.3). However, I have found that in the absence of nematodes (trees not shown) support for a group of tardigrades plus onychophoran is recovered. These findings suggests that when accommodating for LBA artifacts (using CAT instead of homogeneous models or excluding fast evolving lineages) support for an arthropod (and more specifically onychophoran) affinity of the tardigrades is recovered. All my analyses also support priapulid worms to be sister of a group composed of nematodes plus Euarthropoda. This result challenges the idea of monophyletic Cylconeuralia as for example supported by Dunn et al. (2008).

All my analysis support monophyletic Chelicerata, with the pycnogonid *Anoplodactylus* sister to the Arachnida, in accordance with the phylogenomic study of Dunn et al. (2008) but in disagreement with nuclear proteins and ribosomal markers, which support a paraphyletic scenario (Regier et al. 2008, Mallatt and Giribet 2006). It is interesting to note that conditions that worsen the effect of LBA in the dataset (using fast evolving nematodes or distant lophotrochozoans) result in a decrease in support for monophyletic Chelicerata (trees not shown). My interpretation, again, is that distant outgroups to the arthropods tend to attract the relatively fast evolving pycnogonid. Conditions which lessen LBA, such as use of heterogeneous CAT models or close outgroups, strongly support Pycnogonida as basal Chelicerata in all my analyses. This result makes sense on the grounds of morphological evidence, although the only synapomorphy uniting Pycnogonida and other (Eu)chelicerata is the presence of a pincer-like appendage in their first appendage bearing segment (chelifores in Pycnogonida and chelicerae in Euchelicerata). A neuroanatomical study initially proposed that the appendages in the two groups are innervated by different brain regions (Maxmen et al. 2005), but subsequent developmental gene expression analysis has proved that both appendages arise form the same deutocerebral region, supporting a common origin of the two (Jager et al. 2006). Finally, all analyses strongly support a close relationship between insects and collembolans and between hexapods and branchiopod crustaceans, but these relationships will be addressed in detail in the next section of this chapter.

6.3 A group of monophyletic hemimetabolan insects and unresolved crustacean relationships

In the previous section (6.1) I addressed the affinities of basal arthropod groups; in particular the relative positions of myriapods and chelicerates. In this section I will concentrate on the internal phylogeny of the remaining large group of arthropods, the Tetraconata. As discussed in chapter 1.3, crustaceans have been suggested to be paraphyletic, but there is no consensus between various molecular markers as to how the crustaceans are related to each other and how the hexapods fit into the Tetraconata assemblage. While the affinities of the holometabolan insect orders are fairly clear, relationships among hemimetabolan orders are extremely debated and confused, to the extent that different markers support different hypotheses and some markers are even self-inconsistent. A possible explanation of this great level of uncertainty is the lack of suitable molecular datasets. Existing taxonomically dense datasets, such as the ribosomal ones, are short and prone to stochastic errors and larger phylogenomics datasets are poorly taxonomically sampled and prone to systematic errors.

6.3.1 Two phylogenomic datasets centred respectively on Crustacea and Hexapoda.

In an attempt to overcome these problems (treated more extensively in Chapter 1.2) and taking advantage of a considerable number of new ESTs in the public databases, I assembled two large phylogenomic datasets, one centred on insects (51 taxa and 205 genes) and one centred on crustaceans (41 taxa and 149 genes). The generation of 2 distinct datasets has a twofold advantage: first, it minimises the amount of missing data in the dataset and second makes some analyses computationally more tractable (for the same reasons I have separately assembled the “myriapod” dataset used in the previous section). The datasets have been assembled again in collaboration with Hervé Philippe and Henner Brinkmann, following the same procedure used for the “myriapod dataset” of previous section. We used chimeric sequences in order to have as few missing positions as possible. I analysed these datasets using both homogeneous LG (Lee and Gasuel 2008) and heterogeneous CAT (Lartillot and Philippe 2004) models of protein evolution under a Maximum likelihood and a Bayesian framework respectively.

The corresponding phylogenies (figure 6.5) clearly support monophyly of hexapods with the entognathan Collembola sistergroup to the remaining ectognathan insects, reinforcing the unique origin of the six legged body plan and further highlighting the conflict between nuclear (Timmermans et al. 2008) and mitochondrial markers (Carapelli et al 2007), the latter supporting paraphyletic hexapods.

6.3.2 A monophyletic group of hemimetabolan insects.

The most interesting outcome of my analyses is a split of the insects into two distinct monophyletic groups: one comprising holometabolan and the other hemimetabolan insect (Figure 6.5, green lineages and node). Monophyly of holometabolan insects is predicted by morphology, in particular by the shared metamorphosis through pupal stage, and has been confirmed by many molecular markers (Timmermans et al. 2008, Mallat and Giribet 2006, among the others).

On the other hand, a monophyletic group of hemimetabolan insects, challenges the commonly accepted Eumetabola, which groups the hemipteroids with the holometabolans and has found molecular support from the analysis of ribosomal subunits (Kjer 2004, Wheeler et al. 2001). The use of more sophisticated methods (Mallat and Giribet 2006) and larger markers (Timmermans et al. 2008, Roeding et al. 2009) have lead the hemipteroids (the hemipterans in their sampling) to be mutually paraphyletic with the polyneopterans. Furthermore, phylogenomic datasets have recovered a sister relationship between hemipteroids and polyneopterans (monophyletic hemimetabolans), although this support was model dependent (Lartillot and Philippe 2008). Finally, my dataset, which is larger in size and in taxon sampling, supports monophyletic hemimetabolans using both homogenous and heterogeneous models. This evidence, taken together, suggests that the more the stochastic and/or systematic errors are reduced (using more genes and/or more taxa), the more the support for eumetabolans disappears in favour of a monophyletic group of hemimetabolan insects.

This view is corroborated by the analysis of the “crustacean dataset” (discussed in 6.2.3, full tree not shown), which supports the monophyletic group of hemimetabolans with less support (BS 58% using LG) than the “insect” dataset (BS 85%). This discrepancy in support values may be explained by the crustacean dataset containing fewer insects and fewer genes, thus being more prone to systematic errors than the insect dataset. Finally, there seems a low likelihood that the grouping of hemipteroids and orthopteroid is the result of long branch attraction as inspection of figure 6.5 suggests that while hemipteroid lineages are markedly fast evolving, the orthopteroid ones are clearly slow evolving.

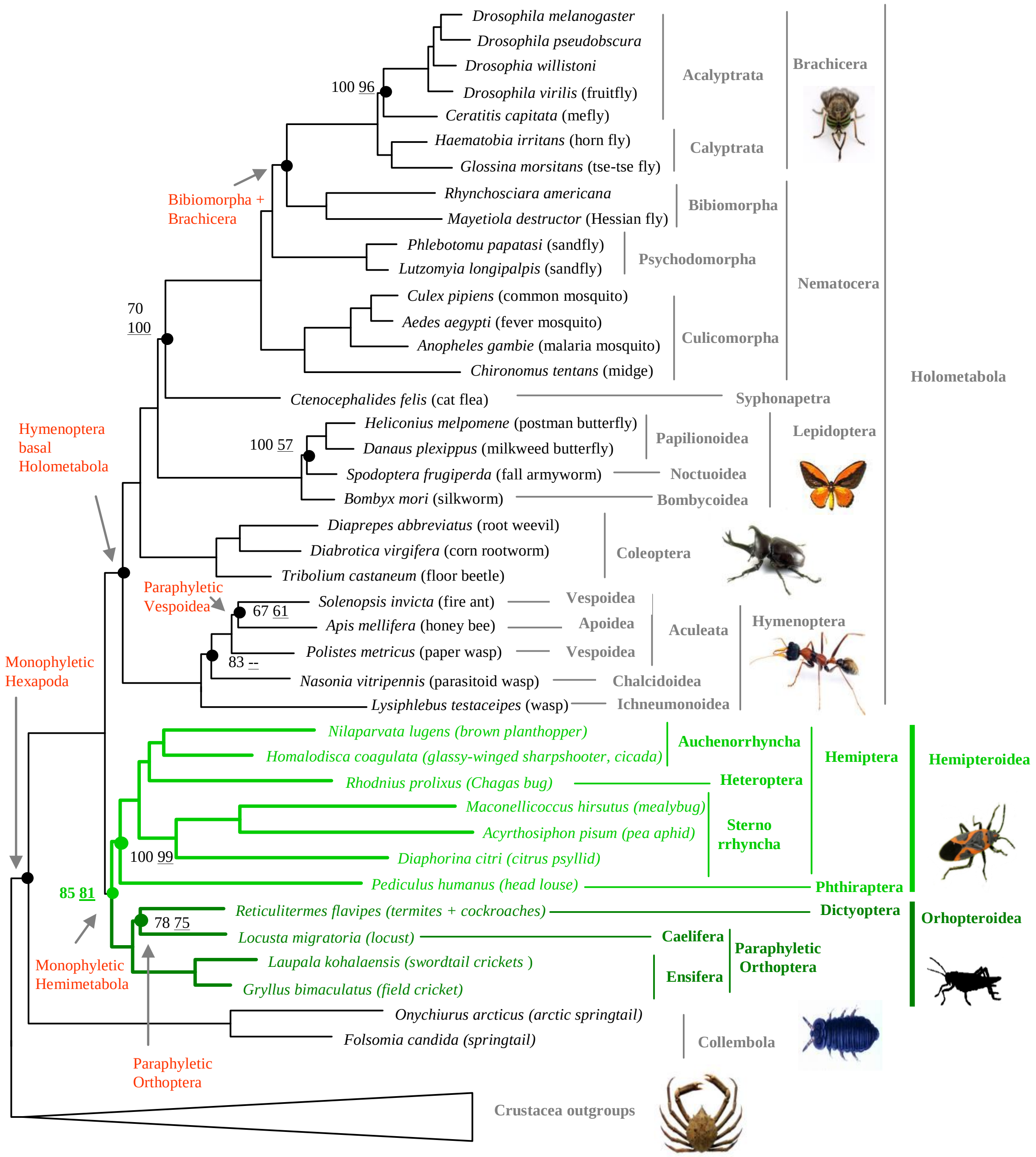

Drosophila melanogaster
Drosophila pseudobscura
Drosophia willistoni
Drosophila virilis (fruitfly)
Ceratitis capitata (mefly)
Haematobia irritans (horn fly)
Glossina morsitans (tse-tse fly)
Rhynchosciara americana
Mayetiola destructor (Hessian fly)
Phlebotomu papatasi (sandfly)
Lutzomyia longipalpis (sandfly)
Culex pipiens (common mosquito)
Aedes aegypti (fever mosquito)
Anopheles gambie (malaria mosquito)
Chironomus tentans (midge)
Ctenocephalides felis (cat flea)
Heliconius melpomene (postman butterfly)
Danaus plexippus (milkweed butterfly)
Spodoptera frugiperda (fall armyworm)
Bombyx mori (silkworm)
Diaprepes abbreviatus (root weevil)
Diabrotica virgifera (corn rootworm)
Tribolium castaneum (floor beetle)
Solenopsis invicta (fire ant)
Apis mellifera (honey bee)
Polistes metricus (paper wasp)
Nasonia vitripennis (parasitoid wasp)
Lysiphlebus testaceipes (wasp)
Nilaparvata lugens (brown planthopper)
Homalodisca coagulata (glassy-winged sharpshooter, cicada)
Rhodnius prolixus (Chagas bug)
Maconellicoccus hirsutus (mealybug)
Acyrthosiphon pisum (pea aphid)
Diaphorina citri (citrus psyllid)
Pediculus humanus (head louse)
Reticulitermes flavipes (termites + cockroaches)
Locusta migratoria (locust)
Laupala kohalaensis (swordtail crickets)
Gryllus bimaculatus (field cricket)
Onychiurus arcticus (arctic springtail)
Folsomia candida (springtail)
Crustacea outgroups
100 96
70
100
100 57
67 61
83 --
100 99
85 81
78 75
Acalyptrata
Calyptrata
Brachicera
Bibiomorpha
Psychodomorpha
Nematocera
Culicomorpha
Syphonapetra
Holometabola
Papilionoidea
Noctuoidea
Bombycoidea
Lepidoptera
Coleoptera
Vespoidea
Apoidea
Vespoidea
Aculeata
Hymenoptera
Chalcidoidea
Ichneumonoidea
Auchenorrhyncha
Hemiptera
Hemipteroidea
Heteroptera
Sterno
rrhyncha
Phthiraptera
Dictyoptera
Orhopteroidea
Caelifera
Paraphyletic
Orthoptera
Ensifera
Collembola
Bibiomorpha +
Brachicera
Hymenoptera
basal
Holometabola
Paraphyletic
Vespoidea
Monophyletic
Hexapoda
Monophyletic
Hemimetabola
Paraphyletic
Orthoptera

Figure 6.5.: Phylogenetic analyses support a monophyletic group of hemimetabolan insects. Consensus tree from the bootstrapped Maximum Likelihood and Bayesian analysis using respectively the LG and the CAT model. Values at nodes are the bootstrap support (BS) for LG (plain text) and CAT (underlined). Were not specified BS are 100 for both models. Common name of species are in brackets and in red novelties or interesting results. Relationships within crustaceans have been collapsed for clarity and correspond to those in figure 6.6A.

The monophyletic hemimetabolan assemblage of figure 6.5 consist of a sister relationship between orthopteroids (Orthoptera plus Dictyoptera in my sample) and hemipteroids (Hemiptera plus Phthiraptera), the two groups being respectivly monophyeltic. Notably, hemipterans are also monophyletic in my analysis in contrast with results from recent phylogenomic datasets (Timmermans et al. 2008, Roeding et al. 2009) thus reinforcing the synapomorphic nature of the hemipteran rostrum, which uniquely within the insects is originated by the fusion of the mandible and the maxillae. In my tree, the hemipteran Heteroptera (true bugs) are sister of the Auchenorrhyncha (other bugs such as cicadas), this group being sister of the Sternorrhyncha (aphids), in accordance with a common view of hemipteran phylogeny (http://tolweb.org/tree?group=Hemiptera). Notably, the dictyopterans are sister of *Locusta migratoria*, making the Orthoptera paraphyletic, though the modest support and the poor taxon sampling at this node, as well as the fast evolving nature of *Locusta* and the dictyopteran lineage, suggest a possible LBA artefact and further analyses, encompassing more taxa, are required. It is, however, clear that the orthopterans and dyctiopterans form a strongly supported monophyletic group, the orthopteroids. The latter group is supported by fossil evidence: while modern dictyopterans possesses short internal ovipositors, the first proto-dictyopteran fossils from the late carboniferous (Grimaldi 1997) had long external ovipositors like the members of the orthopterans.

As for the holometabolans, the hymenopterans (bees, ants and wasps) are basal within the holometabolans, in accordance with a recent large phylogenomic analyses (Savard et al. 2006) and analysis of ribosomal subunits (Mallat and Giribet 2006), but in contrast with a phylogenomic (ribosomal protein based) analysis which give moderate support to a sister relationship between Coleoptera and Hymenoptera (Timmermans et al 2008). Among the hymenopterans, Vespoidea (ants and paper wasps) are paraphyletic, in contrast with morphological characters (Brothers 1999). However, support for this is modest and may be either the effect of a long branch attraction between relatively fast evolving and undersampled bees and ants or of a convergent adaptive evolution of some of my markers as both lienages intriguingly share a similar eusocial population structure and inheritance strategy. Finally, relationships within Coleoptera are consistent with the more detailed work of Hughes and colleagues (2006). As for the Diptera, my analyses confirm the paraphyletic origin of Nematocera, but support the Bibiomorpha (Hessian flies) as closer related to the Brachicera (fruitflies and their kin) than the

Psychodomorpha (sand flies), in partial disagreement with a common interpretation of dipteran evolution (Yeates and Wiegmann 2005).

6.3.3 Crustacean affinities are model and outgroup dependent.

To clarify the relative positions of crustaceans and hexapods, I assembled a second dataset centered on crustaceans and using myriapods and chelicerates as outgroups. Analysis using CAT supports a sister group relationship between Copepoda (Maxillopoda) and Branchiopoda (figure 6.6C). This is in accordance with the Entomostraca assemblage which groups branchiopods, maxillopods and cephalocarids with the exclusion of the malacostracans.

However, this topology is incompatible with the unrooted tree of figure 6.6A (a schematic representation of tree in figure 6.5), which supports a sister relationship between malacostracans and copepods. Interestingly, members of the copepods resemble the malacostracans by having their neurons myelinated (Davis et. al. 99), a character which is a unique feature within the arthropods and can be interpreted as a synapomorphy supporting the malacostracans plus copepods clade. Interestingly, two previous attempts to assess crustacean relationships, consistently supported a group of malacostracans plus copepods (and Cirripedia were sampled, Regier et al. 2008 and Roeding et al. 2009).

The discrepancy between rooted and unrooted trees has been explored by inspecting the branch lengths of the lineages involved because there may be an indication of LBA. According to figure 6.6, hexapods and branchiopods are more slowly evolving than malacostracans and copepods, suggesting that the grouping of the two latter may be an artifact due to LBA. However, the branch leading to the outgroup (in figure 6.6B, C and D) is also long, suggesting that malacostracans may suffer from a reiterated LBA artefact. In order to explore this discrepancy in greater depth I have analysed the dataset using more sophisticated models of evolution. Using the CAT-covarion model, which allow rates to vary among branches, the trees are identical to those found using the non-covarion CAT model (figure 6.6C). Using the CAT+GTR model, the unrooted tree (figure 6.6A) is compatible with the rooted one (figure 6.6D), which supports a sister relationship between branchiopods and hexapods, with copepods more basal and malacostracans close to the outgroup. Analyses using the LG model support a rooted and an unrooted compatible tree (compare figure 6.6A and B) and a monophyletic origin of the crustaceans

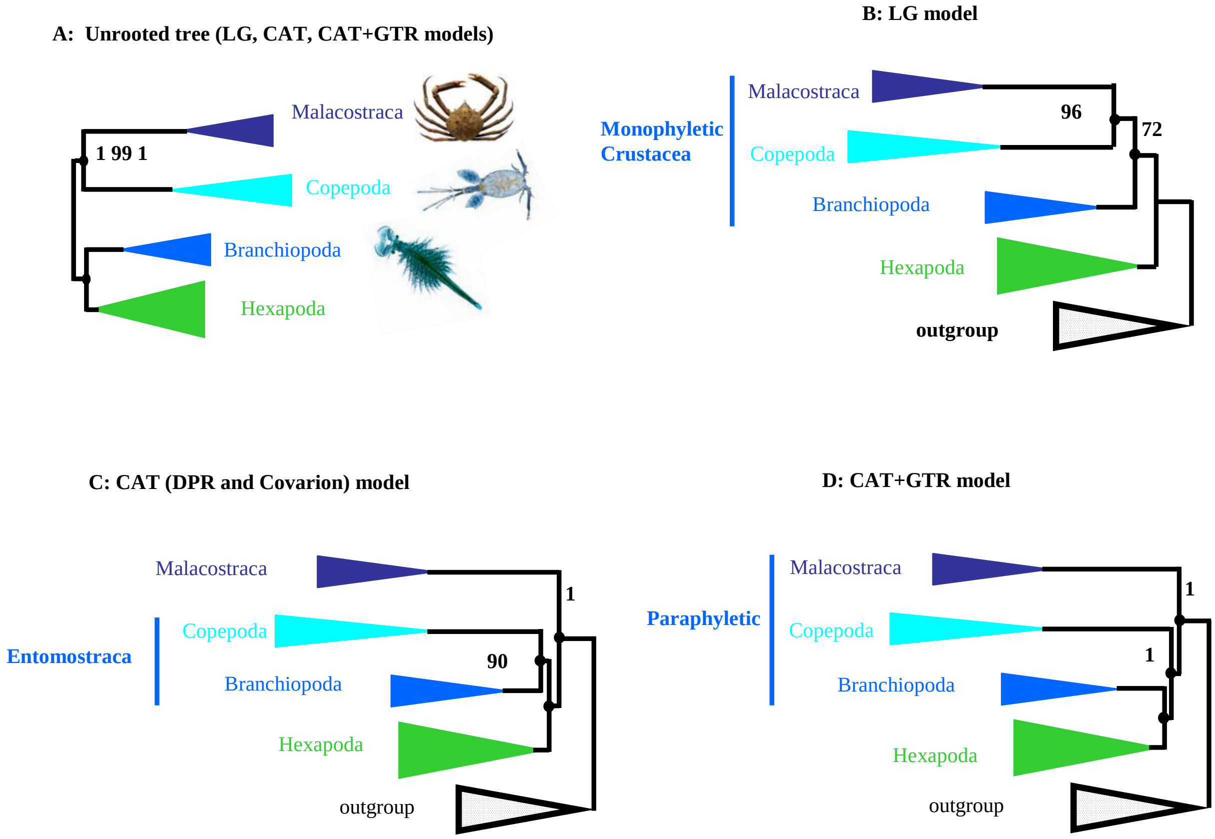


Figure 6.6. Crustacean relationships is model and outgroup dependent. (A) Schematic representation of the Bayesian consensus trees using LG, CAT and the CAT-GTR models using an unrooted dataset. The same dataset, rooted with the myriapods and chelicerates is analysed using LG (in B), CAT (in C) and CAT+GTR (in D). In C, topologies from the analysis of the rooted dataset is inconsistent with that made in the absence of the outgroup sequences. In most of the cases branchiopods are closer related to hexapods than the malacostracans are.

It is clear that different models support different topologies and that, in some cases, the topology in the presence of the outgroup is inconsistent with the unrooted tree. The easiest explanation is that my phylogenetic reconstructions are misled by a systematic error, probably due to a combination of fast evolving lineages (Copepoda and Malacostraca) and poor taxon sampling. I could not, however, observe a clear trend throughout different analyses, as I observed in the analyses of the "myriapod dataset" in chapter 6.1. These results taken together suggest that my crustacean dataset may not posses enough information to address crustacean relationships. However, it is possible to draw at least one conclusions: I do not observe the grouping of malacostracans plus branchiopods (Thoracopoda hypothesis) in any of the trees. Also, the branchiopods, either alone or with the copepods, are always observed as sister to the hexapods (or closer to them than the other crustacean classes), partially polarizing the quartet including hexapods, branchiopods, malacostracans and outgroup.

6.4 Conclusions: support for Mandibulata, Panarthropoda and a monophyletic group of hemimetabolans.

Support for a monophyletic origin of myriapods, hexapods and crustaceans (Mandibulata) from the analyses of my phylogenomic dataset is high to moderate in my phylogenomic analyses using a variety of methods, taxa, models and site selection. I showed that occasional support for Myriochelata (the competing hypothesis to Mandibulata which groups myriapods and chelicerates) may be related to an LBA artefact. Support for Myriochelata is restricted for example to analyses using the poorly fitting WAG model and is associated with a group of onychophorans plus myriapods; however, support for this grouping is limited to fast evolving and/or incomplete positions, while slower evolving sites support Mandibulata. Furthermore, Myriochelata tend to be recovered when fast evolving outgroups are used, while conditions which lessen LBA increase support for Mandibulata. The LBA nature of the Myriochelata group is corroborated by the re-analysis of a published phylogenomic dataset which supported Myriochelata (Dunn et al. 2008). When this dataset is updated to my larger taxon sampling and is analysed under conditions which lessen LBA artefacts, I recover Mandibulata.

My analyses also suggest that studies that have grouped tardigrades with nematodes may have been similarly affected by LBA. When analysed using the CAT model, which has been shown to help in overcoming systematic errors, both my data set and that of Dunn et al. (2008) support a panarthropod affinity for tardigrades. In some of my analyses tardigrades are sister to the onychophorans, and, since the onychophorans are slow evolving and tardigrades are fast evolving, there seems a low likelihood of LBA. However, it is clear that both the reduced taxonomic and gene sampling for either tardigrades and onychophorans suggest that actual signal in my dataset is not probably enough to draw firm conclusions, although it gives some indications of a panarthropod nature of the water bears and a consequent artefactual grouping of tardigrades and nematodes.

My insect-centred phylogenomic dataset, although the largest to date both in terms of genes employed and taxa sampled, describes only a small fraction of the incredible Tetraconata diversity encompassing for example only ten insect orders out of a total of 30. On the other hand, the large number of genes and amino acid positions employed in my datasets allows us confidently to draw some conclusions on the evolution of insects and their crustacean relatives. My analyses strongly support monophyly of some well-established clades such as hexapods and insects and confirms recent findings that the hymenopterans are basal within (monophyletic) Holometabola. My analyses are concordant in supporting an interesting monophyletic group of hemimetabolan insects, composed of a sister relationship between hemipteroids (Hemiptera and Phthiraptera in my sample) and orthopteroids (Orthoptera and Dictyoptera).

Chapter 7
Materials, methods and pipelines

This chapter contains the methodological aspects of my analyses. Each section of this chapter covers the methods of one particular result chapter. For this reason, some of the information presented here are apparently redundant. For example, inference of phylogeny using MrBayes (Huelsenbeck and Ronquist, 2001) as been described more than once. However, similar analyses have been treated slightly differently for two reasons: first of all each dataset needs specific settings; secondly, recent analyses have been carried out with up to date pipelines or solution than earlier analyses. For the sake of clarity, I have numbered sections of this chapter like the thesis chapters (for example section 4 of this chapter contains the materials and methods of chapter 4).

7.1 Wet lab

I sequenced the partial mitochondrial genome of the onychophoran *Euperipatiodes kanagrensis* in order to do phylogenetic studies. I sequenced completely the genes Cox1, Cox2 and Nadh1 and partially CytB, 16S, 12S and Nadh4. I also tested the quality of two cDNA libraries, one of which has been used to sequence 5000 EST, using an external facility. The fragments of interest were first amplified by the polymerase chain reaction (PCR), cloned into plasmid vectors and minipreped for purification of the fragment. The isolated DNA has been sequenced using a BigDye strategy and AB-sequencing.

7.1.1 Polymerase chain reaction

I used standard PCR conditions to amplify fragments for cloning. Reactions were carried out using the Roche Taq DNA Polymerase set (Cat. No. 1 596 594), the AB gene dNTP set (Cat. No. AB-0315) and with primers ordered from Thermo Electron or

MWG. dNTP and primer stocks were diluted to the given concentration with Milli-Q water. Reactions were carried out in a total volume of 30 µl with the following volumes of reagents:

4.0 µl 10x buffer
22.4 µl Milli-Q water
2.0 µl dNTP (5mM)
0.2 µl Forward primer (10nM)
0.2 µl Reverse primer (10nM)
1.0 µl DNA
0.2 µl Taq DNA polymerase

I carried out PCR reactions in a G-Storm Thermal Cycler. Melting temperatures were estimated with "the Wallace rule": Tm (in ºC) = 2(A+T) + 4(G+C). A basic PCR cycle was used consisting of 1 cycle extended DNA denaturation of 2 min at 94ºC, followed by 35 cycles of 30 sec denaturation at 94ºC, 30 sec annealing at a temperature as calculated above and extension for the appropriate length of time at 72ºC, followed by a final extension step of 10 min at 72ºC. DNA was from a cDNA library provided by Joakim Eriksson, Cambridge. Degenerate primers for the amplification of *Euperipatoides kanagrensis* mitochondrial genes have been designed according with a large nucleotide alignment of all the mitochondrial coding genes of arthropods and a list of primers kindly provided by Chuck Cook, Cambridge University.

7.1.2 PCR product isolation and purification

To isolate the fragment, the PCR products were separated by agarose gel electrophoresis. Gels were made with 1% agarose in 1x TBE or 1x TAE. Ethidium bromide was added to the gel (approximately 1 µl (at 10 mg/ml) per 200 ml) and a 1 kb ladder (Invitrogen 1 kb DNA Ladder; Cat. No. 15615-024) was run with the samples. DNA was visualised on a UV light box and bands of the expected size were excised with a scalpel. The excised DNA was purified from the gel using the QIAGEN MinElute Gel Extraction Kit (Cat. No. 28606), which produces a concentrated DNA extract in a volume of 10 µl; 2 µl of the purified PCR product were run on an agarose gel to confirm purification. A typical protocol:

- cut and weight the gel slice, add 3 parts of QG buffer
- keep at 50^{o}C for 10 minutes
- add 1 part of isopropanol and centrifuge for 1 minute at 13k rpm
- discard and add 500 µl of QG, centrifuge for 1 minute at 13k rpm
- discard and add 750 µl of PE buffer, centrifuge for 1 minute at 13k rpm
- repeat centrifugation putting the filter in a clean tube
- elute DNA using 10 µl of TRIS-hcl PH8.5 and 1 minute at 13k rpm

7.1.3 Cloning, colony PCR and Minipreps

Cloning

The purified PCR fragments were cloned into the TOPO TA cloning pCR II-TOPO, which uses a topoisomerase to insert the product into the vector (Cat. No. K4600-40) or Promega pGEM-T Easy vector, which uses a ligase (Cat. No. A1360). For both kits, the products of the cloning reaction were transformed into the TOPO TA cloning TOP10F' chemically competent *E. coli* cells (Cat. No. K4650-40), or New England Biolabs NEB 5-alpha competent *E. coli* cells (Cat. No. C2991H) Transformations were carried out as follow:

- Rapid centrifugation of vectors.
- put 2 µl of vector in 50 µl of frozen cells
- keep on ice 30 minutes
- 42 ^{o}C water bath for 30 seconds
- back to ice
- add 250 µl of SOC medium

Transformed cells were plated after 60 minutes onto LB nutrient agar plates (7.5 g agar per 500 ml LB) containing carbenicillin (60 µg/ml). Plates were prepared by plating 40 µl X-gal (20 mg/ml in dimethlyformamide) and if TOP10F' or NEB 5-alpha cells had been used for the transformation 10 µl 100 mM IPTG. Both the pCR II-TOPO vector and the pGEM-T Easy vector contain an ampicillin resistance gene allowing transformed cells to grow in the presence of ampicillin. Both vectors also have their insert site within the ß-galactosidase gene and when grown in the presence of X-gal, cells with an insert have a disrupted ß-galactosidase and appear white.

Colony PCR

To confirm whether the insert was of the expected size, colony PCR was performed on the colonies, using primers designed to bind to the SP6 and T7 polymerase sites flanking the insert. Reactions were carried out in a total volume of 20 µl with the following volumes of reagents:

2 µl 10x buffer
15.35 µl Milli-Q water
2.0 µl dNTP (10mM)
0.2 µl SP6 primer (100nM)
0.2 µl T7 primer (100nM)
0.25 µl Taq DNA polymerase

Colony PCR was carried out in a thermocycler using a PCR cycle of: 1 cycle extended DNA denaturation of 2 min at 94ºC, 35 cycles of 30 sec denaturation at 94ºC, 45 sec annealing at 50ºC, 1 min extension at 72ºC and a final extension step of 7 min at 72ºC. Colonies were picked with a 10 µl pipette tip or a sterile toothpick and transferred to culture tubes containing 1 ml LB medium and carbenicillin (60 µg/ml) and grown overnight at 37ºC on a shaker at 200 rpm.

Minipreps

To isolate the plasmid DNA from the bacterial cells, minipreps were performed using the QIAGEN QIAprep Spin Miniprep Kit (Cat. No. 27106) according to the manufacturers instructions and the following protocol:

- transfer cells cultures in to 1.5 ml Eppendorf tubes
- pellet cells at 3000 rpm for 10 minutes
- resuspend with P2 and mix 6 times; wait 5 minutes (lysis)
- add N3 and mix 6 times; wait 5 minutes (denaturation)
- spin at 13k rpm for 10 minutes

- place surnatant in QIAprep tubes 13k rpm for 1 minute (binding)
- discard and add 500 µl PB, 3k rpm for 1 minute
- discard and add 750 µl PE, 3k rpm for 1 minute, repeat 3k rpm for 1 minute
- put QIAprep in clean Eppendorf, add 50 µl EB and spin (elution).

7.1.4 Sequence reaction and precipitation

Sequencing reactions were carried out using the Applied Biosystems BigDye Terminator v1.1 (or subsequent versions) Cycle Sequencing Kit (Cat. No. 4337450) in a total volume of 10 µl with the following volumes:

2 µl 5x BigDye sequencing buffer
3.5 µl Milli-Q water
1 µl sequencing primer (3 nM)
2.5 µl plasmid
1 µl BigDye Terminator ready reaction mix

Sequencing primers were designed to bind to the polymerase sites that flank the insert region (T7 and SP6 or T3 in the case of Strigamia cDNA library); each insert was sequenced from both ends. Sequencing was carried out in a thermocycler with a denaturation step of 1 cycle of 3 min at 96ºC, 25 cycles of 20 sec at 96ºC, 10 sec at 50ºC and 4 min at 60ºC. Sequencing reactions products were sent to the Natural History Museum Sequencing Facility or the Wolfson House DNA sequencing facility as dried DNA pellets for the sequences to be read using ABsequencing. To pellet the DNA, the product was precipitated by adding 20 µl Milli-Q water, 70 µl 100% ethanol, 2 µl sodium acetate (3 M) and incubating for 1 hr at room temperature. Precipitated DNA was pelleted by centrifugation at 13000 rpm in a microcentrifuge for 20 min. The liquid phase was discarded and the pellet washed by the addition of 100 µl 70% ethanol, which was then removed and the pellet left to dry by placing in a rack on a 50ºC heating block for approximately 15 min.

7.1.5 Centipede and Onychophora cDNA libraries screening

Libraries were kindly provided by Ariel Chipman (centipede *Strigamia maritima* library) and Joakim Eriksson (onychophoran *Euperipatoides kanagrensis* library) from the Akam lab in Cambridge.

Strigamia library consisted of an high concentration bacteria culture transformed with vector pSK (bluscript) and has been normalised to contain only large fragments (approx 1kb fragment). The best way to plate these cells was to take less than 1µg of frozen cells (the minimum amount as possible on the top of a 10 µl tip), dilute them in at least 1 ml of LB medium and spread 1ul in 40ul XGAL + 10ul IGPT Petri. Approximately 90% of cells were recombinant. *Euperipatoides* library consisted of vector PExCell extracted from LAMBDA phage and required cloning into competent *E. coli* cells following protocol described above. Transformed cells grew sensibly slower than normal and required at least 2 days at 37°C to make reasonable sized colonies, probably because the vector was not directly designed for this kind of cells. Tests suggested that the best dilution for transformation was 1µl of library in 9µl water. The optimal amount of Top10 transformed cells to be platted (in 40 µl XGAL and 10 µl of IGPT Petri) was 50 µl. Approximately 70% of cells wereare recombinant.

The two libraries have been screened to inspect the quality of their inserted fragment. The screening was made on 25 colonies for each of the library. Colony PCR, miniprep and sequence reaction were performed as previously described using the T7 and T3 primers (for the *Strigamia* library) and with T7 and SP6 (for the *Euperipatoides* library). Average length of fragment was 700nn for the *Strigamia* library and 840 nn for the *Euperipatoides* one. A reliable similarity with the NCBI protein bank was detected for 17 out of 25 fragments in *Strigamia*. 80% of the inserts were oriented with 5’ on the side of T7. Following this indication, the external facility Macrogen has been instructed to sequence the 5000 ESTs from the *Strigamia* library using T7.

7.2 Estimation of evolutionary models (methods of chapter 2)

7.2.1 Dataset for the estimation of the models

I assembled two distinct alignments of the 13 mitochondrial coded proteins. For the estimation of MtZoa I carefully chose 108 metazoan species, consisting of 22 lophotrochozoans, 39 deuterostomes, 39 ecdysozoans and 8 non-bilaterians. For the estimation of MtHydro, I assembled an alignment of 100 metazoan species, consisting of 48 protostomes (of which 18 were lophotrochozoans) and 52 deuterostomes (of which 20 were non-vertebrates). The dataset for MtHydro did not contain non-bilaterians because I observed substantial differences between bilaterian and non-bilaterian secondary structures, which would have complicated the calculation of a consensus for the metazoans.

For both alignments, I constructed the corresponding tree in order to best reflect current knowledge of metazoan relationships and the so called “new animal phylogeny” (Telford et al. 2008, Webster et al. 2006, Dunn et al. 2008). The tree for the estimation of MtZoa can be inspected in figure 2.1. I did not incorporate sequences from lineages characterised by extremely accelerated substitution rates, such as urochordates, nematodes and platyhelminthes, in order to minimize the degree of saturation of substitutions in the alignment and to avoid the generation of a corresponding highly saturated substitution matrix. I excluded poorly aligned and unconserved sites using Gblocks (Castresana, 2000) at default settings for MtZoa and with the following settings for MtHydro, B1=N/2=50 B2=N/2=50 B3=6 B4=4 B5=half gaps. Both alignments have been followed by manual refinement resulting in an alignment of 2589 and 2737 amino acid positions for MtZoa and MtHydro respectivly.

7.2.2 Bioinformatic analyses to predict protein secondary structure

The crystal structures of three subunits of complex 1 (Cox1, Cox2 and Cox3 subunits, colored in black in figure 2.3 A) have been characterized from cow (Tsukihara et al. 1996) and one subunit of the Cytb protein has been characterized from yeast (Hunte et al. 2000). I extracted secondary structure information from their corresponding PDB files (respectively 1OCC and 1EZV) using PDB viewer (Guex and Peitsch, 1997) and PyMOLWin (www.pymol.org) and marked the presence of transmembrane helices on the protein alignments (Fig 2.4 C).

For all 13 proteins, I predicted the location of transmembrane helices using three different bioinformatic methods (TMHMM www.cbs.dtu.dk/service/TMHMM-2.0/, HMMTOP www.enzim.hu/hmmtop/ and Memsat http://saier-144-37.ucsd.edu/memsat.html, a tytpical output is in figure 2.4 B). I performed independent predictions on a protostome (the horshoe crab *Limulus Polyphemus*), and a deuterostome (the cow *Bos Taurus*): example results from two of these methods can be seen in the last rows of the alignment in figure 2.4 C. I compared in-silico prediction with information from crystallographic structure (where available) and carefully generated a consensus hydrophobic masking sequence (last line of the alignment in figure 2.4 C). I used the masking sequence to divide the original alignment (figure 2.4 C) into two parts: a hydrophobic partition - corresponding to the putative hydrophobic regions of transmembrane helices (figure 2.4 D) - and a hydrophilic partition corresponding to all other secondary structures, that I have found to be for the most part loops (figure 2.4 E). For each of the two partitions I estimated the corresponding replacement matrix, which I have called respectively MtPhobic (figure 2.4 F) and MtPhilic, (figure 2.4 G).

7.2.3 Estimation of empirical models using the GTR assumption and a ML approach

I used the maximum likelihood approach implemented in PAML (Yang, 2007) to estimate a general reversible amino acid replacement model, assuming reversibility, so that the rate matrix Q={q_{ij}} satisfies the General Time Reversible (GTR) condition π_j r_{ij} = π_i r_{ji} for all the amino acid pairs, where π_j is the stationary frequency of amino acid j and r_{ij} is the replacement rate between amino acids i and j. This makes the replacement matrix Q symmetrical and almost halves the number of free parameters of the model: 20 X 19 (replacement rates r_{ij}) / 2 (matrix is symmetrical) +19 (stationary frequencies π_j = 209). Rate heterogeneity across sites has been shaped by a Gamma distribution with four categories.

For the MtHydro model, I estimated 2 separate replacement matrices, named MtPhobic (figure 2.4F) and MtPhilic (figure 2.4G) using respectively the hydrophobic and the hydrophilic subsets of the original alignment (see above for details). The latter two matrices can be used together as a dual model of evolution called MtHydro.

7.2.4 Dataset used to test the fit to the models.

I analysed two datasets previously treated in the literature: a dataset of 23 arthropods (Rota-Stabelli and Telford, 2008) and a dataset of 41 mammals (Horner et al. 2007). I also constructed 4 mitochondrial protein datasets as follows: one containing 44 species from diverse metazoan groups, one with 24 lophotrochozoans, one with 30 ecdysozoans and one with 30 deuterostomes. I partitioned all the datasets into hydrophobic and hydrophilic subsets in accordance with the partitions used for the construction of my MtHydro model. I modeled the hydrophobic and hydrophilic partitions using respectively MtPhobic and MtPhilic (the MtHydro model), the transmembrane based JTT matrix (Jones et al. 1992) and the globular WAG matrix (Whelan and Goldman, 2001) and using two separate mechanistic GTR models, allowing the replacement matrix for the two partitions to be directly estimated from the data.

All the datasets have been analysed in Bayesian framework using MrBayes3.1 (Huelsenbeck and Ronquist, 2001). I recompiled MrBayes substituting existing matrices with the two matrices of MtHydro (MtPhobic and MtPhylic) and MtZoa models. For comparison reasons I also included MtArt model (Abascal et al. 2007). I ran tree searches on the 6 different metazoan datasets of concatenated mitochondrial proteins under these and other models of evolution using both the original and a partitioned dataset.

For all of runs, I modelled among-site rate heterogeneity with an invariable plus gamma (4 categories) distribution and ran two independent Bayesian tree searches with 4 MCMC chains. I ran the analyses until long after the likelihood of the sampled trees reached a plateau and the standard deviation of split frequencies reached 0.01. In some of the analyses using GTR models, the two runs did not satisfactorily converge even after 2 million generations, but the mean of the LnL distribution of trees sampled in the two distinct runs was similar and, as I was interested in the LnLs more than in the tree topology or other parameters, I stopped the MCMC chains.

7.2.5 Methods to compare replacement empirical matrices

In order to highlight the differences in replacement rates and amino acid frequencies between MtZoa and previous matrices, I generated a subtraction matrix, whose values correspond to the differences in replacement rate (r_{ij}) between MtZoa and MtREV (figure 2.2A upper) and between MtZoa and MtArt (figure 2.2B upper). I also calculated differences in the stationary frequencies (π_j) between the two pairs (lower parts of figure 2.2).

For the comparison of the two sub-matrices of MtHydro, I generated a subtraction matrix, whose values correspond to the normalized differences between replacement rates (r_{ij}) of the two sub-matrices (figure 2.5). I normalized differences according to (Le and Gascuel, 2008) as (X r_{ij} -Y r_{ij})/(X r_{ij} +Y r_{ij}), where X and Y are respectively MtPhobic and MtHydro and r_{ij} are the replacement rates of matrix R.

I used the AIS algorithm of Kosiol and colleagues (Kosiol et al. 2004) to highlight differences between different empirical models (figure 2.6). AIS identifies groups of amino acids with a high probability of change among those of the same group and low probability of exchange with those of other groups. AIS uses eigenvectors (one of the two forms of the spectral decomposition of the matrices) from Q to optimize the amino acid grouping on the basis of the conductance, which is a measure of changes from a group of amino acids to an other in a Markov process (Lio and Goldman 1998). Eigenvectors of different instantaneous rate matrices (MtREV, MtZoa, MtArt, MtPhobic and MtPhilic) were kindly provided by Caroline Kosiol and used as input for the AIS program, together with the stationary frequencies π_j and the matrix R (containing replacement rates r_{ij}).

7.2.6 Test of model fit

I evaluated model fit to the data using the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) defined as follow:

AIC = -2 log-likelihood + 2 K; BIC = -2 Log-likelihood + 2K log N, where K is the number of free parameters in the model and N is the number of sites in the alignment (Huelsenbeck et al. 2004).

The numbers of free parameters used in the AIC and BIC were determined as the number of branches to be estimated plus the number of free parameters in the model. For example I counted 2 parameters for the homogenous empirical models (proportion of invariable sites and the gamma distribution) and 210 for the mechanistic GTR (208 for the replacement matrix plus 2).

The log-likelihood (LnL) value used corresponds to the harmonic mean of the LnLs of trees sampled during a tree search using MrBayes (Huelsenbeck and Ronquist, 2001). For all these Bayesian analyses, I modelled among site rate heterogeneity with an invariable plus gamma distribution with 4 rate categories and ran two separate Bayesian tree searches long after the likelihood of the sampled trees had plateaued. While the likelihood associated with empirical models converged between the two runs after few hundred generations (I have run them for a minimum of 300.000), the mechanistic GTR model required up to two million generations, depending on the dataset.

I estimated the harmonic mean with Tracer (http://beast.bio.ed.ac.uk) using the log-likelihood of the trees sampled after burn-in. As I ran two independent tree searches, I calculated the harmonic mean on the combined LnLs, using Tracer and smoothing with 100 replicates (http://beast.bio.ed.ac.uk/). In a few cases the mean log-likelihood of the two Bayesian runs were slightly different; in this case I kept the highest, in order to be more conservative for the test of model fit. As a general rule I burn-in the first 20% of the sampled trees, but in the cases of GTR analyses the burn-in was set to 50%, after inspecting the LnL distribution. In a few GTR analyses, the LnL of the two runs plateaued at slightly different values (differences up to 50 LnL units) and I calculated the harmonic mean on the run with the highest mean of LnL distribution. While the LnL associated with empirical models plateaued after few hundred generations (I have run them for a minimum of 300.000) mechanistic models required up to two million generations, depending on the dataset. A recent study (Carapelli et al. 2007) shown that the LnL associated with different Bayesian tree searches (from the same dataset and model) may plateau at different values even if the topology of the consensus trees of different runs is almost the same and the standard deviation of split frequency thereby very low. This may imply a possible inaccuracy of the LnLs I have recorded. However, I were reassured by the fact that plateaued LnLs of different models commonly differed in order ranging from 100 to 1000.

7.3 The effect of outgroup choice (methods of chapter 3)

7.3.1 Dataset extraction and preparation.

I downloaded the 13 protein coding gene sequences from the mtDNA genomes of 102 arthropods and 38 metazoan outgroup taxa available on Genbank. I aligned amino acid sequences of each of the 13 mtDNA gene with MUSCLE and then back-aligned the corresponding nucleotide sequences (using TranslatorX - available on request). AGG codons were recoded as NNN because of evidence of parallel evolution of a new variant of the invertebrate mitochondrial genetic code in unrelated arthropods groups (Abascal, 2006(). The alignment was refined by hand and poorly conserved or ambiguously aligned codons were excluded from further analyses. 3rd positions of codons were excluded from the nucleotide dataset as they are prone to saturation in arthropod mtDNA (Rota-Stabelli et al. sadly unpublished data). The final nucleotide alignment (1st and 2nd codon positions) contained 5588 nucleotides corresponding to 2794 codons. I chose outgroups with the invertebrate mitochondrial genetic code (code 5 in NCBI) from the phyla Annelida, Mollusca, Echiura, Brachiopoda, Chaetognatha, Nematoda and Cephalochordata. I also included the recently sequenced priapulid (Webster et al. 2006, Webster et al. 2007) and onychophoran (Podsiadlowski et al. 2007) mtDNAs. Arthropod species were chosen in order to represent major arthropod clades equally and to limit the effects of over-sampled groups such as ticks and flies that are extremely A+T rich and may seriously interfere in the stationarity of the dataset. I used a consideration of compositional characters for all the available arthropods species to guide my selection of a well balanced arthropod dataset including 4 myriapods, 3 chelicerates, 5 crustaceans 7 insects and 2 non-insect hexapods. I excluded species showing exaggerated low GC content and/or abnormal GC skew values (e.g. scorpions, many ticks, the crustaceans *Tigriopus* and *Hutchinsoniella*, the hemipteran insects). I used this dataset of 21 arthropods to test how selected outgroups might affect the ingroup topology. When using a number of different outgroups, the pancrustaceans *Speleonectes, Pollicipes* and *Gomphiocephalus* were found to be prone to LBA tending to branch at the base of the arthropods, closer to the outgroup, rather than within Pancrustacea.

7.3.2 A multi criterion table for the selection of outgroups

In order to choose optimal outgroups I constructed a table (Table 3.1) to be used as a decision maker for the selection of a set of adequate outgroups. I considered a variety of evolutionary characters, such as genetic distances and compositional qualities (see table legend for details). *Limulus* was chosen as a representative arthropod for genetic distance calculations because of its moderate branch length and average composition values. However, distances calculated from other arthropod species or using the mean ML distance to all arthropods gave essentially identical results. The compositional indicators G+C content, GC skew and skew index were calculated on first and second codon positions only. Usually these characteristics are analyzed at the third codon position, which is less constrained and reflects more directly the compositional tendencies of the mtDNA but I have excluded 3rd positions from phylogenetic analyses and wanted to focus my attention only on the effect that composition may have on phylogenetically informative sites. More in detail, nnucleotide content has been calulated as the percentage of G+C and amino acid content as the percentage of amino acid, whose codons are rich in G and C (amino acid G, A, R, P). I also calculated GC Sskew for each gene independently and plotted values in the skew plots of figure 3.2, using the ancient arthropod gene order (AAGO) as a reference. Both in figure 3.1 and 3..2 I used averaged values calculated over the sampled arthopods; for more information on this averaging see section 7.4.3 of this chapter.

I selected at least one and a maximum of 4 species from each outgroup clade on the basis of their compositional similarity to the average for arthropods and with the lowest genetic ML distance. When a selection between alternative taxa was ambiguous I gave precedence to small genetic distance over compositional similarity, however in most cases the two measures coincided (see value in bold in columns GC%, Skew I and Dist of table 4.1). I used selected outgroups independently to root the 21 arthropod dataset (which includes long branched pancrustaceans) in order to detect if a given outgroup was able, used alone, to root the arthropod tree in a credible position, avoiding attraction of unrelated species or making clearly monophyletic groups diphyletic. I used information from these preliminary tree searches as an additional indication of outgroup adequateness.

The table used for the outgroup selection also contains a new value, named “skew index”, that describes how much the overall GC skew values calculated for each gene independently (columns cox1, cox2…nd6 in table1) differ from the mean skew of the arthropods, calculated using the initially selected set of 21 arthropods (which share a similar GC Sskew values for each of the genes).

7.3.3 Phylogenetic analyses

I performed tree searches with MrBayes 3.1.1 (Ronquist and Huelsenbeck 2003) using different sets of outgroups and different character states (nucleotides and amino acids). The nucleotide dataset was partitioned into first and second codon positions and each partition independently modelled under a GTR model with invariable + 8 gamma distribution. The amino acid datasets were analyzed with invariable + 4 gamma distribution using the new amino acid substitution matrix MtZoa (see chapter 3.1, Bourlat et al. 2006, Rota-Stabelli, Yang and Telford 2009). I ran the mcmc for between 250,000 and 1,000,000 generations and discarded trees considerably after the likelihoods had plateaued (as inspected by plotting the logL of the sampled trees against the number of generations). I also performed non-parametric bootstrap analyses of the nucleotide dataset using Treefinder (Jobb et al. 2004) using the same model as in the MCMC analysis, but with no codon partition, and generating 100 replicates.

7.4 Mitogenomic analysis of the Ecdysozoa (methods of chapter 4)

7.4.1 Genome sequencing, annotation and tRNAs inferences

The complete mitochondrial genomes of the onychophorans *Epiperipatus biolleyi* and *Peripatoides* sp. and the priapulid *Halicryptus spinulosus* have been amplified and sequenced as described in Lavrov et al. (2000) by Dennis Lavrov, Mark Blaxter and colleagues. I amplified partial sequences encompassing 5 protein coding genes from *Euperipatoides kanagrensis* using customly designed primers. The open reading frames in the newly sequenced genomes were annotated based on comparisons with protein sequences from closely related species. In addition, the mtDNA available from *Metaperipatus inae* (GenBank: EF624055) has been re-annotated by Dennis Lavrov lab members based on the two other onychophoran mitochondrial genomes. tRNA genes were inferred using the tRNAscan-SE and ARWEN programs (Lowe and Eddy, 1997; Laslett and Canbäck, 2008) and checked manually. tRNA genes not found by the computer programs were searched for manually based on expected anticodon sequences, conserved nucleotides, and potential secondary structures as well as by similarities with known sequences from closely related species when available. Where several potential tRNA gene sequences were found, they preferred the one with a more conserved gene order position.

7.4.2 Compositional analysis

For each species of my dataset, I calculated the nucleotide and the amino acid frequencies of the 13 concatenated coding genes using all three codon positions. Nucleotide content has been calculated as the percentage of G+C % and amino acid content as the percentage of amino acid, whose codons are rich in G and C (amino acid G, A, R, P) or in A and U (amino acid K, L, M, N , I , F, Y). I also calculated the GC skew (Perna and Kocher 1995) and the Skew index (chapter 3 and Rota-Stabelli and Telford 2008), which are two measures of strand asymmetry, on the whole concatenated alignment using all 3 codon positions. The skew index was calculated using the arthropods as a reference (see below for more details). As strand asymmetry affects genes differently, I calculated GC skew for each gene independently using all three codon positions. To test if the strand asymmetry of genes was at equilibrium or not, I calculated GC skew for the 1st+2nd and 3rd codon position separately. GC skew values have been plotted for species of interest, using the arthropod ancestral gene order (AAGO, which is the same of the ancestral ecdysozoan) as a reference to order genes on the abscissa of plots in figure 4.2. I summarize some of these statistics in Table 4.1.

7.4.3 Averaging characters

For comparative reasons, some of the statistics described above have been averaged over all arthropods and/or the four arthropod subphyla. This approach is complicated by the compositional characters not being conserved throughout the arthropods. For example, while most the arthropods are A+T rich, some lineages such as crustaceans are not. Additionally, the nucleotide composition of some species can be very different to that of the larger linage they belong to. This level of heterogeneity results in rather high values of standard deviation for the averaged data. It is possible, however, to find certain patterns in the arthropods, which justify the averaging: compared to their ecdysozoan outgroups, hexapods and chelicerates are, for example, clearly A+T rich, while crustaceans and myriapods are less. Furthermore, all the arthropods are AT enriched compared to the majority of non ecdysozoans; accordingly, as in the case of figure 3.1 (but not figure 4.2), I used an average calculated over the whole arthropods.

Strand asymmetrical properties vary as well in the arthopods. Using the fruitfly genome as a reference, in some species (various arachnids, most of the hemipterans, the cepaholcarid and the copepods for example), the GC skew is inverted for most or all of the genes (Jones et al 2007, Hassanin 2005). In other species, the skew profile of the fruitfly is exaggerated to extreme values as in *Armillifer armillatus* and *Limulus polyphemus*. The majority of arthropods, however, (those which share a similar ancestral gene order) are characterised by a skew pattern in which most of the genes posses a slightly negative GC skew and the genes ND5, ND4, NDL and ND1 posses a positive skew, reflecting the distribution of genes between the two strands. The average arthropod GC skew (used throughout all chapter 4 and for the estimate of the skew index) has been calculated only on arthropods that show the typical arthropod skew pattern.

7.4.4 Alignments and dataset preparation

I downloaded nucleotide sequences of the 13 mitochondrial coding genes for various metazoans species from the Ogre database (http://drake.physics.mcmaster.ca/ogre/compare.shtml) and added complete sequences for the Priapulid *Halicryptus spinulosus*, the two tardigrades *Hyspibius dujardani* and *Thulinia sp*., the onychoproans *Peripatoides sp*. and partial sequences from 5 genes of the onychophoran *Euperipatoides kanagrensis*, resulting in a dataset of 245 metazoan species. I translated nucleotide sequences into their corresponding amino acids, according to the taxonomically appropriate genetic code, and aligned the 13 protein sequences individually with ClustalW (Larkin et al 2007). I back aligned nucleotide sequences to the amino acid alignment and assembled a concatenated alignment of the 13 genes using TranslatorX (downloadable from http://web.mac.com/maxtelford/iWeb/Work/Downloads.html). In order to increase the accuracy of the aligning process, I further realigned each amino acid alignment independently using Muscle (Edgar 2004) followed by a second eye refinement and a successive reconcatenation.

In order to avoid misleading effects due to inadequate outgroup selection (in accordance with results of chapter 3), I compiled a table similar to table 3.1 (table not shown) containing various statistics for each of the 245 ingroup and outgroup species I sampled. I compared characters of this table to select lophotrochozoan and deuterostome outgroups that share the optimal compromise of minimal genetic distance and compositional characters which do not differ too much from the main ecdysozoan ones. The table contained (1) the ML distance to the Ecdysozoa (calculated as the averaged distance to three Ecdysozoans*, Priapulus caudatus, Limulus polyphemus* and *Tribolium castaneum),* (2) the G+C content, (3) the content of GC rich amino acids and two indicators of G/C strand asymmetry (4) GC skew and (5) the skew index (see chapter 4.2.3).

From the 245 taxa alignment, I selected a balanced sample of 66 species (table 4.1) of which 10 were outgroups. The nucleotide and the corresponding amino acid alignments have been processed independently with Gblocks (Castresana 2000) at default settings, follow by insensitive by-eye refinement, to remove poorly conserved regions resulting in datasets of 2016 amino acids and 7482 corresponding nucleotides (note that the length of the two datasets are not consistent to each other because the datasets have been processed separately). To test the affinities of Nematoda, I assembled two extra datasets, based on the 66 taxa alignment containing additional nematodes, in particular slow evolving Enoplea, resulting in two datasets of 88 taxa and 2016 residues and 59 taxa and 2946 resides.

7.4.5 Phylogenetic analyses

I analyzed the 66 taxon dataset using a variety of evolutionary models and phylogenetic tools. I used both nucleotides and corresponding amino acid sequences although most analyses were carried out using amino acids.

The nucleotide alignment was analysed under both a Bayesian and a Maximum likelihood approach using MrBayes and RAxML (Stamatakis 2004) respectively. In both cases I excluded the 3rd codon positions and modeled the 1st and 2nd codon positions separately using two GTR models and gamma distributions with 5 categories (Lanave et al 1984). For the RAxML analysis I used the fast maximum likelihood method and performed a non parametric bootstrapped analysis on 100 pseudo replicates, with bootstrap support reported on the maximum likelihood tree inferred from the whole dataset. The nucleotide dataset was also analysed using the NTE model of Hassanin (2005) with all 3 codon positions recoded accordingly to NTE using the program Recoder from Stuart Longhorn (Masta et al. 2009, http://web.pdx.edu/~stul/Software.html). The NTE dataset was analysed using MrBayes, with the 1st and 2nd codon positions modeled by two distinct GTR models and the 3rd position modeled by a 2 character state model. While two independent runs in the Bayesian analyses using 1st plus 2nd codon positions satisfactorily converged according to the MrBayes manual, the two independent runs using the NTE recoding model did not converge and supported extremely different tree topologies. One run supported a sister relationship between tardigrades and mollusks, while the second run supported monophyly of ecdysozoans. The associated mean log-likelihood of trees was significantly lower in the first run and had not satisfactorily plateaued and I therefore calculated the consensus tree using trees only from the second run.

The amino acid dataset has been analysed more extensively, using homogeneous and heterogeneous model of sequence evolution under both Bayesian and Maximum likelihood frameworks. Initially, I performed a cross-validation analysis to test the fit of different amino acid evolutionary models to my dataset, using PhyloBayes and following the protocol described in the manual. I used the MtREV mitochondrial model (Adachi and Hasegawa 1996) as a reference to test the fit of other models: the CAT model (Lartillot and Philippe 2004), the mechanistic GTR model (Lanave et al 1984, Yang, Nielsen and Hasegawa 1998), MtZoa (which is presented in chapter 2 of this thesis, Rota-Stabelli et al. 2009) and MtArt (Abascal et al. 2007) models, which I have implemented in PhyloBayes. Using the MtREV model as a reference, results of the crossvalidation are as follow: ART versus REV : 80.1 +/- 25.8; GTR versus REV : 85.4925 +/- 25.3; mtZOA versus REV : 91.46 +/- 21.2; CAT versus REV : 169.242 +/- 18.8. Bearing in mind that positive values mean a better fit to the dataset, results clearly show that the heterogeneous CAT model is the model that best fits my 66 taxon dataset. Interestingly, the second best model is MtZoa, which fits the dataset even better than the mechanistic GTR model and which has been shown to fit respectively ecdysozoan and metazoan mitochondrial datasets better than other models (Rota-Stabelli et al 2009). Following these results, I chose CAT and MtZoa models for further analyses and used the other models for comparative reasons only.

Bootstrapped (100 replicates) maximum likelihood analyses have been carried out with the fast maximum likelihood method implemented in RAxML using the MtREV and the MtZoa model (customly implemented) and a 4 categories gamma distribution. Bayesian analyses have been carried out using both MrBayes and PhyloBayes. In both cases we described the among site rate variation with a gamma distribution using 4 categories. I run two independent tree searches and stopped them after the likelihood of the sampled trees had significantly plateaued and the two runs had satisfactorily converged (sd of split frequency lower than 0.02 in MrBayes and maxdiff less than 0.2, but in most of the cases less than 0.01 in PhyloBayes). Analyses using CAT, GTR, MtREV, MtART and MtZoa models have been done with PhyloBayes, analyses using MtHydro (see below for more details) with MrBayes.

I also performed Bayesian analyses using the CAT-BP model implemented in NH-PhyloBayes. CAT-BP accounts for among site heterogeneity and also allows stationary frequencies to vary among branches (Lartillot and Philippe 2004; Blanquart and Lartillot 2007). The number of CAT categories in NH-PhyloBayes was set to a value ranging between 120 and 140, as learned from the corresponding PhyloBayes analyses. I ran a minimum of two separate analyses, but it was impossible to obtain a meaningful convergence even after millions of generations and multiple runs. I therefore sampled trees from each run independently and compared the results of independent runs. I also used the partitioned heterogeneous model MtHydro described in chapter 2.2, which I have implemented in MrBayes. MtHydro is based on a pre-partition of the mitochondrial protein alignment into two sub-alignments: a hydrophobic and a hydrophilic one, which are modeled by two separate empirical sub-models (Bourlat et al. 2009, Rota Stabelli, Horner and Telford, unpublished).

I finally analysed the amino acid dataset after removal of proline and glycine, which are respectively coded by codons CCN and GGN and whose frequencies are expected to be particularly influenced by strand asymmetries (GC skew). For this analysis I used PhyloBayes and the MtZoa model.

7.4.6 Sequential taxa and site removal

In order to explore the signal concerning the affinity of Tardigrada, I removed fast evolving species which show a dubious relationship with the Tardigrada from the 66 taxa alignments. I sequentially removed the two Pycnogonida (dataset of 64 species), the two Symphila (62 species) and the outgroup sequences plus fast evolving arachnids (46 species) and inferred phylogeny from those datasets using PhyloBayes and Raxml and modeling the evolutionary process with the CAT and the MtZoa models respectivley. Results of this analyses are summarised in figures 4.5 and 4.6.

I further explored the signal in sequences by sequentially removing classes of fast and slow evolving sites. I used PAUP to calculate parsimony scores (p-score) at sites, using seven monophyletic groups (Echinodermata, Lophotrochozoa, Aranea, Acari, Myriapoda, Hexapoda and Crustacea). For each site of the alignment, I summed the p-scores of each monophyletic group and sorted sites on the basis of their total p-score, obtaining 34 classes of sites, where 0 correspond to zero changes observed among all the monophyletic groups in the most parsimonious tree and 34 corresponding to 34 observed changes. I generated 9 alignments, whose length ranged from 2014 to 249 amino acids, sequentially removing either the fastest or the slowest evolving sites using percentiles of the frequencies of the 34 classes to guide the construction of the alignments.

I also used quartiles to divide the 34 classes of sites in two groups: one containing sites which fall into the internal quartiles of a quartile distribution and the other containing sites which fall into the external quartiles. The internal quartiles are characterised by moderately evolving sites, thus being homogenous in term of their rate and possibly more adequate markers for the inference of phylogeny. External quartiles are characterised by either fast or slow evolutionary rates, thus being heterogeneous and the fast evolving ones likely to contain homoplastic characters responsible for misleading phylogenetic signal. The nine alignments, plus the two "quartiles" alignments have been analyzed using either CAT or the MtZoa model.

7.5 Reanalysis of published molecular datasets (methods of chapter 5)

I have reanalysed five published molecular datasets: the ribosomal dataset of Mallat and Giribet (2006), the nuclear dataset of Regier et al. (2005), the mitochondrial dataset used in chapter 3 (Rota Stabelli and Telford 2008), the combined dataset of Bourlat et al. (2008) and the phylogenomic dataset of Dunn et al. (2008). These datasets have been chosen as effective representatives of the molecular markers currently used to study the arthropods. I allowed four classes of phylogenetic and analytical parameters to vary: (1) the methods of phylogenetic inference, (2) the model of evolution employed, (3) the taxonomic sampling and (4) the selection of sites. Given the different nature of the various datasets, I provide, for each of the them, a description of the analyses carried out (and the results obtained).

7.5.1 *Nuclear dataset* (Regier et al. 2005)

The nuclear dataset, downloaded from http://www.umbi.umd.edu/users/jcrlab, is composed of three nuclear coding genes: elongation factor-1a (1131 nucleotides long), the largest subunit of RNA polymerase II (2025 nucleotides) and the elongation factor-2 (2178 nucleotides) for a total of 5334 nucleotide positions for 62 taxa. The main analysis of Regier has been carried out at the amino acid level using a Maximum Lihlehood (ML) approach and the WAG model (Regier et al 2005). Under this settings, the most favoured topology support a group of chelicerates plus tetraconates (named by me "Crazypoda"). I reanalysed the dataset at the nucleotide level using two distinct GTR+G models for modelling separately the 1st and 2nd codon positions (excluding the 3rd). Under this conditions, I recover Mandibulata with posterior probability (PP) 100 using MrBayes and an unresolved topology using the WAG+G model under a ML framework implemented in Treefinder. When using a different model of evolution, I recover Mandibulata (PP 72, using amino acids and the CAT+G model implemented in PhyloBayes) or the Crazypoda topology (PP 78 using amino acids and the GTR+G). I also recover discordant topologies using different outgroups sampling in a Bayesian framework: using only Onychophora outgroups, the tree supports Mandibulata (PP 96) when analysed as nucleotides and Crazypoda (PP 100) when analysed as amino acids; using only tardigrade outgroups, I obtained support for Myriochelata (PP 100) using nucleotides or Crazypoda (PP 87) using amino acids. I finally analysed each of the 3 genes of the superalignment independently: while elongation factor-2 support Crazypoda (PP 100 using amino acids and PP 74 using nucleotides), RNA polymerase II and elongation factor-1a were not able to resolve the affinity of the myriapods.

7.5.2 Mitochondrial dataset (Rota Stabelli and Telford 2008)

The mitochondrial dataset comprises 2787 amino acid for 21 arthropods ingroup plus various outgroup sequences (see chapter 7.3 1 for more details). This dataset support Myriochelata when the full set of outgroups is used, but the topology significantly change when subsets of outgroups are used; as fully addressed in chapter 3, when a set of "optimal" outgroups is used, dataset support Mandibulata, at least in a Bayesian framework. The "optimal" dataset analysed using the CAT model for amino acids, which has not be employed in chapter 3, support paraphyletic myriapods with the diplopods sister to chelicerates and the chilopods sister to tetraconates. Using the 2 state model HKY85, which account only for differences between transitions and transversions, the most favoured topology is Myriochelata (PP 100). The signal supporting Mandibulata is found in subunits of complex 1 (Nadh genes, PP 100 and 92 using respectively amino acids and nucleotides), while other genes support Myriochelata (PP 85 and 66 using respectively amino acids and nucleotides).

7.5.3 Ribosomal dataset (Mallat and Giribet 2006).

The ribosomal dataset of Mallat and Giribet (2006) was downloaded from http://www.wsu.edu/~jmallatt/alignments.html. It is composed of the concatenation of the small (18S or SSU) and large (28S or LSU) ribosomal subunits. The alignment used for the majority of the analyses in the original publication (Mallat and Giribet 2006) contains 3852 well aligned, conserved sites for 84 taxa and supported Myriochelta. In collaboration with Andrew Economou, "sequences were aligned to include secondary structural information. 28S and 18S rRNA sequences aligned according to their secondary structure were downloaded from the European Ribosomal RNA database (http://bioinformatics.psb.ugent.be/webtools/rRNA) in the dedicated comparative sequence editor (DCSE) format. These were converted into nexus format using the Ystem software (Telford et al. 2005) and used as a template for the alignment of the Mallat rRNA sequences, using the profile alignment mode in ClustalX. For 28S, the sequences were aligned to the five ecdysozoan taxa present in the original DCSE file, and for 18S, the eleven arthropod taxa were used. The Xstem and Ystem software (Telford, et al. 2005) were used to convert the secondary structure information in the DCSE files into a form that could be used by phylogeny software such as MrBayes. The quorum values for Ystem were set so that for a site to be annotated as a stem site, it had to be present in 3/4 of the annotated taxa” (modified from Andrew Economou 2008). I used a more stringent selection of conserved sites, which resulted in 3736 positions. As for the method of inference, the dataset has been analysed under both a maximum lihlehood framework using PAUP (Swofford 2002) and a Bayesian framework using MrBayes (Huelsenbeck and Ronquist 2001). I reanalysed the dataset using two distinct model of evolution: the GTR+I+G homogenous model and modelled all sites as same (in PAUP, similarly to original analysis of Mallat) and the stem-loop/doublet model using MrBayes (as described in Telford, Gowri-Shankar and Wise 2005). The latter model allow sites which correspond to loops to be modelled by a normal GTR+G model and sites which correspond to the stems to be modelled by a doublet GTR+G model, which take into account for coevolution of pairs of sites. In all cases the phylogenetic trees were in accordance with the original tree of Mallat and Giribet (2006) in supporting Myriochelata, except for the ML tree using PAUP which was unresolved. The effect of taxon sampling (3rd class of parameters in figure 5.1) has been explored by excluding some of the outgroup sequences and a stem/loop Bayesian inference of phylogeny: when excluding fats evolving onychophorans and nematodes the tree was mainly unresolved. Finally, as discussed in the original publication (Mallat and Giribet 2006) use of only SSU resulted in support for Mandibulata, while LSU alone strongly support Myriochelata.

7.5.4 Combined dataset (Bourlat et al. 2008)

The dataset has been provided by Sarah Bourlat and consists of the concatenations of different molecular markers: 8 nuclear coding genes, 12 mitochondrial coding genes and large + small ribosomal subunits, for a total of 8664 mixed characters form 37 metazoan taxa (Bourlat et al. 2008). As the signal in ribosomal and mitochondrial markers has been already analysed in the previous paragraphs, I concentrated my reanalyses on the nuclear coding genes of Bourlat, which mostly differs from those used by Regier et al. (2005). After exclusion of poorly aligned sites the nuclear coded amino acid dataset resulted in 2657 amino acids. The full concatenated dataset of Bourlat supports Mandibulata (PP 100), as well as only the nuclear coded genes analysed using a WAG+G model (PP 100). However, when analysed using the CAT+G model, nuclear genes resulted in an unresolved myriapods relationship. Furthermore, when nuclear genes are analysed at the nucleotide level, they slightly support Myriochelata (PP 60). I have then tested the use of different outgroup sequence (only deuterostomes, only lophotrochozoans, only non bilaterians and slow evolving outgroups): all analyses consistently supported Mandibulata .

7.5.5 Phylogenomic dataset (Dunn et al. 2008)

The dataset has been provided by Casey Dunn and consist of 21150 amino acid position from the concatenation of 150 genes from 77 metazoan taxa. I reduced the taxon sampling to the 9 sampled arthropods plus some ecdysozoan and lophotrochozoans as outgroup sequences. As already shown by Dunn et al. (2008), analyses using both a Bayesian/CAT+G and a ML/WAG+G approach resulted in support for Myriochelata (PP >95, BS>90). I tested the use of different outgroups to root the arthropods (only ecdysozoans, only lophotrochozoans and fast evolving outgroups) and all consistently supported Myriochelata with high support. I tested the effect of using different site selection by sequentially removing fast evolving sites (slow-fast method) using the program SLOWFASTER at default settings (Kostka et al. 2008). I generated nine sub-datasets, which consistently supported Myriochelata with high PP, until the signal in sequences decays as reported by monophyly of tetraconates.

7.5.6 Tests of competing hypotheses

Statistical tests of the robustness of tree topologies have been evaluated with two bootstrapped based likelihood tests: the Approximately Unbaised (AU) and the Shimodaira –Hasegawa (SH) test (Shimodaira and Hasegawa 2001) and were performed on each dataset in the form in which it has been published with the exception of the combined dataset of Bourlat et al. (2008) and the phylogenomic dataset of Dunn et al. (2008) which have been analysed in a reduced taxonomic version. AU and SH tests were performed with CONSEL (Shimodaira and Hasegawa 2001) based on site wise likelihood values calculated by PAML (Yang 1997), using the same model and method of inference as in corresponding publication. Four categories of the gamma distribution have been used to model rate heterogeneity and stationary frequencies have been inferred directly from the datasets.

7.6 Phylogenomic analyses (methods of chapter 6)

7.6.1 EST Sequencing and Data Assembly.

Approximately 5000 ESTs of the myriapod *Strigamia maritima* have been sequenced by Macrogen, from a cDNA library provided by Michael Akam and Ariel Chipman (Chipman Arthur and Akam 2004). The ESTs are publicly available in dbEST/GenBank (http://www.ncbi. nlm.nih.gov/dbEST/). The data assembling has been carried out by my collaboration Herve Philippe and Henner Brinkmann from the Montreal University . They assembled a phylogenomic dataset of 59 taxa, consisting of 48 Ecdysozoa and 11 outgroups from within the Lophotrochozoa. The dataset has been built by merging orthologs from two previously assembled datasets (Philippe et al. 2009, Dunn et al. 2008), via the protocol described in (Bapteste et al 2002), and adding orthologs found amongst the the 5000 *Strigamia maritima* sequences as well as from other recently sequenced ESTs available for various Ecdysozoa species in the Trace Archive and the dbEST (**http://www.ncbi.nlm.nih.gov/Traces**, **http://www.ncbi.nlm.nih.gov/dbEST/**). They evaluated orthology by inferring single-gene phylogenies and looking for conflict with the super-matrix tree according to the protocol described in Rodriguez-Ezpeleta et al. (2007) and rejecting orthology below a bootstrap threshold of 70%. Sequence selection and concatenation were performed with SCaFoS (Roure, Rodriguez-Ezpeleta and Philippe 2007). In order to increase the gene sampling and to reduce the effects of missing data, we decided to include only genes sampled for at least two-thirds of the species and species whose genes, covered at least 15% of the total alignment. For the same reason 26 out of the 59 taxa were composed of chimeric sequences produced by merging two or more species belonging to non-controversial clades (see table for details). In most cases we merged species of the same genus and in a few cases of the same (super)family or (infra)order. Only in the case of Onychophora did they merge two distantly related species as a consequence of the Peripatidae species being poorly sampled. They assembled a total of 198 genes, of which 61 were exclusive to the dataset of Dunn and colleagues (Dunn et al. 2008), 126 exclusive to the dataset of Lartillot and Philippe (2008) and 44 shared by the two datasets. The final dataset displays 29% of missing data.

7.6.2 Alignments preparation and Phylogenetic analyses.

For computationally demanding analyses, I selected a balanced alignment of 30 species from my original 59. I excluded species with poor gene sampling where possible. This dataset contains 4 lophotrochozoan outgroups, 2 scalidophorans, 6 nematodes, 2 tardigrades, 1 onychophoran and 15 arthropods (2 myriapods, 4 chelicerates, 4 crustaceans and 5 hexapods) and 40,100 unambiguously aligned positions.. For analyses using the CAT model, the dataset was processed to exclude constant sites (those conserved throughout all species and less likely to carry phylogenetic signal; they can only influence on the shape of the distribution of rate across sites, which is circumvented here using a non-parametric approach – see below), resulting in 24345 amino acid positions. I also constructed a new dataset using my 30 taxa sampling (including *Strigamia*) and genes used by Dunn and colleagues (2008). The resulting alignment contained 18829 positions, fewer than the 21152 positions in the original study because the new sampling covers only part of their original large metazoan sampling and I used a more stringent criterion to select unambiguously aligned positions.

Bayesian analyses were done with PhyloBayes for CAT, CAT-Covarion and CAT+GTR models (Lartillot and Philippe 2004, Zhou et al. 2007) with a posterior consensus tree obtained by pooling the tree lists of two independent runs. I generated consensus trees from the Bayesian analyses of 100 bootstrapped pseudoreplicates using PhyloBayes (CAT and CAT+GTR models) and the fast maximum likelihood search (WAGGAMMAF and GTRGAMMA models) implemented in RAxML (Lartillot and Philippe 2004, Stamatakis 2006, Whelan and Goldman 2001, Yang, Nielsen and Hasegawa 1998). In all the analyses stationary frequencies have been estimated from the datasets and a discrete Gamma distribution with four categories has been used for modeling rates across sites, except for the CAT model where rate heterogeneity was modeled using a Dirichlet process (Huelsenbeck and Suchard 2007). In all PhyloBayes analyses, the posterior consensus tree was obtained by pooling the tree lists of two independent runs, stopped when the observed larger discrepancy across bipartitions (maxdiff) was less than 0.2, and discarding a sufficient number of initial sampled trees in order to minimise the maxdiff.

I explored the effect of taxon sampling on both my dataset (figure 6.2) and that of Dunn et al. (2008) (data not shown, but see figure 6.3). I generated four sub-datasets containing all the arthropods plus different outgroups to them: (i) all outgroups, (ii) only distant Lophotrochozoa (iii) only fast evolving nematodes and finally (iv) short branch Ecdysozoa (Onychophora, Priapulida and Kinorhyncha). I analysed these four sub-datasets using Bayesian (CAT model) and bootstrapped Maximum Likelihood (WAGGAMMAF and GTRGAMMA models).

I (re) analysed the original 150 genes dataset of Dunn et al. (2008) using a sub-set of their 77 metazoan dataset to contain the 16 ecdysozoans they sampled plus 6 slow evolving lophotrochozoan outgroups. The number of positions sampled in this reduced dataset was 20079 (original Dunn is 21152), as some positions were missing in the new alignment and/or a few others were poorly represented (due to reduced sampling).

The updated Dunn dataset and my 30 taxa, 24345 amino acid dataset have been used to explore the effect of fast evolving sites removal. For each of six putative monophyletic groups – Myriapoda/Chilopoda, Chelicerata, Tetraconata, Tardigrada, Nematoda and outgroups - I calculated the sitewise parsimony scores using PAUP (Swofford 2002) and used their sum to define classes of sites. I generated 10 alignments by sequentially removing classes of fast evolving sites from the original alignment and stopped generating sub-alignments when 75% of the sites had been removed. The slow-fast sub-alignments were analysed with RaxML using the fast maximum likelihood search and the WAGGAMMAF model.

Saturation analysis was carried out using PAUP on my 30 taxon set and the updated Dunn dataset. Observed pairwise differences have been plotted against pairwise substitutions according to a WAG (GAMMA+F) model of evolution to check for the level of saturation. In the absence of mutational saturation the coefficient of the regression line should be 1; the lower the coefficient, the higher is the level of saturation in the dataset.

7.6.3 Phylogenomic analyses of Tetraconata (methods of chapter 6.2)

Data assembly

I assembled two distinct phylogenomic datasets using ESTs available in the Trace Archive and the dbEST (http://www.ncbi.nlm.nih.gov/Traces, http://www.ncbi.nlm.nih.gov/dbEST/). The two datasets have been centred respectively on insects and on crustaceans in order to focus the sampling on a specific taxonomic level and to minimize the proportion of missing data. Orthology of genes and concatenation has been carried as in section 7.6.1, in collaboration with Herve Philippe and Henner Brinkmann. In order to increase gene sampling and to reduce the effects of missing data, I only included genes sampled in at least 71% and 76% of the species for respectively the insect and the crustacean dataset. For the same reason I generated chimera sequences for various species belonging to recognised monophyletic clade. I typically merged species belonging to the same genus or family, except for the case of the Dictyoptera, where I sampled two distinct orders, Blattaria and Isoptera.

The insect dataset comprises 201 genes and 45353 amino acid positions from 51 Tetraconata taxa of which 41 are hexapods. None of the genes is missing in more than 15 taxa, resulting in 25% missing data. I selected insect species in order to avoid over representation of some lineages such as Diptera, Coleoptera and Lepidoptera and used the crustacean taxa as outgroups. Species/chimeras have been chosen on the basis of a compromise between minimising missing data and emphasizing taxonomic diversity taxonomically .

The crustacean dataset contains 149 genes (none is missing in more than 10 species) and 33,833 positions (27% missing data). I sampled 41 species of which 11 are crustaceans sampled from 3 classes, Malacostraca, Branchiopoda and Copepoda (former Maxillopoda) and the remaining hexapods, myriapods and chelicerates (the latter two clades have been used as outgroups).

Phylogenetic analyses

Phylogenetic analyses have been carried out on both the “insect” and the “crustacean” dataset using two distinct approaches: (i) the fast maximum likelihood method implemented in RAxML using the recent LG model (Stamatakis 2004, Lee and Gascuel 2008) and (ii) the Bayesian approach implemented in PhyloBayes using the CAT model (Lartillot and Philippe 2004). For each model and method I performed a non parametric bootstrap analysis based on 100 pseudo-replicates and calculated the consensus tree. In the LG analyses, stationary frequencies have been estimated from the datasets and in the CAT analyses, these were modeled using a Dirichlet process. For all the analyses I modeled across rate variation using a discrete Gamma distribution with four categories. The consensus of each of the CAT analysis was derived from the pooling of two independent runs, which had satisfactorily converged. In order to further explore the phylogenetic incongruence from the analyses of the crustacean dataset, I have additionally analysed it using (i) the CAT-covarion model which allows site rates of evolution to vary among the trees (Zhou et al 2007) and (ii) the CAT-GTR model which uses the stationary frequencies of the CAT model categories and the replacement probabilities of the homogenous GTR model.

Chapter 8
Discussion and perspectives

8.1 The actual novelty of my new models of evolution: are they genuine improvements?

A theory has only the alternative of being wrong. A model has a third possibility, it might be right but irrelevant.
Manfred Eigen

In chapter 2, I presented two new models of amino acid evolution estimated from and intended for mitochondrial proteins: MtZoa and MtHydro.

MtZoa is a simple GTR empirical matrix and its innovation is the large taxonomic sampling it has been estimated from. According to tests of model fit to various datasets, MtZoa should be used for the analysis of deuterostome and lophotrochozoan datasets and for datasets containing diverse or basal metazoan groups. In the case of poor phylogenetic signal or problematic nodes, the use of a more appropriate model such as MtZoa results in a better fit to the dataset and may lessen possible systematic biases. In the light of this MtZoa is an effective and useful advance.

However, I show that MtZoa only modestly differs to MtArt in terms of replacement rates and stationary frequencies, suggesting that MtArt, although derived from an arthropod dataset, describes the evolutionary pattern of the whole metazoans to a reasonable extent. There are some key differences between the two of a subtle compositional nature, but quantitatively speaking these differrences are not as significant as those between MtZoa and MtREV, which has been the reference model for the last fifteen years. I was able to find an example in which MtZoa is able to recover a tree topology more in line with the morphological point of view than MtART

(Chapter 2.1). My final and impartial consideration is that, strictly speaking, the big jump of quality over MtREV has been done by MtArt, making MtZoa a useful, even if not an essential improvement.

MtHydro, on the other hand, addresses an important aspect of protein evolution: the among sites heterogeneity of the replacement process. Accordingly, the probability of observing a certain amino acid substitution depends not only on the nature of the two amino acids involved in the substitution, but also on the "residual environment" in which the substitution is taking place. Existing empirical models of mitochondrial amino acid evolution do not, however, discriminate between structural or chemical characteristics that are known to vary within the mitochondrial proteins, while MtHydro with its structural partitioning does. Accordingly, tests of model fit showed that my structural partitioning strategy, either as MtHydro or as two GTRs, is a legitimate improvement over all other empirical and mechanistic homogeneous models and may promote more reliable phylogenies. I also advocate that a similar partitioning strategy (and a corresponding partitioned model) should be estimated from and applied to other protein datasets, such as chloroplast coded proteins or even large phylogenomic datasets (perhaps to discriminate between ribosomal ones and other proteins).

I have shown that MtHydro can be considered superior to homogenous models, but what about other heterogeneous models? The current trend in accounting for among site heterogeneity is to allow different stationary frequencies (π) at different sites, while the replacement probabilities (r) are all equal; this has been effectively done by defining a certain number of categories and assigning each site of the alignment to a certain category, as for example in the CAT and the CAT related models (Le et al. 2008; Lartillot et al. 2007; Wang et al. 2008). In this way the exchangeability rates (q) are defined only by the frequency of the amino acids π (replacement rates are all equal and defined by a Poisson distribution), making the likelihood calculation extremely fast.

But what happen when two sites have a similar amino acid frequency but different replacement behaviour, for example because they belong to two different secondary structures? The CAT model will assign both sites to the same category and the exchangeability rate of the two sites will be regarded as same, even if is not. MtHydro has the advantage of discriminating between these two kinds of site, when they fall in different partitions. However, MtHydro is based on only two classes of sites, while heterogeneous models tend to have many more classes/categories (the CAT model generates approximately 110 categories to describe a typical mitochondrial dataset).

Mt126 is a complex heterogeneous model designed to implement 10 structural classes of sites. It accounts therefore for many more classes than my MtHydro, which has only two classes (Liò and Goldman 2002). However, one risk of using too many classes (or partitions) is that a single class may contain too few sites (and insufficient information) to allow a correct estimation of the corresponding GTR replacement matrix. For this reason Lio and Goldman have been forced to estimate all but one of their replacement matrices from *nuclear* proteins, an approach that can be extremely problematic for mitogenomics due to the genetic code differing between nuclear and mitochondrial DNA. This is the main reason why I decided to restrict the MtHydro partitions to two classes: one of trans membrane and the other of all other structures (mainly hydrophilic). The second reason has been of a more bioinformatic nature: programs for the prediction of transmembrane domain are very accurate, while prediction of other secondary structure types are less so. Having chosen only two partitions allowed me to have (i) an accurate prediction of the partitions and (ii) many sites for a precise inference of the replacement matrices. Of course my choice is in contrast with the actual complexity of proteins and the vast heterogeneity of replacement types among proteins. If only two partitions is probably a forcing (but every model is a forcing), more partitions would have been perhaps unfeasible.

Unfortunately, I was not able to compare my heterogeneous model directly with other ones such as CAT or MT126 because of their implementation. CAT is implemented only in PhyloBayes, which does not allow pre-partitions of datasets and MT126 is implemented in a relatively old Maximum Likelihood framework, not designed for large datasets such the ones I used. However, MT126 has been shown to be slightly superior or equivalent to homogenous models depending on the kind of dataset used (Lio and Goldman 2002), while I show that MtHydro (or the associated two GTR partitions) is always a great improvement in term of model fit to datasets. On the other hand, CAT has been repeatedly shown to be an effective and efficient way of reducing systematic problems such as LBA (Brinkmann et al. 2006). As a matter of fact, in chapter 4 and in Bourlat et al. (2009)), I show that while MtHydro seems unable to lessen putative LBA artefacts, the CAT(BP) model fully overcame this artefact. This suggests that the CAT approach is indeed superior to MtHydro. It is still clear, however that CAT is unable to discriminate between two sites with similar composition and different replacement rate. A possible interesting improvement is to include in a CAT framework the structurally based pre-partition of sites which is the main feature of MtHydro.

Finally, I would like to point an interesting and unexpected observation. According to AIC and BIC tests of model fit, empirical models should be preferable in mitogenomic studies to the mechanistic GTR one, as a moderate increase in the log-likelihood of GTR trees may not justify the much larger amount of time needed for computation and the vast number of free parameters in the model. This is probably due to the small size of the datasets I have used to test fit of models: small datasets may not contain sufficient substitutional information for a correct estimation of all the replacement rates of the GTR mechanistic matrix. This is reinforced by the fact that, according to table 2.1 and 2.2, the smaller the dataset, the worse is the performance of the mechanistic GTR models. Notably, the datasets I used to test the fit of models range between 20 and 44 taxa, which is the typical number of OTUs used in current mitogenomic studies. This, together with consideration of convergence (GTR parameters are very slow to be estimated and may keep independent runs from converging satisfactorily), reinforce the idea that empirically derived models should still be preferred in mitogenomic studies.

8.2 Is mitogenomics dead? Considerations over the utility of mitogenomics in deep metazoan phylogeny

Mitogenomics, from its phylogenetic point of view, has been repeatedly declared sick (Cameron et al. 2004, Shao and Barker 2006, Whitfield and Kjer 2008), if not dead, in the past decade. It is true that some important advances in animal systematics came from mitochondrial analyses, but the vast majority of those advances were based on the analysis of rare changes such as gene order rearrangements, large sequence signatures or variation in the genetic code (Boore et al. 1998, Papillon et al. 2005, Telford et al. 2000). On the other hand, some sensational evolutionary hypotheses from the analysis of mitochondrial coding sequences (for example paraphyly of hexapods and the relationship of the guinea pigs) have since been dramatically disproved by nuclear and morphological analyses (D'Erchia et al. 1996 and Nardi et al. 2003).

One invaluable aspect of mitogenomics is the "almost perfect" orthology of mitochondrial genes and the relative ease of generating new data. On the other hand mitogenomics clearly showed us its problems: striking acceleration of the evolutionary rate compared to nuclear sequences, across lineages heterogeneity of both composition and rate of sequence evolution and even the confounding effect of *Wolbachia* for inter-generic studies (Gibson et al 2005, Hassanin et al. 2005, Blouin et al. 1998, Whitwort et al. 2007). Furthermore, metazoan mitogenomes are intrinsically small as mitochondrial gene content is typically 13 genes in metazoans, which translates as a data matrix of only approximately 2000-3000 amino acid positions, leaving space for possible stochastic problems. For these reasons, and according to some of my results (chapter 3, 4 and 5), mitochondrial sequences seems to be easily prone to systematic and especially stochastic errors, particularly in the presence of ancient and/or fast radiations. I have also shown that my improved models of evolution (chapter 2) fail to eliminate some of these problems; however, a better model can lessen problems of systematic nature, not stochastic: if a dataset such as the mitochondrial one is too short to carry enough information for certain nodes, there is no analytical condition that can extract signal from where there is no signal. Finally, it has been recently shown that mitochondrial sequences may be prone to convergent adaptive evolution (non-neutral substitutions), a fact which can mislead all existing evolutionary models (Castoe et al. 2009).

I suggest that in the light of this and in the absence of useful rare changes, mitogenomics may represent an obsolete approach for deep phylogenetic studies and should be abandoned in favour of more reliable approaches such as the phylogenomic one (Philippe and Telford 2006). This is particularly true because new sequencing technologies such as pyrosequencing (Margulis et al. 2005) are beginning to allow a phylogenomic analysis for the price and the effort of sequencing a mitochondrial genome. While it seems to me sacrosanct and justified to attempt phylogenetic studies with the mitogenomes we already have, it is probably better to plan further metazoan and eukaryotic evolutionary studies using a phylogenomic approach.

This said, we now have more than 1500 complete animal mitochondrial genomes and although their sampling is taxonomically biased toward vertebrates and arthropods, they now cover most of the major metazoan lineages. This large sampling may justify a new effort in understanding how a miniaturised genome such as the mitochondrial one evolved and maintains itself. Some effort has been done in this direction, in particular during the 1990s, but today the function and the "ecology" of the mtDNA is still little known, although these genomes play a central role in many pathologies and are fundamental for the maintenance of the cell (Reyes et al. 1998). Paradoxically, if we are still unable fully to understand a small genome such as the mitochondrial one, how could we ever comprehend the much vaster and undoubtedly much complex nuclear one? In the light of this, mitogenomics may represent a good training for the understanding of genomic mechanisms and I advocate that mitogenomic studies should focus more on the potential genomic aspect rather than the phylogenetic one.

As cases in point, I have explored some mitogenomics characters from a more "genomic" point of view. Some analyses of mine, not present in this thesis, show that the large sequence signature in ND5, which has been used for phylogenetic purpose only (Papillon et al. 2004, Telford et al. 2008), mostly lies in a highly conserved cytosolic loop, leading to the possibility that the signature may correspond to a putative binding/regulative motifs. While the signature is very different between protostomes and deuterostomes, it is incredibly conserved within the two groups, suggesting a strong evolutionary constraint and a possible important role in mitochondrial homeostasis. I also show in Bourlat et al. 2009 that some structural features of the mitochondrial regulatory region of vertebrates exist in the enigmatic *Xenoturbella* in a reduced and extremely derived form, suggesting that these characters may be key for the regulation and maintenance of these mitogenomes. Puzzlingly, other deuterostomes such as cephalochordates and echinoderms do not seem to posses these structural features: is it because they evolved a different mitogenomic regulatory system? And if yes, why lose characters which seem to work efficiently, as seen in both distant *Xenoturbella* and vertebrates? The answer probably needs further and more accurate studies in this direction. I advocate that a similar approach should be carried out in the arthropods.

8.3 “Better models...and more genes”: is this enough?

Although it is true that stochastic errors will naturally vanish in a phylogenomic context, systematic errors will not disappear. Indeed, they should become even more apparent
Nicolas Lartillot and Hervè Philippe

The evolutionary relationships of major animal groups are now well understood. However, at least three types of nodes in the metazoan tree of life are still unresolved. Clearly, the first kind are those (not yet investigated) describing the terminal nodes and thus the relationships at the order, family or genus level. As first Darwin and later Dawkins pointed out, it is merely a matter of time before all these nodes will be satisfactorily described. A second and unjustified kind of unresolved nodes are those describing relationship of neglected or practically challenging minute or rare animal groups, such as the tiny ecdysozoan Loricifera, the lophotrochozoan Cycliophora, but also some of the crustaceans classes or insect orders. In the next few years most of these nodes will probably be addressed and possibly solved thanks to reductions in sequencing costs due for example to 454 and Illumina technologies. Unless they will dramatically fall in a third kind of node.

This third kind of unresolved node are the problematic ones, which describe lineages which are subject to phylogenetic artefacts due to systematic errors. They are indeed the most important nodes in the tree of life still under debate. Typical examples are fast evolving lineages such as nematodes and tardigrades, which are prone to LBA artefacts. Other (related) examples are lineages, such as myriapods and chelicerates whose speciation occurred in very ancient times and close to each other, promoting evolutionary scenarios which have been referred to as soft polytomies or bushes (Rokas and Carrol 2006). Others of these nodes are those describing lineages characterised by a long stem branch, due to recent radiation of the extant species and/or extinction of most of the stem lineages; the number of autapomorphies in these lineages are extremely high and may confound the phylogenetic signal. In all these kinds of node, the natural phylogenetic signal has been veiled by non historical signal to the extent that even large data matrices and sophisticated models of evolution currently used in systematics may fail unambiguously to solve them. These problems are generally caused by systematic errors due to model violation. The nodes describing the affinities of myriapods and tardigrades, which have been extensively analysed in this thesis are likely to be clear cases of problematic, systematic error-prone nodes.

The use of more genes has been shown to reduce the effect of stochastic errors (Philippe and Telford 2006) and I have indeed shown in chapter 5 that datasets containing many genes can be more consistent over the phylogeny they support. Moreover, better models of evolution and larger taxon sampling have promised to reduce the effect of systematic biases and promote more reliable phylogenies, in accordance with various results from this thesis. As a consequence, using the largest and longest available dataset and applying a “suitable enough” model of evolution is a tempting way to assure that all possible has been done, in particular if the inferred phylogeny is highly statistically supported. However, in the light of some of my results and as suggested by recent literature, this may not be enough to describe problematic nodes (Rodriguez-Ezpeleta et al. 2007). Also, the model may not be suitable enough, even if it is the best available.

I have repeatedly shown in this thesis that certain phylogenetic relationships may depend on the model applied, the taxonomic sampling and/or the kind of positions used, even if a “suitable enough” model supports a certain topology with high confidence. The mere presentation of the tree obtained with the “optimal” settings may be a simplistic view of the phylogenetic signal in sequences. On the other hand, a comparative approach, based on the evaluation of how the phylogenetic signal changes over different analyses may give a broader and more realistic view of the phylogenetic problem. Different phylogenies, obtained under different settings, should be compared against the context of possible systematic and/or stochastic errors. For example it may be of great interest to compare tree topologies obtained under settings which minimise potential sources of errors against those settings which maximise them. To a certain extent, it may be possible to predict phylogenetic artifacts using setting which exaggerate them, as I shown in chapter 6.1 and in chapter 4.

If the results of such comparative approaches are consistent, then we may have an indication that my phylogeny is robust. This is not, however, always true: the phylogenomic dataset of Dunn et al. (2008) as (re)analysed in chapter 5 consistently supports Myriochelata under various parameters and only the addition of new taxa resulted in a change of signal (figure 6.3). The apparent robustness of the Myriochelata node in the Dunn dataset, was therefore probably due to a systematic error (related to poor myriapod sampling and subsequent higher likelihood of LBA) which was too strong to be detected in the absence of additional myriapod.

Conversely, if the results of a comparative approach are inconsistent, it may mean that systematic and/or stochastic problems affect the focal phylogenetic reconstruction. However, a detailed examination of the alternative results may suggest which is the correct topology. In chapter 6.1 for example, I showed that, using phylogenomics, different models and different outgroups promote different positions of the myriapods. From a superficial point of view this result can be interpreted as a lack of resolution due to insufficient phylogenetic signal. I was, however, able to show that there is valid explanation for the "jumping" of myriapods: the myriapod lineage is attracted by the chelicerate lineage, as suggested by observing how the topology (or the support at nodes) varies when better fitting models and closer/slower evolving outgroups are sequentially used. The vast majority of the findings presented in this thesis wouldn't be possible without the well-considered comparison of many trees obtained under different analytical/parametrical settings.

According to the results of my thesis, taxonomic sampling seems to be the parameter which most affects the phylogenetic reconstruction. It has already been shown that increased taxon sampling greatly reduces errors in phylogeny (Zwickl and Hillis 2002); my point is that taxon sampling (not only the gain, but also the exclusion of taxa) is useful because it may reveal possible systematic errors and show the most tenable hypothesis. In chapter 4 for example, I showed that the LBA nature of the tardigrades/chelicerates grouping can be enlightened by sequential taxa removal. A similar strategy has been used in chapter 6 to study the affinities of myriapods.

I would like to point out that there are other (probably more effective ways) to explore problematic nodes than the "comparative approach" I suggest. For example, increasing the taxon sampling can potentially reduce the effect of systematic errors and certainly helps in recovering the correct topology. Of course, one may wait to address a certain node (writing the paper) until the taxon sampling is very dense and the complete genome is available for all the taxa; however, given the current trend of academic research, waiting too long is probably not the most productive idea. Also, adding more sequences is obviously a costly solution which still relies on a large budget. Increased taxon sampling is also not always feasible: for example Placozoa and Xenoturbellidae phyla are comprised by only one (or few closely related) extant species each.

In conclusion, I advocate a detailed investigation of the phylogenetic signal at highly problematic nodes. This can be done by exploring how the signal changes while sequentially changing some key analytical parameters, in particular, as suggested by my thesis, the taxonomic sampling. This should be followed by a comparison of different tree topologies (if any) against the specific attributes of the data in light of possible systematic and/or stochastic problems in order to assess which topology is the most likely. Recent published molecular analyses are in line with this point of view, but outside the specialised field of molecular systematics such an approach is still quite underestimated or even unknown (Brinkmann et al. 2005, Lartillot et al. 2007, Rodriguez- Ezpeleta et al. 2007, Pisani et al. 2009). Recently, while presenting some of the results included in this thesis at the meeting of the Systematic Association in Leiden, a critical (and frank) morphologist asked me: “How can you trust your data if the myriapods keep jumping around?” I answered: “ Because I know why they jump and, in some cases, I can predict where they will jump”.

8.4 The ancestral ecdysozoan

As discussed in the introduction, it is extremely important for drawing a picture of the ancestral ecdysozoan to determine whether Cycloneuralia (nematodes, priapulids and their kin) have a paraphyletic or a monophyletic origin.

My phylogenomic analyses of chapter 6 support a paraphyletic origin of the Cycloenuralia, with the Scalidophora (priapulids and their kin) basal to a group of nematodes plus panarthropods. This is in accordance with ribosomal markers (Garey 2001, Mallat and Giribet 2006), but in contrast with a previous phylogenomic study, which instead supported monophyly of Cycloneuralia (nematodes + Scalidophora, Dunn et al. 2008). Notably, when updating the gene selection of Dunn and colleagues (2008) to my larger taxon sampling, a paraphyletic origin of the Cycloneuralia is recovered. It has to be remarked that the dataset of Dunn, while including fewer nematodes and arthropods than my dataset, contains a key species, the nematomorph *Spinochordodes tellinii.* This species is extremely important for addressing affinities of Cycloneuralia, as it is believed to be the sister of nematodes and therefore may shorten the long nematode stem branch and lessen possible systematic artefacts such as LBA. However, gene sampling for *Spinochordodes* is exceptionally reduced (in the Dunn dataset covers only 11% of the alignment for approximately 2000 aa positions) and may promote stochastic artifacts due to reduced phylogenetic information. For this reason, in the phylogenomics study of chapter 6, I have set a cut-off in order to exclude species which covered less than 15% of the concatenated alignment (30% for the 30 taxa dataset) and accordingly excluded species such as the nematomorph.

The paraphyletic nature of the Cycloneuralia, as supported by my analyses, suggests that the ancestral Ecdysozoa was cycloneuralian-like, thus possessing a circumpharyngeal brain and an introvert. This implies that the panarthropods are the most derived of the Ecdysozoa and evolved from an introverted collar-brained ancestor as pointed out by Garey (Garey 2001). This view is reinforced by the analysis of Eriksson and Budd (2000), which suggested that the onychophoran brain evolved from a circumpharyngeal nerve ring (the cycloneuralian state) by expansion of the dorsal part of the ring. Similarly, the tardigrade brain consists of a circumesophageal ring with dorsal lateral lobes. Furthermore, the tritocerebral commissure of (eu)arthropods loops around the pharynx, suggesting that many tritocerebral ganglia were originally located behind the mouth and migrated dorsally during evolution. The overall scenario may be that the panarthropods dorsally enlarged a putative ancestral ecdysozoan circumpharyngeal brain. A gradual achivment of neuronal characters in the panarthropods is somehow reinforced by a recent neuro anatomical study which has shown that ganglia are not ancestrally segmented in the onychophoran (Mayer and Whitington 2009b). What happened to the introvert in panarthropods is rather less clear and even more speculative. Analyses of gene expression in the introvert of cycloneuralians and comparison with possible orthologs in the (pan)arthropods may give exciting answers.

8.5 Tardigrada: finally Panarthropoda, perhaps Lobopoda?

Despite their potential importance as outgroups to the euarthropods, the position of tardigrades is far from being uncontentious. Recent phylogenomic studies even challenged the (pan)arthropod nature of tardigrades, grouping them with nematodes (Lartillot and Philippe 2008), although it seems that their affinity is model dependent (Dunn et al. 2008). In this thesis, using two different types of dataset, I gave evidence that tardigrades should be grouped in a monophyletic panarthropod clade and that previous support for their nematode affinity is likely the effect of systematic error. The most likely scenario, as suggested by my results, is a sister relationship between onychophorans and tardigrades, a clade which can be regarded as extant Lobopodia.

The mitogenomic studies of chapter 4 show that tardigrades tend to branch either with fast evolving arthropods or with the outgroups, suggesting a reiterated LBA. Furthermore, I showed that the CAT model, which has been shown to lessen the LBA artefact (Lartillot and Philippe 2008), consistently supports a group of onychophorans plus tardigrades in a monophyletic panarthropod clade. As further evidence, the signal nesting tardigrades within long branched arthropods is found in the least reliable source of signal, the fastest evolving and/or slow evolving sites.

This view is corroborated by my phylogenomic analyses of chapter 6 which also groups tardigrades and onychophorans. While the CAT model supports this grouping consistently, homogenous models tend to group tardigrades with nematodes. Exclusion of fast evolving nematodes from the dataset leads to an inconsistent tree topology: tardigrades became sister to the onychophorans rather than basal panarthropods, as expected by the mere removal of nematodes, suggesting an LBA artifact between nematodes and tardigrades. However, one caveat of my phylogenomic study is that data completion in both tardigrades and onychophorans cover only approximately 50% of the original alignment, suggesting a possible effect of missing data and a consequent artefactual attraction. New sequences from the two lineages and improved gene coverage may solve the problem and exclude possible “missing data attraction” between tardigrades and onychophorans. Still, the simple observation of tardigrades and onychophoran branch lengths excludes the possibility of an LBA: tardigrades are fast evolving while onychophorans have a moderate rate of evolution. Assuming an LBA artefact involving tardigrades, we should expect attraction to occur towards more distant outgroup sequences for example, rather than with onychophorans.

There are, however, no commonly accepted synapomorphies of a tardigrade - onychophoran clade, though morphologists are divided over whether one of the two is the sister group of the Euarthropoda. A tentative character uniting the tardigrades and the onychophorans is their shared possession of non-articulated clawed appendages, as in the Cambrian lobopodian *Aysheaia*, but in contrast with arthropods which have articulated ones. However a lack of information from panarthropod stem group (and/or the difficulty to assess their phylogenetic position) prevents from possible polarisation of this character. Tardigrades lack certain characters shared by Onychophora and Euarthropoda such as an ostiate heart, but are plausibly seen to have secondarily lost such characteristics through miniaturisation. Evidence from cuticular and nuero developmental structures suggests instead a sister relationship between the arthropods and the tardigrades (Nielsen 2001, Mayer and Whitingotn 2009b). It is however clear that morphological comparisons are complicated by the extremely reduced and derived nature of the tardigrades.

Regardless of Onychophora and Tardigrada being truly sister groups or not, my analyses support monophyletic Panarthropoda. This has great support from the morphological point of view as all the three panarthropod lineages share a segmented body and paired walking appendages, which are unique features within ecdysozoans and should be regarded as synapomorphies of the panarthropods.

In any case it is clear that the divergences leading to the three main panarthropod lineages occurred very deep in time, in the middle Cambrian or even earlier (according with some unpublished results of mine and to Regier et al. 2005). Consequently, a significant level of mutational saturation is expected in all the lineages in particular in the tardigrades one, which is fast evolving. Moreover, the tardigrades sampled in my thesis (*Hypsibius dujardini*, *Richtersius coronifer* and *Thulinia sp*.) are all members of the same class, Eutardigrada, a fact that may have promoted an extremely long stem tardigrades branch (and consequent higher level of autapomorphies). Consequently, a broader taxon sampling, possibly from the Heterotardigrada class, may help to reduce the length of tardigrades stem branch, lower the likelihood of systematic artifact and draw firmer conclusions over the monophyletic group of extant lobopodian as observed throughout my analyses.

8.6 The hexapods and their origin.

Although I was not able to come to a conclusive hypothesis over their relationships (chapter 6.2), the most likely scenario is that hexapods evolved from within paraphyletic crustaceans. Most of the analyses suggest that the hexapods are sister to branchiopodan crustaceans. Branchiopods are fresh water crustaceans and the few species which inhabit salt waters appear to do so as a secondary adaptation, suggesting that the ancestral branchiopod lived in fresh water This may imply that hexapods colonised land not form the sea, but rather from a lake, as recently suggested (Glenner et al. 2006). Intriguingly, a river or a lake has far more points of contact with land than a sea has plus can be more prone to the effects of drought, increasing the likelihood of tentative landings of a putative branchiopod-hexapod intermediate. Interestingly, this is reinforced by the fossil records of the first putative hexapods from the Devonian Rynie

Lagerstaette (*Ryeniella*), which has been found in a deposit of land (mountainous) rather than costal origin. On the other hand, neuroanatomical evidences (Fanenbruck et al. 2004) support a closer relationship between the hexapods and a malacostracans/remipede clade, which has been recently corroborated by the analysis of respiratory hemocianin proteins in this groups (Ertas et al. 2009). These authors suggests that the enigmatic lower Devonian fossil *Devonohexapodus boksbergensis* presents features shared by both hexapods and remipedes. Recent molecular studies (Regier et al. 2008, Economou 2008) supported a remipede affinity for the hexapods, but these evidences, based on various molecular markers, are complicated by the fast rate of evolution observed in the remipede. Notably, remipedes inhabit only anchialine marine caves, in which salt water from a subterranean sea connection mixes with fresh water from the top of the cave, still not completely sea. The overall scenario is contentious and extremely interesting because the relative position of rempiede and branchiopods in the Tetraconata assemblage may shed interesting light on the terrestralisation routes of the arthropods.

An interesting outcome of my phylogenomic analyses is the monophyly of Hemimetabola or Exopterigota insects (chapter 6.2). I advocate that this hypothesis should be taken into serious account because it seems unlikely to be the result of systematic error; the competing Eumetabola hypotheses seem to be the result of a series of artefactual reconstructions in previous phylogenies. On the other hand, the insects have been rooted in my tree with the collembolans, which are quite distantly related and may have promoted a rooting problem. Surprisingly, a rather recent phylogenomic study of the arthropods, using closer outgroups (dragon fly, mayfly and silver fish, respectively Odonata Ephemeroptera and Archaeognatha) supported the hemipterans as basal Pterygota, with the orhropterans (however with the hemipteroid Phtiraptera) closer to holometabolans, in complete contrast with my phylogeny (Karen Meusemann, Bonn, personal communication).

Monophyly of hemimetabolans as supported by my results (but also the results of Karen Meusemann) notably challenges the widely accepted Eumetabola which groups instead hemipteroids with holometabolans. There is however, only one valid synapomorphy uniting the Eumetabola; a shared sclerotisation in the hind wings of hemipterans and holometabolans (Kristensen 1975). A 1965 textbooks from Ross (referenced in Rasnitsyn, 1998) and Paulus (1979) also reports the “juvenile ocelli suppression” as a eumetabolan character. On the other hand, my monophyletic group of hemimetabolans resuscitate the Paurometabola assemblage, which groups all the hemimetabolans except for the Plecoptera (Korschelt and Heider 1895). Some work has shown that the holometabolan insects are characterised by a different pattern of diversification and more complex feeding strategies than both hemipteroid and orthopteroid (see Yang 2001). However, these characters should be regarded as synapomorphies of the holometabolans and cannot give evidence for a monophyletic origins of hemimetabolans.

In any case, a monophyletic assemblage of hemimetabolans has profound implications for our understanding of insect evolution, in particular concerning their evolution of development and their ancestral states. The Eumetabola hypothesis suggests a gradual evolution from incomplete (as in some orthopteroid hemimetabolans) to complete metamorphosis (as in all the holometabolans) as suggested by the complex ontogenic pattern of most of the hemipteroids (especially hemipterans), which can be interpreted as an intermediate stage between the typical hemimetabolan and the holometabolan ontogenic process. In the light of my results, the most likely scenario is that the holometabolism is a complete novelty and that the complex ontogenic patterns observed in most of the hemipteroids are an independent , although partial gain.

8.7 Affinities of the myriapods: back to Mandibulata?

One of the most debated systematic issues of the last decade has been the relative position of myriapods and chelicerates with respect to other arthropod lineages. A notable outcome of my thesis is that phylogenomic (chapter 6.1) and to a lesser extent mitogenomic (chapter 2, 3 and 4) studies support a monophyletic origin of myriapods, crustaceans and hexapods (the Mandibulata), in contrast with the grouping of myriapods and chelicerates (the Myriochelata).

The possibility that the molecular phylogenies supporting Myriochelata might have been affected by a systematic error was highlighted by the occasional contradictory results (Regier et al. 2008, Pisani 2004, Rota Stabelli and Telford 2008) and in particular by the conflict with morphological data (see introduction). The contradictory nature of myriapod affinities is reinforced by the reanalyses of various phylogenetic datasets in chapter 5, which show that signal at this node is sparse. Furthermore, in all the trees throughout my thesis, the branch leading to Mandibulata (or Myriochelata) is short, making it particularly susceptible to the effects of LBA, a systematic error that could unite the slowly evolving myriapods and chelicerates in the midst of several other fast evolving taxa.

Accordingly, in my phylogenomic study, experiments designed to reduce the effects of systematic error – increased taxon sampling, additional data of lower saturation, exclusion of outgroups with the longest branches, removal of the fastest evolving positions and the use of improved evolutionary models – all resulted in increases in support for Mandibulata (chapter 6.1). In addition to phylogenomic studies, two different mitogenomic datasets of mine support Mandibulata under some reasonable conditions: when a better fitting model is used (as in chapter 2.1.5), when an optimal outgroup is used (chapter 3.6), and when fast evolving species and both fast and slow evolving sites are excluded from the analyses (as in chapter 4.2.5). In conclusion, the most tenable position of the myriapods, from the analysis of my mitogenomic and especially large phylogenomic datasets, is as the sister group of the Tetraconata.

To my knowledge, these are among the first robust molecular studies in support of monophyletic Mandibulata and represent a significant contribution for the understanding of basal arthropod relationships. Remarkably, my analyses tend to reconcile molecules and morphology, as Mandibulata is manifestly sustained by the majority of morphological and developmental studies, but have always found poor support in molecules. However, very recently, the similar way in which ganglia forms in spiders, centipedes and millipedes (Chipman and Stollewerk 2006) have been polarised by the work of Mayer and Withington (2009) as a synapomorphy of the Myriochelata. Mayer also suggested that the cumulus, a group of cells which determine the dorsal region, may be a novelty of the Myriochelata as it seems absent in onychophorans and has never been reported in Tetraconata. While this latter character should be addressed definitively using *decapentaplegic* gene expression (Mayer, personal communication), the scenario is becoming very interesting: if Mandibulata is a true clade, as my molecular studies suggest, the above mentioned characters may have been gained by the arthropod common ancestor and lost in the tetraconatan ancestor. It is much less parsimonious, according to the Myriochelata hypothesis, that all the characters uniting the Mandibulata arose two times independently or have been ancestrally gained in the arthropod and secondarily lost in the chelicerates.

The Mandibulata is by far the largest clade of animals on earth, but the origin of this successful bodyplan in terms of the evolution of its development remains obscure. The picture from palaeontology is, somewhat clearer. Cambrian fossils that have been identified as a grade of stem-group mandibulates (Richter and Wirkner 2004) indicate a crustacean-like *habitus* for basal members of the Mandibulata and shed light on how a mandible is likely to have evolved. The limb on the third cephalic segment (the mandible homologue) in Cambrian stem-group mandibulates such as *Martinssonia* displays a stronger development of a movable, setose process at the limb base ("proximal endite"; Waloszek et al. 2007) than that on the adjacent limbs (Moura et al 1996). The more elaborated proximal endite used for food manipulation is viewed as a precursor to the fully differentiated coxal chewing surface in the mandibulate crown group (Zhang et al. 2008). Further studies of fossils and embryos in the light of what I suggest is a reliable phylogeny of euarthropod classes should clarify the evolution of the mandibulate bodyplan (Telford and Budd 2003).

8.8 Final remarks

> *...(1) What is the immediate sistergroup to the arthropods? (2) Can the position of the insects be clarified, perhaps as sistergroup to a particular clade of malacostracans crustacean? (3) Where do the affinities of the myriapods lie? (4) How much can we learn from the fossil record about arthropod evolution? Given the degree of disagreement at the conference on all of these topics, no early solution is in sight.*
>
> *Graham E. Budd*

The above quotation is from a comment published after “the” conference on arthropod relationships organised by Richard Thomas and Richard Fortey at the Natural History Museum of London in 1996 (Budd 1996). After 13 years, the four issues discussed by Budd are still extremely current and neither a decade of evo-devo studies, nor the advent of phylogenomics, have unambiguously solved them.

The work presented in this thesis may help to address the first three of these questions. According to my results, “*the immediate sistergroup to the arthropods”* should be a monophyletic group of extant Lobopoda (onychophorans plus tardigrades) and “*the affinities of the myriapods lie”* close to the insects and crustaceans in a monophyletic Mandibulata clade. My analyses failed to unambiguously describe “*the position of the insects”*, although it is clear that this position does not have to be found in “*a particular clade of malacostracans crustacean”*. My analyses also gave convincing evidence that Cycloneuralia are paraphyletic and that panarthropods originated form a cycloneuralian ancestor. I acknowledge that these hypotheses have to be further tested using more genes and more taxa before drawing conclusions. I advocate, however, that the monophyletic origin of Mandibulata should be regarded as credible, not only because it is robustly supported by my large phylogenomic dataset, but also because I was able to show that the competing hypothesis (Myriochelata) is associated with conditions that promote systematic errors.

When addressing the phylogenetic issues of the ecdysozoans, I encountered various methodological problems, which I have overcome by creating new tools and methods. I have developed a pipeline and a new statistic for the selection of suitable outgroups and estimated two new models of evolution that describe more accurately the evolution of animal sequences.

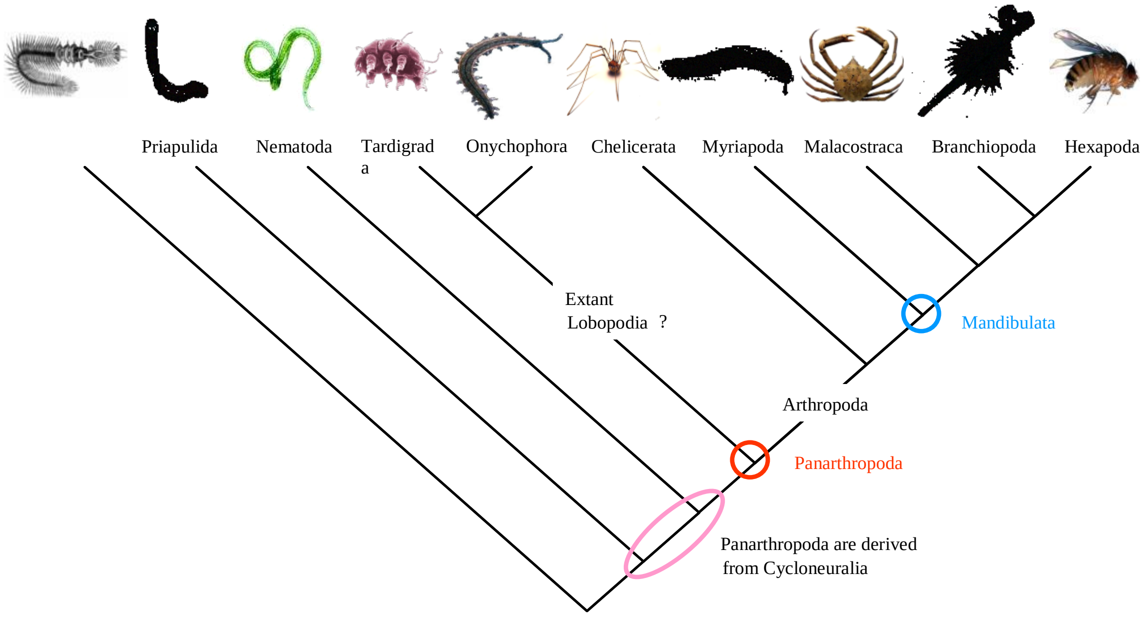


Figure 8.1: Major hypothesis presented in this thesis. Myriapods are sister to a group of hexapods plus crustaceans in a monophyletic Mandibulata clade (blue circle). Panarthropods (arthropods, tardigrades and onychophorans) form a monophyletic group (red circle) in which tardigrades and onychophorans are sistergroup. The cycloenuralians are paraphyletic implying that panarthropods evolved from a cycloneuralian ancestor (pink circle).

It is my hope that these methodological advances, as well as the phylogenetic hypothesis I have presented, will move forward our understanding of ecdysozoan and animal evolution.

FINE

References

Abascal, F., Posada, D. and Zardoya, R. 2007. MtArt: a new model of amino acid replacement for Arthropoda. Mol Biol Evol 24:1-5.

Abascal, F., Posada, D., Knight, R.D. and Zardoya, R. 2006. Parallel evolution of the genetic code in arthropod mitochondrial genomes. PLoS Biol 4:711-718.

Abzhanov, A. and Kaufman, T. C. 2000. Homologs of Drosophila appendage genes in the patterning of arthropod limbs. Developmental Biology, 227:673-680.

Adachi, J., and Hasegawa. M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459-68.

Adkins, R.M., Walton A.H., and. Honeycutt. R.L. 2003 Higher-level systematics of rodents and divergence time estimates based on two congruent nuclear genes. Mol Phylogenet Evol, 26:409-420.

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet,O., Prud'homme, B. and De Rosa, R. 2000. The new animal phylogeny: Reliability and implications. Proc Natl Acad Sci USA. 25.97:4453-6.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff and J. A. Lake.1997. Evidence for a clade of nematodes, arthropods, and other moulting animals. Nature 387:489–493.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6):716-723.

Atkinson, H.J. 1973. The Respiratory Physiology of the Marine Nematodes Enoplus brevis (Bastian) and E. communis (Bastian): I. The Influence of Oxygen Tension and Body Size. J. Exp. Biol. 59(1): 255–266.

Bapteste, E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller M and Philippe H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A. 99(3):1414-9.

Baurain, D. Brinkmann, H. and Philippe. H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol. 24:6-9.
Bitsch, J. 2001. The arthropod mandible: morphology and evolution. Phylogenetic implications. Annales de la Société Entomologique de France (N.S.), 37 :305-321.

Bitsch, J. and Bitsch, C. 2007. The segmental organization of the head region in Chelicerata: a critical review of recent studies and hypotheses. Acta Zoologica, 88:317-335.

Blair, J. E., Ikeo, K., Gojobori, T. and Hedges, S. B. 2002 The evolutionary position of nematodes. BMC Evol. Biol. 2:7.

Blanquart S. and N. Lartillot. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 2008 May;25(5):842-58.

Blouin, H. S., Yowell, C. A., Courtney, C.H. and Dame, J. B. 1998 Sysyetmatic bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol. Biol. Evol. 15,12:1719-1727.

Boore, J. L., Lavrov, D. V. and Brown, W. M. 1998. Gene translocation links insects and crustaceans. Nature 392: 667-668.

Bourlat S.J, Rota-Stabelli, O. ,Lanfear, R. and Telford. M. J. 2009. The mitochondrial genome of Xenoturbella is ancestral within the deuterostome. BMC Evol Biol 9:107.

Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., Thorndyke, M., Nakano, H., Kohn, A.B., Heyland, A., Moroz, L.L., Copley, R.R. and Telford, M.J. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature. 444(7115):85-88.

Braddy, S.J., Markus Poschmann, M., and Tetlie, O.E. 2008. Giant claw reveals the largest ever arthropod. Biology Letters 4: 106–109.

Brinkmann, H. and Phillipe, H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817-825.

Brothers, D. J. 1999. Phylogeny and evolution of wasps, ants and bees (Hymenoptera, Chrysidoidea, Vespoidea and Apoidea). Zoologica Scripta 28 (1-2): 233-249.

Brusca, R.C. and Brusca, G.J. 2001. Invertebrates. Sinauer associates .

Budd, G. E. 2004. Palaeontology: Lost children of the Cambrian. Nature 427:205-207.

Budd, G. E. 1996. Progress and problems in arthropod phylogeny. TREE 11:356-358.

Cameron, S.L., Miller, K.B., D'Haese, C.A., Whiting, M.F. and Barker, S.C. 2004. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda). Cladistics 20:534-557.

Cameron, S.L., Barker, S.C. and Whiting, M.F. 2006. Mitochondrial genomics and the new insect order Mantophasmatodea. Mol. Phylogenet. Evol.38(1):274-9.

Carapelli, A., Liò, P., Nardi, F., Van der Wath, E. and Frati, F. 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. BMC Evol. Biol. 16;7 Suppl 2:S8.

Castoea, T. A., Jason de Koninga, A. P., Kima, H., Gua, W., Noonanb, B.P., Naylorc, G., Jiangd, Z. J., Parkinsond, C. L., and Pollocka. D. D. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. PNAS. 106. 8986-8991.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Chapman, A. 2005. Numbers of living species in Australia and the World. Report, Department of the Environment and Heritage, Canberra, Australia.

Chipman, A. D. & Stollewerk, A. 2006. Specification of neural precursor identity in the geophilomorph centipede Strigamia maritima. Dev. Biol. 290:337-350.

Chipman, A. D., Arthur, W. & Akam, M. 2004. Early development and segment formation in the centipede Strigamia maritima (Geophilomorpha). Evol. Dev. 6:78-89.

Chuan, M., Liu, C., Yang, P. and Kang. L. 2009. The complete mitochondrial genomes of two band-winged grasshoppers, Gastrimargus marmoratus and Oedaleus asiaticus BMC Genomics, 10:156:10.1186/1471-2164-10-156.

Cook, C. E., Smith, M. L., Telford, M. J., Bastianello, A. and Akam, M. 2001. Hox genes and the phylogeny of the arthropods. Curr. Biol. 11:759-63.

Copley, R. R., Aloy, P., Russell, R. B. & Telford, M. J. 2004 Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. Evol. Dev. 6,3:164-169.

D'Erchia, A., Gissi, C., Pesole, G., Saccone, C. and Arnason, U.1996. The guinea pigi s not a rodent. Nature. 381:597-600.

Curole, J.P. and T. D. Kocher. 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. Trends Ecol Evol. 14:394-398.

Davis, A.D., Weatherby, T.M., Hartline, D.K. and Lenz, P.H. 1999. Myelin-like sheaths in copepod axons. Nature 15,398 (6728):571.

De Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M., Carroll, S. B. & Balavoine, G. 1999 Hox genes in brachiopods and priapulids: implications for protostome evolution. Nature 399:772–776.

Delsuc, F., Phillips, M. J. and Penny, D. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?" Science 301:1482.

Dohle W. 2001. Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name 'Tetraconata' for the monophyletic unit Crustacea+Hexapoda. Ann. Soc. Entomol. France. 37:85–103.

Dohle, W. 1997. Are the insects more closely related to the crustaceans than to the myriapods? Entomol. Scand. Suppl. 51: 7–16.

Douzery E.J.P., Delsuc, F., Stanhope, M.J. and Huchon, D. 2003. Local molecular clocks in three nuclear genes: divergence times for rodents and other mammals and incompatibility among fossil calibrations. J Mol Evol, 57:S201-S203.

Dowton, M. and Austin, A. D. 2001. Simultaneous analysis of 16S, 28S, COI and morphology in the Hymenoptera: Apocrita - evolutionary transitions among parasitic wasps. Biological Journal of the Linnean Society 74:87-111.

Dunn, C. W., Hejnol, A., Matus, , D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q. and Giribet, G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-9.

Eck, R. V. and Dayhoff, M. O. 1966. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science 152:363-366.

Economou, A. D. 2008. Phylogenetic and Developmental Studies into the Evolution of an Insect Novelty. Thesis submitted for the Degree of Doctor of Philosophy, UCL, August 2008.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.

Edgecombe GD, Giribet G. 2007. Evolutionary biology of centipedes. Ann Rev Entomol 52:151-170.

Edgecombe, G. D., Richter, S. and Wilson, G. D. F. 2003. The mandibular gnathal edge: Homologous structures across the Mandibulata? In Hamer, M. (ed). Proceedings of the 12th International Congress of Myriapodology. African Invertebrates 44:115-135.

Edman, P. 1950. Preparation of phenylthiohydantoins from some natural amino acids. Acta Chem. Scand. 4:283.

Edwards, A. W. F. and Cavalli-Sforza, L. L.1964. Reconstruction of evolutionary trees. pp. 67-76 in Phenetic and Phylogenetic Classification, ed. V. H. Heywood and J. McNeill. Systematics Association Volume No. 6. London: Systematics Association.

Eernisse, D.J., Albert, J.S. and Anderson, F.E. 1992. Annelida and Arthropoda are not sister taxa: A phylogenetic analysis of spiralian metazoan morphology, Syst. Biol. 41: 305–330.

Eriksson, B. J. & Budd, G. E. 2000. Onychophoran cephalic nerves and their bearing on our understanding of head segmentation and stem-group evolution of Arthropoda. Arthropod Struct. & Developm. 29: 197–209.

Ertas, B., von Reumont, B. M., Wagele, J.W., Misof, B. And Burmester, T. 2009. Hemocyanin suggests a close relationship of Remipedia and Hexapoda. Mol. Biol. Evol. Doi:10.1093/molbev/msp186.

Fanenbruck, M., Hartzsch, S. and Wagele, J.W. 2004. The brain of Remipedia (Crustaceans) and an alternative hypothesis on their phylogenetic relathionhips. Proc.Natl. Acad. Sci. USA. 101:3868-3873.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Foster, P. G., L. S. Jermiin, and Hickey, D. A. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 44:282-8.
Foster, P.G. and Hickey, D.A. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48:284-90.

Friedrich, M. and Tautz, D. 1995. rDNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165–167.

Garey, J. R. 2001. Ecdysozoa: the relationhip between Cycloenurlaia and Panarthropoda. Zool. Anz. 240:321-330.

Gibson, A., Gowri-Shankar, V., Higgs, P.G. and Rattray, M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol 22 251-64.

Giribet, G., Carranza, S., Baguna, J., Riutort, M. and Ribera, C. 1996. First molecular evidence for the existence of a Tardigrada + Arthropoda clade. Mol Biol Evol 13:76-84.

Giribet, G., Edgecombe, G.D. and Wheeler, W.C. 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature 413:157-61.

Giribet, G., Richter, S., Edgecombe, G. D. & Wheeler, W. C. 2005. The position of crustaceans within the Arthropoda - evidence from nine molecular loci and morphology. In S. Koenemann & R. A. Jenner (Eds), Crustacean Issues 16: Crustacea and Arthropod Relationships. Festschrift for Frederick R. Schram (pp. 307-352). Boca Raton: Taylor & Francis.

Glenner H., Thomsen, P. F., Hebsgaard, M. B., Sorensen, M.V. and E. Willerslev. 2006. Evolution: the origin of insects. Science 22 December 314:1883-1884

Goldman, N., Thorne, J. L. and. Jones, D. T 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Biol 263:196-208.

Grimaldi, D. 1997. A fossil mantis (Insecta: Mantodea) in Cretaceous amber of New Jersey, with comments on the early history of Dictyoptera. Am Mus Novitates, 3024:1-11.

Grimaldi, D. and Engel, M. S.. 2005. Evolution of the Insects. Cambridge: Cambridge University Press. ISBN 0-521-82149-5. pp. 139-141.

Guex, N., and Peitsch, M. C.1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714-23.

Haase, A., Stern, M., Wa¨chtler, K. & Bicker, G. 2001 A tissue-specific marker of Ecdysozoa. Dev. Genes Evol. 211:428–433.

Halanych K.M., Bacheller J.D., Auginaldo A.M.A., Liva S.M., Hillis D.M. and Lake, J.A. 1995. Evidence from 18S ribosomal DNA that the Lophphorates are protostome animals. Science 267:1641-1643.

Hamilton, K.G.A. 1972. The insect wing, Pt. IV. Venational trends and the phylogeny of the winged orders. Journal of the Kansas Entomological Society 45:295-308.

Harzsch, S. 2002 The phylogenetic significance of crustacean optic neuropils and chiasmata: a re-examination. J. Comp. Neurol. 453:10–21.

Harzsch, S. 2004. Phylogenetic comparison of serotonin-immunoreactive neurons in representatives of the Chilopoda, Diplopoda, and Chelicerata: implications for arthropod relationships. Journal of Morphology, 259:198-213.

Harzsch, S., Muller, C.H. and Wolf, H. 2005. From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and "Myriapoda" but favour the Mandibulata concept. Dev Genes Evol 215:53-68.

Hassanin, A., Leger, N. and Deutsch, J.: Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. Systematic Biology 54 (2005):277-298.

Hassanin, A.: Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol Phylogenet Evol 38 (2006):100-16.

Heming B. S. 2003. Insect Development and Evolution. Cornell University Press.

Hessler, R. R. 1992. Reflections on the phylogenetic position of the Cephalocarida. Acta Zoologica 73:315-316.

Hejnol, A., M. Obst, A. Stamatakis, M. Ott, G.W. Rouse, G.D. Edgecombe, P. Martinez, J. Baguñà, X. Bailly, U. Jondelius, M. Wiens, W.E. Müller, E. Seaver, W.C. Wheeler, M.Q. Martindale, G. Giribet and C.W. Dunn. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods.Proc. Biol. Sci. 276:4261-70.

Heymons, R., 1901. Die Entwicklungsgeschichte der Scolopender. Zoologia, Heft 33, S.1-244, Taf.I-VIII.

Horner, D. S., Lefkimmiatis, K., Reyes, A., Gissi, C., Saccone, C. and Pesole, G. 2007. Phylogenetic analyses of complete mitochondrial genome sequences suggest a basal divergence of the enigmatic rodent Anomalurus. BMC Evol Biol 7:16.

Hovmöller, R., Pape, T., and Källersjö, M. 2002. The Palaeoptera problem: basal pterygote phylogeny inferred from 18S and 28S rDNA sequences. Cladistics 18: 313–323.

Huelsenbeck JP and Suchard MA. 2007. A Nonparametric Method for Accommodating and Testing Across-Site Rate Variation. Syst Biol. 2007 Dec 56(6):975-87

Huelsenbeck, J. P., and Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Huelsenbeck, J. P., Larget, B. and Alfaro M. E. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol Biol Evol 21:1123-33.

Hughes, C.L. and Kaufman, T. C. 2002. Hox genes and the evolution of the arthropod body plan. Evolution and Development, 4:459-499.

Hughes, J., S. J. Longhorn, A. Papadopoulou, K. Theodorides, A. De-Riva, M. Mejia-Chang, P. G. Foster, and A. P. Vogler. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Mol. Biol. Evol. 23:268 - 278.

Hunt T., Bergsten J., Levkanicova Z., Papadopoulou A., John O.S., Wild R., Hammond P.M., Ahrens D., Balke M., Caterino M.S., Gómez-Zurita J., Ribera I., Barraclough T.G., Bocakova M., Bocak L. and Vogler, A. P.. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science. 318(5858):1913-6.

Hunte, C., J. Koepke, C. Lange, Rossmanith, T. and Michel, H. 2000. Structure at 2.3 A resolution of the cytochrome bc(1) complex from the yeast Saccharomyces cerevisiae co-crystallized with an antibody Fv fragment. Structure 8:669-84.

Ingemar Jönsson, K., Rabbow, E., Schill, R. O., Harms-Ringdahl, M. and Rettberg, P. 2008. Tardigrades survive exposure to space in low Earth orbit. Current Biology 18 (17): 729–731

Irimia, M., Maeso, I., Penny,D., Garcia-Ferna`ndez, J. and Roy, S. W. 2007. Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. Mol. Biol. Evol. 24:1604–1607.

Jager, M., Murienne, J., Clabaut, C., Deutsch, J., Le Guyader, H. and Manuel, M. 2006. Homology of arthropod anterior appendages revealed by Hox gene expression in a sea spider. Nature. 441(7092):506-8.

Jobb, G., Von Haeseler, A. and Strimmer, K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4:18.

Jones, D. T., Taylor, W. R. and Thornton J. M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-82

Jones, M., Gantenbein, B., Fet, V. and Blaxter, M. 2006. The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions. Mol Phylogenet Evol. 3(2):583-95.

Kadner, D. and Stollewerk, A. 2004. Neurogenesis in the chilopod Lithobius forficatus suggests more similarities to chelicerates than to insects. Dev Genes Evol 214:367-79.

Kjer, K. M. 2004. Aligned 18S and insect phylogeny. Syst Biol. 53(3):506-14.

Korschelt, E. and K. Heider. 1895. Text-Book of the Embryology of Invertebrates. London: Swan Sonnenschein & Co.

Korschelt, E. and Heider, K. 1892. Lehbuch der vergleichenden Entwicklungsgeschichte der wirbellosen Thiere. Specieller Theil, 2. Heft. Jena: 309-908.

Kosiol, C., I. Holmes, and N. Goldman. 2007. An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464-79.

Kosiol, C., Goldman, N. and Buttimore, N. H. 2004. A new criterion and method for amino acid classification. J Theor Biol 228:97-106.

Kostka, M., Uzlikova, M., Cepicka, I., Flegr, J. 2008: SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics, 9:341.

Kristensen, N. P. 1975. The phylogeny of hexapod 'orders'. A critical review of recent accounts. Z Zool Syst Evol. 13:1–44.

Kristensen, N. P. 1991. Phylogeny of extant hexapods. In The insects of Australia. Vol. 1. 2nd ed. Edited by J.D. Naumann, P.B. Carne, J.F. Lawrence, J.P. Spradbery, R.W. Taylor, M.J. Whitten, and Littlejohn, M.J. Carlton, Victoria, Australia:Melbourne University Press. pp. 125–140.

Kristensen, N. P. 1999. Phylogeny of endopterygote insects, the most successful lineage of living organisms. Eur. J. Entomol. 96: 237–253.

Labandeira, C. C. 1998. Early history of arthropod and vascular plant associations. Annu. Rev. Earth Planet. Sci. 26: 329–377.

Lanave, C., Preparata, G., Saccone, C. and Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93.

Lartillot, N., Brinkmann , H. and Philippe, H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 8;7 Suppl 1:S4.

Lartillot N. and Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria.. Philos Trans R Soc Lond B Biol Sci. 363(1496):1463-72.

Lartillot, N. and Philippe, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-109.

Lartillot, N., Brinkmann, H. and Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 8;7 Suppl 1:S4.

Laslett, D. and Canbäck, B. 2008. ARWEN, a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24:172-175.

Lavrov D.V., Boore, J.L. and Brown, W.M. 2000. The Complete Mitochondrial DNA Sequence of the Horseshoe Crab Limulus polyphemus. Mol. Biol. Evol. 17:813-824.
Lavrov D.V. and Lang B.F. 2005. Poriferan mtDNA and animal phylogeny based on mitochondrial gene arrangements Systematic Biology 54:651-659.

Le S.Q., Gascuel, O. and Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics, 24:2317-2323.

Le, S. Q. and Gascuel, O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-20.

Lio, P. 2005. Phylogenetic and structural analysis of mitochondrial complex I proteins. Gene 345:55-64.

Lio, P., and Goldman, N. 1998. Models of molecular evolution and phylogeny. Genome Res 8:1233-44.

Lio, P., and N. Goldman. 2002. Modeling mitochondrial protein evolution using structural information. J Mol Evol 54:519-29.

Lo, N., Tokuda G., Watanabe, H., Rose, H., Slaytor, M., Maekawa, K., Bandi, C., and Noda, H. 2000. Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. Current Biology 10(13):801-804.

Loesel, R., Nässel, D. R. and Strausfeld, N. J. 2002. Common design in a unique midline neuropil in the brains of arthropods. Arthropod Structure & Development, 31: 77-91.

Lopez, P., Casane, D. and Philippe, H. 2002. Heterotachy, an Important Process of Protein Evolution. Mol. Biol. Evol. 19:1-7.

Lowe, T. M. and Eddy, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955-64.

Machida, R. 2000. Serial homology of the mandible and maxilla in the jumping bristletail Pedetontus unimaculatus Machida, based on external embryology (Hexapoda: Archaeognatha, Machilidae). Journal of Morphology, 245:19-28.

Mallatt, J. & Winchell, C. J. 2002 Testing the new animal phylogeny: first use of combined large-subunit and smallsubunit rRNA gene sequences to classify the protostomes. Mol. Biol. Evol. 19:289–301.

Mallatt, J., and Giribet, G. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phylogenet. Evol. 40:772–794.

Mallatt, J.M., Garey, J.R. and Shultz, J.W. 2004. Ecdysozoan phylogeny and bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phyl. Evol. 31:178-191.

Manton, S. M. 1977. The Arthropoda. Habits, functional morphology, and evolution. Oxford: Oxford University Press.

Margulies, M., Egholm, M., Altman, W.E. et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Martin, J.W. and Davis, G.E. 2001. An updated classification of the recent Crustacea. Science Ser., Nat. Hist. Mus. L.A. Co., No. 39.

Masta, S. E., Longhorn, S. J. and Boore, J. L. 2009. Arachnid relationships based on mitochondrial genomes: asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Mol Phylogenet Evol. 2009 Jan 50(1):117-28

Mayer, G. and P.M. Whitington. 2009a. Neural development in Onychophora (velvet worms) suggests a step-wise evolution of segmentation in the nervous system of Panarthropoda. Dev Biol. 335:263-75. 2.

Mayer, G. and P.M. Whitington. 2009b. Velvet worm development links myriapods with chelicerates. Proc. Biol. Sci. 2009. 276:3571-9.

Maxmen, A., Browne, W.E., Martindale, M.Q. and Giribet, G. 2005. Neuroanatomy of sea spiders implies an appendicular origin of the protocerebral segment. Nature 437(7062):1144-8.

Mittmann, B. and Scholtz, G. 2003. Development of the nervous system in the "head" of Limulus polyphemus (Chelicerata: Xiphosura): morphological evidence for a correspondence between the segments of the chelicerae and of the (first) antennae of Mandibulata. Development Genes Evolution, 213:9-17.

Mooers, A.O. and E.C. Holmes. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15:365-369.

Moura, G. & Christoffersen, M. L. 1996.The system of the mandibulate arthropods: Tracheata and Remipedia as sister groups, "Crustacea" non-monophyletic. J. Comp. Biol. 1:95-113.

Müller, C. H. G., Rosenberg, J., Richter, S. and Meyer-Rochow, V. B. 2003. The compound eye of Scutigera coleoptrata (Linnaeus, 1758) (Chilopoda: Notostigmophora): an ultrastructural reinvestigation that adds support to the Mandibulata concept. Zoomorphology, 122:191-209.

Müller, C. H. G., Sombke, A. and Rosenberg, J. 2007. The fine structure of the eyes of some bristly millipedes (Penicillata, Diplopoda): additional support for the homology of mandibulat ommatidia. Arthropod Structure & Development, 36:463-476.

Müller, K. J. and Walossek, D. 1986. Martinssonia elongata gen. et sp.n., a crustacean-like euarthropod from the Upper Cambrian 'Orsten' of Sweden. Zoologica Scripta, 15:73-92.

Nardi, F., Spinsanti, G., Boore, J. L., Carapelli, A., Dallai, R. and Frati F. 2003. Hexapod origins: monophyletic or paraphyletic? Science 299, 1887–1889.

Negrisolo, E., Minelli, A. and Valle, G. 2004. The mitochondrial genome of the house centipede scutigera and the monophyly versus paraphyly of myriapods. Mol Biol Evol 21:770-80.

Nichols, S.W. 1989. The Torre-Bueno Glossary of Entomology. New York: New York Entomological Society.

Nielsen, C. 1995 Animal evolution. Interrelationships of the living phyla, 1st edn. Oxford, UK: Oxford University Press.

Nielsen, C. 2001 Animal evolution. Interrelationships of the living phyla, 2nd edn. Oxford, UK: Oxford University Press.

Nilsson, D.E. and Kelber, A. 2007. A functional analysis of compound eye evolution. Arthropod Structure & Development, 36:373-385.

Novotny, V., Miller, S. E., Hulcr, J., Drew, R.A, Basset, Y., Janda, M., Setliff, G. P., Darrow, K., Stewart, A. J., Auga, J., Isua, B., Molem, K., Manumbor, M., Tamtiai, E., Mogia, M. and Weiblen G. D. 2007. Low beta diversity of herbivorous insects in tropical forests. Nature 448(7154):692-5.

Papillon, D., Perez, Y., Caubit, X. & Le Parco, Y. 2004 Identification of chaetognaths as protostomes is supported by the analysis of their mitochondrial genome. Mol. Biol. Evol. 21:2122–2129.

Paps, J., Baguna. J. and Riutort, M. 2009. Bilaterian Phylogeny: A Broad Sampling of 13 Nuclear Genes Provides a New Lophotrochozoa Phylogeny and Supports a Paraphyletic Basal Acoelomorpha. Mol Biol Evol. 26: 2397-2406.

Park S.J., Lee Y.S. and Hwang U. W. 2007. The complete mitochondrial genome of the sea spider Achelia bituberculata (Pycnogonida, Ammotheidae): arthropod ground pattern of gene arrangement. BMC Genomics. 1 8:343.

Paulus, H. F. 1979 Eye structure and the phylogeny of the Arhtopoda. In Arthropod phylogeny. Edited by A.P.Gupta Van Nostrand Company.

Perna, N.T, Kocher, T.D. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol. 41:353–35.

Philip, G. K., Creevey, C. J. & McInerney, J. O. 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol. Biol. Evol. 22:1175–1184.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D.J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. and Manuel, M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 19(8):706-12.

Philippe, H. and Laurent, J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616-23.

Philippe, H., and M. J. Telford. 2006. Large-scale sequencing and the new animal phylogeny. Trend. Ecol. Evol. 21:614-20.

Philippe, H., Lartillot, N. and Brinkmann, H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22:1246–1253.

Phillipe, H., Delsuc, F., Brinkmann, H. and Lartillot, N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541-562.

Pisani, D., Poling, L.L., Lyons-Weiler, M. and Hedges, S.B. 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol 2 1.

Pisani, D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst Biol 53:978-89.

Podsiadlowski L, and Braband A. 2006. The complete mitochondrial genome of the sea spider Nymphon gracile (Arthropoda: Pycnogonida). BMC Genomics. 7:284.

Podsiadlowski, L., Braband, A. and Mayer, G.: The Complete Mitochondrial Genome of the Onychophoran Epiperipatus biolleyi Reveals a Unique Transfer RNA Set and Provides Further Support for the Ecdysozoa Hypothesis. Mol Biol Evol 25 (2008):42-51.

Popadic, A., Panganiban, G., Rusch, D., Shear, W. A. and Kaufman, T. C. 1998. Molecular evidence for the gnathobasic derivation of arthropod mandibles and for the appendicular origin of the labrum and other structures. Development Genes and Evolution, 208:142-150.

Posada, D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253-6.

Posada, D., and Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793-808.

Prpic, N.-M. and Tautz, D. 2003. The expression of the proximodistal axis patterning genes Distal-less and dachshund in the appendages of Glomeris marginata (Myriapoda: Diplopoda) suggests a special role of these genes in patterning the head appendages. Developmental Biology, 260, 97-112.

Prpic, N.-M., Wigand, B., Damen, W. G. M. and Klinger, M. 2001. Expression of dachshund in wild-type and Distal-less mutant Tribolium corroborates serial homology in insect appendages. Development, Genes and Evolution, 211:467-477.

Rasnitsyn A. P. 1998. Problem of the basal dichotomy of the winged insects. In: R.A. Fortey and R.H. Thomas (eds.) Arthropod relationships. Systematic Association Special Volume Series 55. Chapman and Hall: 237-248

Regier, J.C., Shultz. J.W. and Kambic, R.E. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc Biol Sci. 272(1561):395-401.

Regier, J. C. et al. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57:920-938.

Regier, J.C., Wilson, H.M. and Shultz, J.W. 2005. Phylogenetic analysis of Myriapoda using three nuclear protein-coding genes. Mol Phylogenet Evol 34:147-58.

Reyes, A., Gissi, C., Pesole, G. and Saccone, C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol 15:957-66.

Richter, S. & Wirkner, C. S. 2004. Kontroversen in der phylogenetischen systematik der Metazoa (eds. Richter, S. & Sudhaus, W.) 73-102 (Sitzungs-Berichte der Gesellschaft naturforschender Freunde zu Berlin, Berlin, 2004).

Richter, S. 2002. The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. Organisms, Diversity & Evolution, 2:217-237.

Roeding, F., Borner, J., Kube, M., Klages, S., Reinhardt, R. and Burmester, T. 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator). Mol Phylogenet Evol. doi:10.1016/j.ympev.2009.08.014

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F. and Philippe, H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56(3):389-99.

Rogozin, I. B., Wolf, Y. I., Carmel, L. & Koonin, E. V. 2007 Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol. Biol. Evol. 24:1080–1090.

Rokas, A. and S.B. Carroll. 2006. Bushes in the tree of life. PLoS Biol. 4(11):e352.

Rota-Stabelli, O. and Telford, M. J. 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol Phylogenet Evol 48:103-11.

Rota-Stabelli, O., Yang, Z. and Telford, M. J. 2008. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol Phylogenet Evol. 52:268-272.

Roure, B., Rodriguez-Ezpeleta, N. and Philippe, H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7 Suppl 1:S2.

Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, U., Bagun˜a`, J. & Riutort, M. 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc. Natl Acad. Sci. USA 99, 11 246–11 251.

Ryu, S.H., Lee, J.M., Jang, K.H., Choi, E.H., Park, S.J., Chang, C.Y., Kim,W. and Hwang, U.W. 2007. Partial mitochondrial gene arrangements support a close relationship between Tardigrada and Arthropoda. Mol Cells. 24(3):351-7.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. and Reyes, A. 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene 238:195-209.

Saccone, C., Gissi, C., Reyes, A., Larizza, A., Sbisa, E. and Pesole, G. 2002. Mitochondrial DNA in metazoa: degree of freedom in a frozen event. Gene 286: 312.

Savard, J., D. Tautz, S. Richards, G. M. Weinstock, R. A. Gibbs, J. H. Werren, H. Tettelin and M. J.Lercher. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. 16(11):1334-8.

Schmidt-Rhaesa, A. 1996. The nervous system of Nectonema munidae and Gordius aquaticus, with implications on the ground pattern of the Nematomorpha. Zoomorphology 116:133–142.

Schmidt-Rhaesa, A. 1998. The position of the Arthropoda in the phylogenetic system. J. Morphol. 238:263–285.

Schmidt-Rhaesa, A., Ehlers, U. Bartolomaeus, T, Lemburg, C. and Garey, J.R. 1998. Then phylogenetic position of the Arthropoda. J. Morphology 238:263-285.

Scholtz, G. & Edgecombe, G. 2006.The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. Dev. Genes Evol. 216:395-415.

Scholz, G., Mittmann, B. and Gerberding, M.: The pattern of Distal-less expression in the mouthparts of crustaceans, myriapods and insects: new evidence for a gnathobasic mandible and the common origin of the Mandibulata. Int. J. Dev. Biol. 42 (1998):801-810.

Schwartz, G. 1978. Estimating the dimension of a model. Annals of Systematics 6 (2):461-464.

Shao, R., Barker, S. C. 2006. Mitochondrial genomes of parasitic arthropods: implications for studies of population genetics and evolution. Parasitology, 134:153-167.

Shimodaira, H. and Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 12467.

Shultz, J.W. and Regier, J.C. 2000. Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. Proc Biol Sci 267:1011-9.

Snodgrass, R. E. 1950. Comparative studies on the jaws of mandibulate arthropods. Smithsonian Miscellaneous Collections, 116:1-85.

Sperling, E.A., Peterson, K.J. and Pisani, D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol. 26:2261:2274.

Stamatakis, A. 2006. Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective". In Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006), Rhodos, Greece.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stollewerk, A. and Chipman, A. 2006. Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships. Integrative and Comparative Biology 46:195-206.

Strausfeld, N. J. 1998. Crustacean-insect relationships: the use of brain characters to derive phylogeny amongst segmented invertebrates. Brain, Behavior and Evolution 52:186-206.

Strausfeld, N. J., Strausfeld, C. M., Stowe, S., Rowell, D. and Loesel, R. 2006. The organization and evolutionary implications of neuropils and their neurons in the brain of the onychophoran Euperipatoides rowelli. Arthropod Structure & Development 35:169-196.

Strausfeld, N.J., Strausfeld, C.M., Loesel, R., Rowell, D. and Stowe, S. 2006. Arthropod phylogeny: onychophoran brain organization suggests an archaic relationship with a chelicerate stem lineage. Proc Biol Sci 273:1857-66.

Telford, M. J. et al. 2000. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. Proc Natl Acad Sci USA. 97 (21):11359-64.

Telford, M. J. and Budd, G. E. 2003. The place of phylogeny and cladistics in Evo-Devo research. Int. J. Dev. Biol. 47:479-90.

Telford, M. J., and Thomas, R. H. 1995. Demise of the Atelocerata? Nature 376, 123-124.

Telford, M. J., Bourlat, S. J., Economou, A., Papillon, D., and Rota-Stabelli, O. 2008. The evolution of the Ecdysozoa. Philos Trans R Soc Lond B Biol Sci 363:1529-37.

Telford, M. J.and Thomas, R. H. 1998. Expression of homeobox genes shows chelicerate arthropods retain their deutocerebral segment. Proceedings of the National Academy of Sciences USA, 95:10671-10675.

Telford, M. J., Gowri-Shankar, V. and Wise, M.J., 2005. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the Bilateria. Mol. Biol. Evol. 22:1129–1136.

Timmermans, M. J., D. Roelofs, J. Mariën and N. M. van Straalen. 2008. Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. BMC Evol. Biol.12:8:83.

Tsukihara, T., Aoyama, H., Yamashita E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R.,Yaono, R. and Yoshikawa, S. 1996. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science 272:1136-44.

von Reumont, B.M., Meusemann, K., Szucsich, N.U., Dell'Ampio, E., Gowri-Shankar, V., Bartel, D., Simon, S., Letsch, H.O., Stocsits, R.R., Luan, Y.X., Wägele, J.W., Pass, G., Hadrys, H. and Misof, B. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. BMC Evol Biol. 2009 May 27;9:119.

Waeschenbach, A., Telford, M. J., Porter, J. S., and Littlewood, D. T. 2006. The complete mitochondrial genome of Flustrellidra hispida and the phylogenetic position of Bryozoa among the Metazoa. Mol Phylogenet Evol 40:195-207.

Wägele, J.W. 1993. Rejection of the "Uniramia" hypothesis and implications of the Mandibulata concept. Zoologische Jahrbucher, Abteilung für Systematik, Ökologie und Geographie der Tiere, 120:253-288.

Waloszek, D., Maas, A., Chen, J. Y. & Stein, M. 2007. Evolution of cephalic feeding structures and the phylogeny of Arthropoda. Palaeogeography Palaeoclimatology Palaeoecology 254:273-287.

Wang, H-C., Li K. M., Susko, E. and Roger, A. J. 2008. A class frequency mixture model that adjust for site-specific amino aicd frequencies and imporves inference of protein phylogeny. BMC Evolutionary Biology 8:331.

Webster, B. L., Copley, R. R., Jenner, R. A., Mackenzie-Dodds, J. A., Bourlat, S. J., Rota-Stabelli, O., Littlewood, D. T. and Telford, M. J. 2006. Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. Evol Dev 8:502-10.

Webster, B.L., Mackenzie-Dodds, J.A., Telford, M.J. and Littlewood, D.T. 2007. The mitochondrial genome of Priapulus caudatus Lamarck (Priapulida: Priapulidae). Gene 389:96-105.

Wheeler, W. C., M. Whiting, Q. D. Wheeler and Carpenter, J. M. 2001. The phylogeny of the extant hexapod orders. Cladistics 17:113-169.

Whelan, S., and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. MBE 18:691-699.

Whitfield, J. B. and Kjer, K.M. 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. Annu Rev Entomol. 53:449-72.

Whiting, M. F., Carpenter, C. J., Wheeler, Q. D. and Wheeler, C. W. 1997. The Strepsiptera Problem: Phylogeny of the Holometabolous Insect Orders Inferred from 18S and 28S Ribosomal DNA Sequences and Morphology Systematic Biology 46(1):1-68.

Whitworth, T. L., Dawson, R. D., Magalon, H. and Baudry, E. 2007. DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). Proceedings of the Royal Society B. 274:1731-1739.

Wiegmann, B. M., Trautwein, M. D., Kim, J. W., Cassel, B. K., Bertone, M. A., Winterton, S. L. and Yeates, D. K. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 24:7:34.

Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. 2004 Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.

Wolff, C. and Scholtz, G. 2006. Cell lineage analysis of the mandibular segment of the amphipod Orchestia cavimana reveals that the crustacean paragnaths are sternal outgrowths and not limbs. Frontiers in Zoology 2006, 3:19.

Yang, A. S. 2001. Modularity, evolvability, and adaptive radiations: a comparison of the hemi- and holometabolous insects. Evolution and Development 3(2): 1-14.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution, 11, 9:367-372.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-91.

Yang, Z., and Nielsen, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568-79.

Yang, Z., Nielsen, R. and Hasegawa, M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-11.

Yeates, D. K. and Wiegmann, B. M. 2005. Phylogeny and evolution of Diptera: recent insights and new perspectives, in Yeates, D. K. and B. M. Wiegmann, eds. The Evolutionary Biology of Flies. Columbia University Press.

Zhang, X. G., Siveter, D. J., Waloszek, D. and Maas, A. 2007. An epipodite-bearing crown-group crustacean from the Lower Cambrian. Nature 449:595-598.

Zhou, Y., Rodrigue, N., Lartillot, N. and Philippe, H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol. 2007 Nov 1;7:206.

Zwickl D. J. and D.M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology: 51(4): 588-598

Appendix 1: Anatomical evidences in support of Mandibulata

The following list of putative apomorphic characters in support of the Mandibulata has been complied by Greg Edgecombe from the NHM London during a collaborative analyses. In summation they represent a large body of complex detail from numerous anatomical systems.

1. Mandible

i. Position: The mandible is the appendage of the post-tritocerebral segment, embedded in a chewing chamber between the clypeolabrum and hypopharynx (Wägele 1993; Bitsch 2001).

ii. Similarity of gnathal edge: Mandibles have their gnathal edge modified for chewing. In the three mandibulate groups the gnathal edge is differentiated into a dentate incisor part and a molar part with a surface formed from rows of fused spines (Edgecombe *et al*. 2003). Evidence from musculation (Snodgrass 1950) and ontogenetic development (Machida 2000) that the gnathal edge of the mandible is a coxal endite is corroborated by gene expression (see below).

iii. Gene expression:

a. *Distal-less*: Mandibles are unique among gnathal appendages in showing a gradient of decreasing *Distal-less* expression through ontogeny. *Distal-less* is expressed only in the palp when a palp is present, or may have a transient expression in crustacean and myriapod taxa that lack a palp (Popadić *et al*. 1998; Scholtz *et al*. 1998). In contrast to other gnathal appendages, *Distal-less* expression is wholly lacking along the inner margin of the mandible, i.e., in the area corresponding to the gnathal edge.

b. *Dachschund*: Expression of *Dachschund* is characteristically strong in the area corresponding to the tooth-like parts of the mandible in myriapods (*Glomeris*: Prpic and Tautz 2003), hexapods (*Tribolium*: Prpic *et al.* 2001) and crustaceans (*Porcellio*: Abzhanov and Kaufman 2000) and this gene appears to specify mandibular identity (Prpic and Tautz 2003). In the homologous appendage of chelicerates (leg 1), *Dachshund* has an expression that instead indicates a role in patterning the proximal-distal axis (Abzhanov and Kaufman 2000), as in locomotory legs throughout the Arthropoda.

The suggestion that the mandible could be a basal character of euarthropods that secondarily reverses to an unmodified locomotory limb in chelicerates (Cook *et al*. 2001; Prpic and Tautz 2003) forces an outstanding degree of reversal in all the details listed above, which collectively show that the mandible is a profoundly modified coxal endite that functions in a specialised chewing chamber.

2. Head segmentation

i. Composition and appendages: Mandibulates have the appendage of the deutocerebral segment modified as an antenna. The antenna is variably identified as apomorphic for Mandibulata (Scholtz and Edgecombe 2006) or a symplesiomorphy inherited from stem-group euarthropods (Waloszek *et al*. 2007). The mandibulate head consists of antennal/antennular, intercalary/second antennal, mandibular, first maxillary, and labial/second maxillary segments. The boundary of the head capsule behind the second maxillary segment, i.e., shared possession of five appendage-bearing head segments, defines crown-group mandibulates relative to fossil taxa in the mandibulate stem group that have only four appendage-bearing head segments (Zhang *et al*. 2007).

ii. Hox expression: Compared to the more generally broadly overlapping Hox expression domains in the prosoma of chelicerates, myriapods display a trend towards the narrow expression domains that more precisely unite hexapods and crustaceans (Hughes and Kaufman 2002). The same pattern manifests itself in the trunk, in which *Antennapedia* expression is restricted from the posterior of the embryo in mandibulates, but strongly expressed throughout the opisthosoma in chelicerates (Hughes and Kaufman 2002).

3. Sternal buds of mandibular segment

A classical hypothesis that the paragnaths of crustaceans could be homologous with the hypopharyngeal superlinguae observed in various groups of hexapods and myriapods is reinforced by new studies of the development of paragnaths (Wolff and Scholtz 2006). Mandibulates generally share paired lateral buds on the mandibular sternum that give rise to either the paragnaths or components of the hypopharynx. Such sternal anlagen on the posterior stomodaeal region are reasonably identified as homologous in and apomorphic for Mandibulata.

4. Differentiation of first maxillae as a mouthpart

By outgroup comparison to the undifferentiated locomotory limb in the corresponding position in chelicerates and onychophorans, the shared presence in all mandibulates of a first maxilla as a gnathal appendage is synapomorphic. The millipede *Glomeris* demonstrates that the gnathal identity of the maxilla is expressed at the molecular level by *Distal-less* being expressed where sensory organs (primordia of the sensory palps) form, rather than this gene having a role in proximal-distal axis patterning as it does in the antenna and locomotory limbs (Prpic and Tautz 2003).

5. Brain anatomy

Brain morphology defends a grouping of chilopods, hexopods and crustaceans to the exclusion of chelicerates and onychophorans (Loesel *et al*. 2002; Strausfeld *et al*. 2006a, b), although other myriapods (Diplopoda) do not share the putative mandibulate apomorphies. Specific neural characters that serve as putative autapomorphies of Mandibulata include:

i. A conserved midline neuropil is embedded in the protocerebral matrix rather than lying superficial to the protocerebrum as in chelicerates and onychophorans (Strausfeld *et al*. 2006a). This neuropil is uniquely lacking in diplopods.

ii. The central body of the brain has a unique neuropil named midline neuropil 2 by Loesel *et al*. (2002) in chilopods, hexapods and crustaceans.

iii. The somata that supply cerebral neuropils are variable in size in chilopods, hexapods and crustaceans but are uniform in chelicerates and onychophorans (Strausfeld *et al*. 2006a).

iv. The deutocerebrum contains the olfactory lobe. This contrast with protocerebral olfactory glomeruli in onychophorans and variable positions in chelicerates depending on which appendage is equipped with olfactory receptors (Strausfeld *et al*. 2006a).

v. The stomatogastric and labral nerves are connected to the tritocerebrum, rather than to the deutocerebrum as in onychophorans and chelicerates (Scholtz and Edgecombe 2006) according to the view that the cheliceral segment is homologous with the antennal segment, i.e., the cheliceral segment is deutocerebral (Telford and Thomas 1998; Mittmann and Scholtz 2003). An alternative view in which the cheliceral segment is identified as tritocerebral rather than deutocerebral (Bitsch and Bitsch 2007) posits that the stomatogastic ganglia are invariably connected to the tritocerebrum throughout Euarthropoda.

6. Ommatidial ultrastructure

Special similarity in the ommatidium of mandibulates is informed by re-description of the compound eyes of scutigeromorph centipedes (Müller *et al*. 2003) and penicillate diplopods (Müller *et al*. 2007). Details that are apomorphic for Mandibulata are:

i. A crystalline cone is developed in the dioptric apparatus, as in hexapods and crustaceans. The crystalline cone is *Scutigera* is composed of four cones cells (Müller *et al*. 2003), precisely as in general condition for the common ancestor of crustaceans and hexapods according to the Tetraconata hypothesis (Dohle 2001; Richter 2002). Functional speculations that the cone of scutigeromorphs is convergent with that of Tetraconata (Nielsen and Kelber 2007) do not overrule the morphological arguments for their homology (Müller *et al*. 2007).

ii. The lateral eyes of scutigeromorphs and penicillates have dozens of cells in each subunit and, although cell numbers are variable, some individual cells (e.g., cone cells and proximal retinula cells) can be identified (Harzsch *et al*. 2005). This is intermediate between the low, fixed cell numbers shared by hexapods and crustaceans and the higher, more variable cell numbers in chelicerates.

iii. Interommatidial pigment cells in scutigeromorphs share detailed similarity with those of crustaceans and hexapods (Müller *et al*. 2003). Correspondences include longitudinal extension of the cell bodies, distal positioning of the nuclei, the cytoplasm absorbing pigment granules, and the specific mode of attachment of the cornea and basement membrane (Müller *et al*. 2003).

7. Serotonin-reactive neurons in ventral nerve cord

i. Myriapods, hexapods and crustaceans share a reduced and more fixed number of serotonergic neurons than are observed in chelicerates (Harzsch 2004), in which clusters of ca 10 somata are present. In mandibulates, cells are individually identifiable and typically developed singly or in pairs, to a maximum of four neurons in a group. The reduced, more stable number may be viewed as apomorphic for Mandibulata (Harzsch *et al*. 2005).

ii. A specific apomorphic character shared by chilopods, hexapods and entomostracan crustaceans is a posterior pair of serotonergic neurons with neurites that cross to the contralateral side (Harzsch 2004).