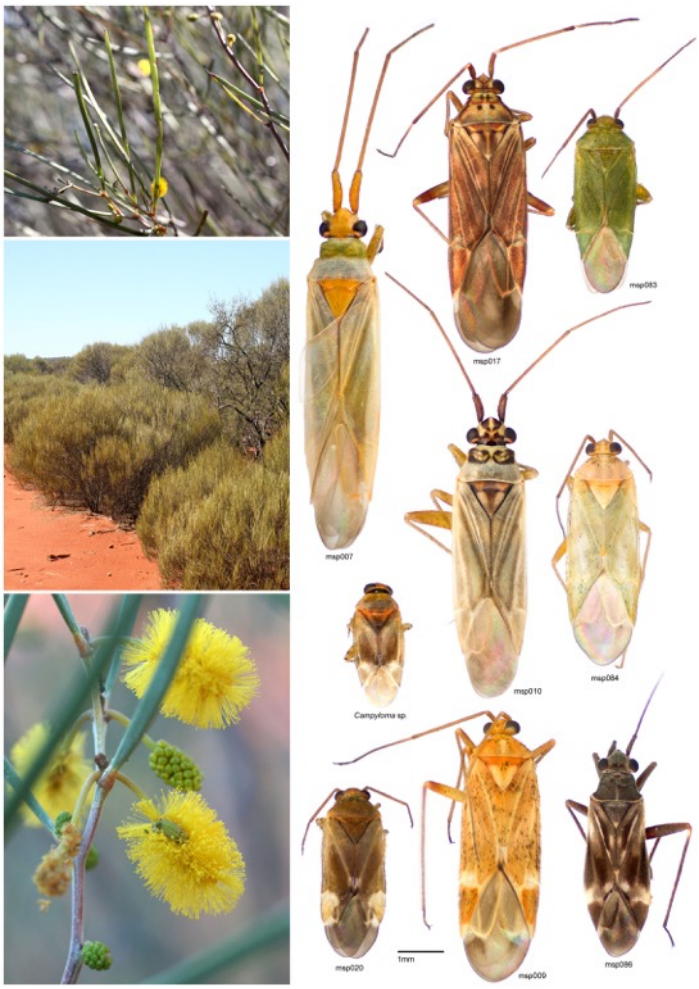
FACULTY OF SCIENCE

School of Biological, Earth and Environmental Science

BIOS3221

ASSEMBLING THE TREE OF LIFE



Course Coordinator: Prof. Gerry Cassis

A course designed and taught by scientists from the University of New South Wales, the Australian Museum and Royal Botanic Gardens (Sydney) and

2022

Table of Contents

Faculty of Science - Course Outline

1. Information about the Course

And BEES website: https://www.bees.unsw.edu.au/study-us/courses/assembling-tree-life

Year of Delivery	2022
Course Code	BIOS3221
Course Name	Assembling the Tree of Life
Academic Unit	School of Biological, Earth and Environmental Science
Level of Course	3rd UG; elective
Units of Credit	6UOC
Session(s) Offered	T3
Assumed Knowledge, Prerequisites or Co-requisites	Assumed knowledge: HSC science (including biology); English language proficiency. Understanding of biological principles appropriate for a 3rd year level course. Prerequisites: BIOS1101
Commencement Date	Week 1: 16 September 2022

Summary of Course Structure (for details see 'Course Schedule')

Component	HPW	Time	Day	Location
Lecture	1	-	All lectures available on course commencement	Moodle, pre-recorded
Practicals	4	1-5 pm	Friday	TeachLab3 Ground (K-E26-G001)

2. Staff Involved in the Course

Role	Name	Contact Details – room and email	Consultation Times
Course Convener	Prof. Gerry Cassis	Mathews 1314 gcassis@unsw.edu.au	Email for appointment
Additional Course lecturers	Dr Frank Koehler	Australian Museum frank.koehler@Australian.Museum	Email for appointment
	Dr Chris Reid	Australian Museum Chris.Reid@Australian.Museum	Email for appointment
	Dr Richard Jobson	Royal Botanic Gardens (Sydney) Richard.Jobson@rbgsyd.nsw.gov.au	Email for appointment
	Prof. Simon Ho	University of Sydney simon.ho@sydney.edu.au	Email for appointment

3. Course Details

Course Description (Handbook Entry)	Systematics investigates historical aspects of evolution and establishes evidence-based classifications and genealogical relationships between organisms. Phylogenetic systematics also known as cladistics provides a basis for hierarchical classification and a framework for examining other evolutionary and biological events and phenomena, such as historical biogeography, sexual coevolution, and host-parasite coevolution. This course is designed to introduce the principles and application of phylogenetic systematics, using a variety of organisms. The practicals will place emphasis on the use of computer software and examples from major clades of the Tree of life.
Course Aims	This course is designed to train undergraduate students in the principles and application of phylogenetic systematics. Students will learn about the conceptual basis of comparative biology, using morphological and molecular data. Coupled with this conceptual framework students will learn how to determine character homologies, construct and interpret phylogenies by hand and using the latest software programs (TNT, Mesquite, BioEDIT and MEGA).
Student Learning Outcomes	By the end of the course students should have achieved the following outcomes: 1) Understand the core concepts in phylogenetic systematics (characters, taxa, phylogeny). 2) Understand the nature of morphological and molecular data. 3) Undertake and interpret phylogenetic analyses. 4) Understand importance of fieldwork and collections 5) Be able to understand the processes and material used in identifying taxa. 6) Interpret and evaluate scientific ideas and express them in the Term oral presentation. 7) Knowledge of the above elements will be assessed through practicals, field trip and oral presentation. There is no final exam – assessment is continuous.

Science Graduate Attributes	Select the level of FOCUS *0 = NO FOCUS* *1 = MINIMAL* *2 = MINOR* *3 = MAJOR*	Activities / Assessment
Research, inquiry and analytical thinking abilities	*3*	The course assignments necessitate an understanding of theory and practice of systematics, and the ability to synthesise and analyse given scientific data and literature.
Capability and motivation for intellectual development	3	Despite the fact that there is only one Tree of Life, its determination requires the ability to resolve conflicts in character information and assess alternative theories. Student participation in course discussions will enhance their interest in the course materials and the relevant scientific literature.
Ethical, social and professional understanding	3	The role of systematics and the Tree of Life in society will be an ongoing theme in this course. Students will be encouraged to make this connection and contribute to a discussion of this nexus. In addition, an understanding of the role of systematics in evolutionary theory will be encouraged.
Communication	3	The course will require communication in written and oral formats. Communication between lecturers and fellow students is necessary to maximise learning outcomes.
Information literacy	3	Students are provided with citations to reference material. Students are expected to be able to evaluate the scientific information and theories and present their findings in the compulsory assignments. Standard citation of literature and references is essential.

Major Topics	The major topics that will be taught in this course will be: 1) Fundamental principles in systematics, including those in taxonomy and phylogenetics. 2) Core principles of comparative biology, including the use of comparative methods. 3) In depth examination of characters, character states and homology. 4) Detailed examination of the nature of morphological and molecular data. 5) Practical demonstration and exercises in the use of phylogenetic tree-building algorithms. 6) Theory and practice of how to evaluate phylogenetic results, particularly when competing solutions are found. 7) Divergence dating.
Relationship to Other Courses within the Program	The Tree of Life course bears a close relationship with Evolution BIOS3071 and Biology of Invertebrates (BIOS2031). In this course - BIOS3221 - the focus is on macroevolutionary patterns. For those students who have taken the former two courses you will have had some exposure to key systematics concepts. In this course, these will be explored in much greater depth with an emphasis on hands-on training in the use of phylogenetic techniques.

4. Rationale and Strategies Underpinning the Course

Teaching Strategies	The teaching strategy involves a weekly unit of one lecture coupled with a practical on the Friday of weeks 1-4 and 6-10, with the subject matter interrelated. The lectures serve as the theoretical component of the course and the practicals are primarily concerned with applications. In the practicals the students will be taught **how to undertake phylogenetic analysis**. This will involve exposure to different phylogenetic software packages. The student oral presentation will involve selecting a topic of the list provided. These teaching strategies are designed so that students at the end of the course can understand the why, when and how of phylogenetics.
Rationale for learning and teaching in this course	BIOS3221 has been designed to be a hands-on style course, where students will be engaged with all course materials. It has a strong emphasis on providing the theoretical underpinnings of systematics, but always coupled with applications.

5. 2022 Course Schedule

Date	Week		Time and place	Lecture/Lab topic
Fri 16 Sept	Week 1	Lecture 1	Moodle, recorded	Course introduction and core systematic concepts
Fri 16 Sept	Week 1	Practical 1	Teaching lab 3, 1-5 pm	Collections and systematics
Fri 23 Sept	Week 2	Lecture 2	Moodle, recorded	Similarity, characters and character states
Fri 23 Sept	Week 2	Practical 2	Teaching lab 3, 1-5 pm	Proteaceae lab 1
Fri 30 Sept	Week 3	Lecture 3	Moodle, recorded	Homology and character polarity
Fri 30 Sept	Week 3	Practical 3	Teaching lab 3, 1-5 pm	Proteaceae lab 2
Fri 7 Oct	Week 4	Lecture 4	Moodle, recorded	Phylogenetic reconstruction and parsimony
Fri 7 Oct	Week 4	Practical 4	Teaching lab 3, 1-5 pm	Mesquite/TNT Computer lab
Fri 14 Oct	Week 5	Lecture 5	Moodle, recorded	Molecular phylogenetics lecture 1
Fri 14 Oct	Week 5	Practical 5	Teaching lab 3, 1-5 pm	Molecular phylogenetics lab 1
Sat 15-19 Oct	Week 6	Field trip	Smiths Lake	Insect collections and phylogenetics
Fri 28 Oct	Week 7	Lecture 6	Moodle, recorded	Molecular phylogenetics lecture 2
Fri 28 Oct	Week 7	Practical 6	Teaching lab 3, 1-5 pm	Molecular phylogenetics lab 2
Fri 4 Nov	Week 8	Lecture 7	Moodle, recorded	Divergence dating
Fri 4 Nov	Week 8	No practical	No practical	No practical
Fri 11 Nov	Week 9	Lecture 8	Moodle, recorded	Animal phylogenetics
Fri 11 Nov	Week 9	Practical 7	Teaching lab 3, 1-5 pm	Student presentations
Fri 18 Nov	Week 10	Lecture 9	Moodle, recorded	Insect phylogenetics
Fri 18 Nov	Week 10	No practical	No practical	No practical

6. Assessment Tasks and Feedback

Task	Knowledge & abilities assessed	Assessment Type	% of total mark	Date of		Feedback		
				Release	Submission	WHO	WHEN	HOW
Practical exercises								
Practical 2 (Proteaceae 1)	Phylogenetic methods – plant lab 1	Written exercises	5	23 September	30 September	Cassis	7 October	Marks/comments
Practical 3 (Proteaceae 2)	Phylogenetic methods – plant lab 2	Written exercises	10	30 September	7 October	Cassis	14 October	Marks/comments
Practical 4 (TNT/Mesquite)	Phylogenetic methods – software	Written exercises	10	7 October	14 October	Cassis	28 October	Marks/comments
Practical 5 (Molecular phylogenetics 1)	Molecular methods	Written exercises	10	14 October	28 October	Cassis	4 November	Marks/comments
Practical 6 (Molecular phylogenetics 2)	Molecular methods	Written exercises	10	28 October	4 November	Cassis	11 November	Marks/comments
Field trip report								
Smiths Lake	Knowledge of systematics and entomology	Insect collection, phylogenetics, group work	35	15 October	19 October	Cassis	19 October	Marks/comments
Student presentation								
Oral presentation	Knowledge of systematics	Oral presentation	20	16 September	11 November	Cassis	18 November	Marks/comments

7. Additional Resources and Support

Course Manual	The course manual will be made available through **UNSW Moodle**.
Additional Readings	During Term your lecturers may assign additional readings and will be listed on **UNSW Moodle** and students will be notified in lectures.
Societies	If students are interested in phylogenetics beyond the course content, especially if they are considering undertaking an Honours of postgraduate degree in systematics, phylogenetics, taxonomy or biogeography, the following societies and their journals are worth exploring: 1) Willi Hennig Society (journal: *Cladistics*) 2) Society of Systematic Biologists (journal: *Systematic Biology*)
Computer Laboratories or Study Spaces	The software for the practicals are available through the UNSW MyAccess portal. The software that you will need to access are: TNT, Mesquite, BioEdit and MEGA.

8. Required Equipment, Training and Enabling Skills

Enabling Skills Training Required to Complete this Course	All students will be required to have competent writing and PowerPoint presentations skills, as well as being computer literate (e.g., Microsoft Office, standard internet skills).

9. Administration Matters

Information about each of the following matters is best presented in a generic School handout or webpage. Reference should be made in every course handout to where the information can be found, and the importance of being familiar with the information.

Expectations of Students	You are expected to attend the scheduled practicals. The lectures are pre-recorded and will be made available at the commencement of the course. Any alterations to the schedule will be announced via Moodle. Although systematics is arguably the oldest of the biological disciplines it is always subject to revision both theoretically and practically. Students are expected to understand the dynamics and future developments of the discipline. In this sense, rote memory is not a recommended strategy to learning. Students need to comprehend and be able to articulate theory and applications. In the process, they should have a deeper understanding of the history of life and the Tree of Life as we know it, including controversial points. Students are given a week to complete practical exercises. Notwithstanding, the practicals are the best time to take advantage of an instructor's time, where you are encouraged to interpret results with their input.
Assignment Submissions	Due dates, word-lengths, percentages, and the staff member responsible are shown in the schedule. Assignments are due as per the Table on page 8. **They must be submitted at UNSW Moodle BIOS3221**. **LATE SUBMISSIONS WILL REQUIRE SPECIAL CONSIDERATION APPLICATIONS.**
Assessment Procedures	Assessment procedures are in the table above. Assessment task and feedback'. If you have a problem which affects your work, you should immediately apply for special consideration.
Equity and Diversity	Those students who have a disability that requires some adjustment in their teaching or learning environment are encouraged to discuss their study needs with the course Convenor prior to, or at the commencement of, their course, or with the Equity Officer (Disability) in the Equity and Diversity Unit (http://www.studentequity.unsw.edu.au/) Issues to be discussed may include access to materials, signers or note-takers, the provision of services and assessment arrangements. Early notification is essential to enable any necessary adjustments to be made.

10. UNSW Academic Honesty and Plagiarism

What is Plagiarism?

The following definition of plagiarism is given at the UNSW website: https://student.unsw.edu.au/what-plagiarism

"**Plagiarism** at UNSW is using the words or ideas of others and passing them off as your own. Plagiarism is a type of intellectual theft.

Plagiarism can take many forms, from deliberate cheating to accidentally copying from a source without acknowledgement Consequently, whenever you use the words or ideas of another person in your work, you must acknowledge where they came from."

PLEASE READ CAREFULLY THE UNSW POLICY ON PLAGIARISM. DO NOT PLAGIARISE – THERE ARE SERIOUS CONSEQUENCES FOR PLAGIARISING AND YOUR SUBMISSIONS WILL BE CHECKED!

INTRODUCTION

The course 'Assembling the Tree of Life' (BIOS3221) is one of only a handful of courses currently offered in Australia on systematic methods and applications, with an emphasis on phylogenetic reconstruction.

This course is complementary to the other biological sciences course, particularly those concerning evolutionary biology. The teaching of systematics, incorporating biological classification and phylogenetics, is poorly represented on all campuses in the Sydney Basin. Moreover the 2007 'National Forum on the Future of Taxonomy in Australia', funded by the Commonwealth Government, identified this shortfall in tertiary training (https://www.science.org.au/support/analysis/decadal-plans-science/discovering-biodiversity-decadal-plan-taxonomy) and that it was of strategic importance to promote systematics through new undergraduate courses in Australian universities, with the aim of training the next generation of systematists in Australia. You are now participating in a course that addresses this shortfall!

Systematics is the science of biological classifications based on the determination of relationships of taxa. Classifications are the cornerstone of organismal biology and provide a hierarchical organisation of species based on their genealogical relationships. As such classifications/phylogenies provide a framework for generating and testing hypotheses in evolutionary biology, biogeography and conservation biology.

The course is designed to integrate three major themes: systematics methods, current theories and controversies concerning the evolutionary relationships of organisms (= Tree of Life), and natural history. To accommodate these themes, each in-session week comprises one lecture interlocked with a practical. Students will produce a student presentation in Week 9.

This course has been designed in conference with research scientists from the Australian Museum, Royal Botanic Gardens and the University of New South Wales.

The course assessment will involve practical exercises, the Smith Lake field trip exercises and the oral presentation.

There is no exam for this course.

ASSESSMENT IS CONTINUOUS.

LATE SUBMISSIONS WILL REQUIRE SPECIAL CONSIDERATION APPLICATIONS.

LECTURERS

In this course you will have a diversity of lecturers from the Australian Museum, the Royal Botanic Gardens (Sydney), and the University of New South Wales. You will have exposure to entomologists, malacologists, and botanists. Although each of their taxa have specialised research needs, all of these scientists apply the same systematic methods.

Your lecturers are as follows (arranged in lecture order as per the schedule):

Prof. Gerry Cassis (gcassis@unsw.edu.au) of the University of Sydney. He is a world leader in molecular phylogenetics and divergence dating.

Dr Richard Jobson (Richard.Jobson@rbgsyd.nsw.gov.au) is a systematic botanist whose research incorporates theoretical and practical aspects of systematic botany and the broader uses to which phylogenetic knowledge may be applied.

Dr. Frank Köhler (Frank.Koehler@Australian.Museum) is a systematic zoologist with special interest in the systematics, phylogeny and evolution of non-marine gastropods.

Dr Chris Reid (Chris.Reid@Australian.Museum) is a systematic zoologist who works on beetle systematics and natural history.

Prof. Simon Ho (simon.ho@sydney.edu.au) of the University of Sydney. He is a world leader in molecular phylogenetics and divergence dating.

LECTURES

Lectures are provided as Powerpoint online files, which can be viewed and/or downloaded from Moodle.

The following pages provide a summary of each lecture content. There are nine lectures, finishing in week 10.

Lecture 1 (Week 1) - Course outline, core phylogenetic concepts

Prof. Gerry Cassis (UNSW)

Pre-recorded (download from Moodle)

The 'Tree of Life' course focuses on the principles and methods for determining the **identity** and **relationships** of all organisms – the science of **systematics**. Systematics is arguably the oldest of the biological sciences and yet it has undergone a revolution in the past 50 years with the advent of **cladistic methodology** and **DNA sequence data and model-based analytical techniques**. Systematics (inclusive of the descriptive science of **taxonomy**) has never been more relevant as our species strives to document all of life on Earth in the face of the biodiversity crisis. There has been a surge of effort in understanding the evolutionary relationships at all scales of the taxonomic hierarchy, from the macroevolutionary patterns of the major clades of the 'Tree of Life' to species-level relationships.

This lecture gives an outline of the course and the assignments, as well as core principles of systematics.

Lecture 2 (Week 2) – Similarity, characters and character states

Dr Richard Jobson - Royal Botanic Gardens (Sydney)

Pre-recorded (download from Moodle)

"Similarity" is a common-sense concept that is an integral aspect of human perception as well as being fundamental to comparative biology. Pre-scientific approaches to taxonomy relied on the holistic, intuitive assessment of similarity between organisms in establishing classifications. Science, however, demands that we be able to specify and quantify similarity for communication and hypothesis testing. Systematists do this by atomising similarities and differences between parts of different organisms as characters. In phylogenetic methodology, characters are heritable variables composed of two or more discrete states (binary or multistate) that distinguish different taxa. At the outset of a study they are usually not polarised and, if they have more than two states, these are usually unordered – they are said to be primary hypotheses of homology. The character states of primary homologies can be conceptualized as potentially transformed versions of one another. The classical criteria (or tests) for recognising attributes of different organisms as potentially homologous are:

1. similarity of structure (including similarity of composition and developmental origin)
2. similarity of relative position

What if these two criteria disagree (homeosis)? Neither test is as rigorous as the test of congruence, which will be dealt with in the next lecture.

Lecture 3 (Week 3) – Homology and character polarity

Dr Richard Jobson - Royal Botanic Gardens (Sydney)

Pre-recorded (download from Moodle)

"Homology" was originally defined by Richard Owen in 1843 as "the same organ in different [organisms] under every variety of form and function". The concept of homology subsequently acquired a phylogenetic interpretation with the acceptance of the theory of evolution: "a character state of two or more taxa is homologous if this character state is found in the common ancestor of these taxa, or, two character states (or a linear sequence of character states) are homologous if one is directly (or sequentially) derived from the others". In the previous lecture we saw how the homology of characters and their states is weakly tested by similarity of composition and position. A more rigorous test of homology is the test of congruence, in which data from different characters are cladistically analysed to discover whether they agree with each other in specifying the same phylogenetic relationships amongst taxa. Incongruence between different characters indicates that at least some of our primary hypotheses of homology are false (or that the phylogenetic model we have assumed is false). The criterion of parsimony is used to arbitrate between conflicting hypotheses of relationships: the cladogram requiring the lowest number of postulated character state transformations (and thus the fewest mistaken hypotheses of homology) is preferred. Shared character states that have passed the test of congruence are said to be secondary (corroborated) hypotheses of homology.

Conventional cladistic analysis yields unrooted trees that have no time dimension. Any node or branch on such a tree could conceivably be its base and all of the character state transformations reconstructed along its branches are not polarised and could have proceeded in either direction. The method that is most commonly used to root trees and thus polarise characters (determine ancestral or plesiomorphous states from derived or apomorphous states) is outgroup comparison. This involves appealing to higher level phylogenetic analyses that have already been conducted to identify the root of our ingroup of interest. The main limitation of this method is that when pursued to its logical conclusion, it gives us an unrooted tree of life that needs to be rooted using some other reasoning than outgroup comparison. A method of character analysis that does not rely on pre-existing trees, and which can be applied in some circumstances, involves the direct observation of the nested relationship between a less general character state and its more general homologue. The simplest such case involves an ontogenetic transformation from a more general to a less general state, in which case it is most parsimonious to conclude that the more general state is ancestral and the less general derived. Similar logic can be applied to relations between more and less general serial homologues and to nucleotide sites shared by molecular paralogues.

Lecture 4 (Week 4) – Phylogenetic reconstruction and parsimony

Prof. Gerry Cassis (UNSW)

Pre-recorded (download from Moodle)

Under the principle of Maximum Parsimony, we strive to reproduce the evolutionary history of organisms by creating cladograms of the shortest length (as measured by character state changes across the tree). This is no easy task. Estimating phylogenies is computationally difficult, as there are often many possible tree topologies to choose from, and it becomes computationally impossible to derive a direct solution once we include more than 18 taxa (e.g. for 4 taxa, there are 15 possible rooted trees, for 10 taxa there are 34,459,425 possible trees). As a result, phylogenetic analyses (regardless of methodology) rely on complex heuristic (i.e. short-cut) methods to speed up tree searches.

In this lecture we will discuss how to sift through the near-infinite forest of possible topologies in tree space, in order to arrive at a most parsimonious solution. We will also discuss how the relative importance of different characters is determined, and how support for sister-relationships, overall topologies and alternative solutions are measured through tree statistics and resampling methods.

Lecture 5 (Week 5) – Molecular Phylogenetics 1

Dr Frank Köhler - Australian Museum

Pre-recorded (download from Moodle)

Sequence alignment is a way of arranging nucleotide or amino acid sequences as rows within a matrix by identifying regions of similarity, which may be a consequence of functional, structural or evolutionary relationships. Gaps may be inserted between the residues so that identical or similar characters are aligned in successive columns. The resulting character matrix is required for subsequent phylogenetic analyses, such as Maximum Parsimony, Maximum Likelihood or Bayesian Inference algorithms.

If two sequences in an alignment share a common ancestor, mismatches in their nucleotide sequence can be interpreted as point mutations introduced in one of the sequences in the time since they diverged from one another. These mutations may have either caused a nucleotide to be replaced by another (**transition** or **transversion**) or by the loss or gain of a nucleotide (**Indel**) being expressed by an alignment gap.

Sequence alignments are a homology statement; nucleotides in each row of a character matrix are hypothesized to represent different states of the same character (i.e., at this nucleotide position). Alignments of sequence data can have 5 such possible character states (A, T, C, G for the four different nucleotides, or a gap for an Indel); alignments of amino acid sequences have 21 potential character states (20 amino acids plus gap).

While very short or similar sequences can be aligned by hand, most alignments are produced by computerised algorithms. One differentiates between **local** and **global** alignments with respect to lengths of compared sequences (i.e., **partial** or complete) and between **pairwise** and **multiple** alignments.

In **structural alignments** the secondary structure of the analysed gene (portion) is taken into consideration. For the alignment of protein-coding genes the codon usage may also provide clues as to how nucleotides are most credibly aligned.

Methods, applications and problems of sequence alignment are discussed. Once a data matrix has been produced, it can be subjected to phylogenetic analyses. Specifics of sequence data, with respect to assigning differential costs to various types of changes in maximum parsimony analyses (**weighting**) are discussed.

Readings

Essential reading:
Giribet, G. & W. C. Wheeler. 1999. On gaps. *Molecular Phylogenetics and Evolution* 13:132-143.
Recommended reading:
Wheeler, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Systematic Biology 44:321-331.

Lecture 6 (Week 7) – Molecular Phylogenetics II

Dr Frank Köhler - Australian Museum

Pre-recorded (download from Moodle)

Parsimony is the analytical method most often applied to morphological data, however for molecular data model-based methods are more commonly used. This is because much is known about the patterns and processes of evolution of DNA and protein sequences, and this knowledge can be used to improve our methods of estimating phylogenies from molecular data.

For example, we know that transition base substitutions occur much more frequently (i.e. they have a higher rate of evolution) than transversion base substitutions. Transitions are therefore more likely to be homoplasious (due to multiple hits) than transversions, so it makes sense to give more weight to transversions than to transitions. However the optimal weighting will vary among datasets, depending on many factors. Instead of making arbitrary decisions on a weighting scheme, we can use a simple mathematical model to estimate the transition/transversion ratio from the data set, and apply this model in a phylogenetic analysis.

There are two main classes of methods that use models of sequence evolution:
1) Distance methods and 2) Character methods.

1) Distance Estimates attempt to estimate the mean number of changes per site since 2 taxa last shared a common ancestor based upon a model of how the sequences may have evolved, i.e. they correct for "multiple hits."
2) Character methods include parsimony, maximum likelihood (ML) and Bayesian methods. Although distance models are often based upon some of the same assumptions as the models in ML they are implemented in a very different way.

Models of sequence evolution

- Jukes Cantor model (JC): assumes all changes equally likely
- Kimura 2-Parameter (K2P): allows different rates for transitions and transversions
- General Time Reversible model (GTR): assigns different probabilities to each type of change
- LogDet /Paralinear distance model: was devised to deal with unequal base frequencies in different sequences

All of these models include a correction for "multiple hits" (multiple substitutions superimposed on the same site in a gene). All (except Logdet/Paralinear distances) can be modified to include corrections for site rate heterogeneity and invariant sites.

Distance Methods

One major approach to tree building is to use a measure of distance between taxa, joining together those taxa that are closest to each other (a distance matrix is calculated from the original character matrix, so some information is lost in the process).

Two factors spurred the early development of methods based distances:

- The existence of data which were intrinsically distances, so couldn't use character data (e.g., DNA-DNA hybridization)

- Molecular clock: It was noticed early on (1960's) that if one examines the degree of amino acid sequence divergence among species with a known phylogeny:
    - one sees approximately equal rates of evolution, and
    - the degree of divergence more or less indicates recency of common ancestry.

Early methods applied UPGMA clustering (Unweighted Pair-Group Method using Arithmetic mean) to various measures of sequence similarity. UPGMA is very sensitive to differential rates of evolution and produces bogus results under these circumstances. A confusing profusion of methods was developed to deal with this problem: Neighbour-Joining (NJ) is one such method, subtly but significantly different from UPGMA.

Although very fast, distance methods suffer from a number of deficiencies, therefore statistical methods such as Maximum Likelihood and Bayesian Inference were developed to provide more rigor. Their downside is that they are computationally expensive.

Maximum Likelihood (ML) and Bayesian Inference (BI)

ML: The tree with the highest likelihood is the better one, i.e., the one that gives the highest probability of observing the (sequence) data, given the model.

Probability, in technical sense, is something that events or experimental outcomes have, whereas likelihood is something that only theories or hypotheses or models have.

The greatest advantage of likelihood-based methods is the ability to deal with the **long-branch problem**.

Bayesian analysis has a number of advantages over ML, the most obvious of which is that estimating support for the tree is straightforward. However, it is argued that Bayesian Posterior Probabilities overestimate support relative to ML bootstrap values.

Readings

Essential Reading:
Baldauf, S.L. (2003). Phylogeny for the faint of heart: A tutorial. Trends in Genetics 19 (6): 345-351.
Recommended Reading:
Whelan, S., Liò, P. and Goldman, N. (2001). Molecular phylogenetics: state-of-the art methods for looking into the past. Trends in Genetics 17(5): 262-272.

Lecture 7 (Week 8) – Divergence dating

Prof. Simon Ho – University of Sydney

Pre-recorded (download from Moodle)

The estimation of evolutionary time-scales is pivotal to a wide range of biological studies. It provides a temporal context for our analyses and interpretations, allowing us to study rates of speciation, population divergence, and other evolutionary processes. Traditionally, evolutionary timescales have been estimated using the fossil record, with absolute dates provided by radiometric analysis. Fossils can provide an estimate of when different groups of organisms first appeared and when various lineages diverged from each other. Such data are often unavailable, however, making it necessary to look elsewhere for a means to estimate evolutionary timescales.

Dating the divergences of lineages can be done by analysing DNA, using methods based on the "molecular clock". This hypothesis, proposed in 1962, states that the rate of molecular evolution is approximately constant among lineages. A corollary of the molecular clock is that the genetic difference between two species is proportional to the time since they diverged from their most recent common ancestor. This relationship allows us to estimate evolutionary timescales using genetic data, provided that we have an estimate of the rate of molecular evolution.

Molecular clocks have provided important insights into major evolutionary events, including the radiations of animal phyla and flowering plants. Our current understanding of human evolution has also been greatly informed by analyses of genetic data. However, molecular clocks can be drastically affected by calibration errors and rate variation among lineages.

Over the past decade, there have been numerous developments in this field, allowing the molecular evolutionary process to be modelled with increasing sophistication. In this talk, I will provide an overview of the current state of the field.

Readings

Essential Reading:
Chapters 13 and 14 in Bromham, L. (2016). An Introduction to Molecular Evolution and Phylogenetics. Oxford University Press.

Recommended Readings:

Donoghue, P.C.J., and Yang, Z. (2016) The evolution of methods for establishing evolutionary timescales. Philosophical Transactions of the Royal Society of London B, 371: 20160020.

Lee, M.S.Y., and Ho, S.Y.W. (2016) Molecular clocks. Current Biology, 26: R387–407.

Bell, C.D. (2015) Between a rock and a hard place: Applications of the "molecular clock" in systematic biology. Systematic Botany, 40: 6–13.

Lecture 8 (Week 9) – Animal Phylogenetics

Prof. Gerry Cassis - UNSW

Pre-recorded (download from Moodle)

The phylogeny of animals has a long phylogenetic history. This lecture will provide information from the traditional morphological view of the monophyly of higher groups to the most recent phylogenetic reconstructions. Key characters are discussed.

Lecture 9 (Week 10) – Insect Phylogenetics

Prof. Gerry Cassis - UNSW

Pre-recorded (download from Moodle)

The phylogeny of insects has a vexed history. This lecture will provide information from the traditional morphological view of the monophyly of higher groups to the most recent phylogenetic reconstructions. Competing hypotheses are compared and discussed. Key morphological characters and molecular data are discussed. Details will also given on the impact of evolutionary development data, as well as brain and visual system. The lecture will conclude by the argumentation for nesting of insects within the Pancrustacea.

PRACTICALS

PRACTICALS ARE FACE-TO-FACE IN E26 TEACHING LAB 3.

PLEASE FOLLOW COVID-19 RULES AS FOLLOWS:

When on campus, please ensure that you follow the rules and guidance below to help keep our UNSW community safe:

- Do not come to campus if you have tested positive for COVID-19, have any symptoms at all or are unwell.
- If you are a household or close contact of a COVID-19 positive case, you should work or study from home for seven days from the last time someone in your household tested positive. If you do need to come to campus for any reason, you must do a RAT test prior and wear a mask at all times while on campus.
- Masks are mandatory on public transport.
- We strongly encourage you to wear masks in indoor settings on campus, including in offices, classrooms, labs, communal spaces and especially where physical distancing cannot be maintained.
- Keep a safe distance from others, and if you can't, we strongly recommend that you wear a good quality mask – N95, P2 or surgical mask.
- Wash or sanitise your hands regularly.

PRACTICAL ASSIGNMENTS

Attendance to all practicals is compulsory.

Each Practical has exercise notes and data files that can be downloaded from Moodle.

Practical 1 is not assessable but will have exercises that you will need to complete. This introduces key concepts that are necessary for understanding the lectures, practicals and field trip that follow.

Practicals 2, 3, 4, 5 and 6 are assessable. You will need to submit your answers electronically via UNSW Moodle.

Practical 7 is in week 9 of the course. This session is designed for the student individual oral presentations.

LATE SUBMISSIONS WILL REQUIRE SPECIAL CONSIDERATION APPLICATIONS.

Hereafter the practical notes for each week are given in full. It is advisable to read these practical exercise notes prior to commencement of the practical.

The practical notes are given in both the Course Manual (which is uploaded to the first tab of the course on Moodle) and each practical tab in Moodle.

There may be adjustments to the practical notes and exercises. Check regularly for announcements.

PRACTICAL 1 (Week 1) – Collections and systematics

Prof. Gerry Cassis – UNSW

E26 Teaching Lab 3

Practical 1 is a series of 10 exercises for you to undertake, accompanied by videos and other materials for you to view, read and analyse. The questions in this lecture will be discussed in the practical. **There is no assessment for Practical 1.**

Exercise 1 – Collections and their uses

Natural History collections are the backbone of Taxonomy and Systematics. There are about 1000 museums in the world that house animal collections, and there are over 3000 herbaria that house plant and fungal collections. The herbaria collections house over 387 million specimens. Universities also house collections, and at UNSW we have an herbarium with about 22,000 specimens, and I am responsible for the insect collection, that includes specimens belonging to UNSW.

These collection institutions are both large and small. The Natural History Museum in London, one of the oldest natural history collections and largest, houses over 80 million specimens. Watch the following video on the natural history collection

NATURAL HISTORY MUSEUM (LONDON)

https://www.youtube.com/watch?v=6eIRuuUVnbk

Please watch two other videos, one from the Australian Museum, and the other from National Herbarium of Victoria (note there is no suitable video available from the NSW Herbarium).

THE AUSTRALIAN MUSEUM

https://www.youtube.com/watch?v=qxIsyOUIJZ8

ROYAL BOTANIC GARDENS AND HERBARIUM OF VICTORIA

https://www.youtube.com/watch?v=gbNs1NN4W9U

QUESTION 1: From these videos what do you see that museums and herbaria have in common? What are the uses of the collections? What role does taxonomy play in society?

Exercise 2 – Gall wasps and Alfred Kinsey

Building collections is not an end in itself, it is a means to an end. That is, collections serve the research task of building natural classifications of living organisms. That is under the theory that there is one history of life on earth, and as a result all organisms are inter-related. This means that we can reconstruct the evolutionary history of life, which is represented in hierarchical classifications and codified through biological nomenclature (scientific names arranged in a hierarchy).

Alfred Kinsey was a very famous entomologist, who became a pioneering sexologist, undertaking US nationwide surveys of human sexual behaviour. The Kinsey reports became highly influential in the US, uncovering aspects of social life that were relatively unknown at the time. You can read about the extraordinary life of Kinsey in the wiki: https://en.wikipedia.org/wiki/Alfred_Kinsey

His wasp work was equally remarkable. He travelled throughout the US and Mexico collecting gall forming wasps, mostly from oaks (more than half of the world's oak species are endemic to Mexico!). I want you to watch this video and understand what the collections reveal, including the natural history of the cynipid wasps.

https://www.youtube.com/watch?v=lHc5l4gQsro

One of his most famous synthetic works on cynipid wasps can be found at:

http://digitallibrary.amnh.org/handle/2246/1148

On page 298 of this paper, you will see the description of a new species, *Aulacidea annulata*. You will notice that this scientific name is composed of two words, the genus name and the species name (a bit like your family name in reverse!).

QUESTION 2: Examining Kinsey's cynipid paper, analyse the layout and content of the species description of *Aulacidea annulata*. In the treatment of this species, there is also what is called a *differential diagnosis* – can you point out which paragraph it is in. Why do you think it is called a differential diagnosis, and how does it differ from the *description*?

Exercise 3 – Classification – taxonomic description and diagnosis

The above species description, with its two names (genus and species), are referred to as *binomial nomenclature*. This was the invention of the great Swedish biologist, Carl Linnaeus, who in 1758, published the 10th edition of a book called the *Systema Naturae*, and is the starting point of zoological nomenclature.

Here is an extract from the *Systema Naturae*, which describes key differences between genera of bugs.

II. HEMIPTERA.

195. CICADA *Roſtrum* inflexum. *Pedes* poſtici ſaltatorii.
196. NOTONECTA *Roſtrum* inflexum. *Pedes* poſtici natatorii (ciliati.)
197. NEPA *Roſtrum* inflexum. *Pedes* antici capitis cheliferi.
198. CIMEX *Roſtrum* inflexum. *Pedes* curſorii.
199. APHIS *Roſtrum* inflexum. *Abdomen* bicorne.
200. CHERMES *Roſtrum* pectorale. *Pedes* poſtici ſaltatorii.
201. COCCUS *Roſtrum* pectorale. *Abdomen* poſtice ſetoſum maribus.
202. THRIPS *Roſtrum* obſoletum. *Alæ* incumbentes abdomini reflexiſ

QUESTION 3: The above bug genus-group treatments are in the form of a *taxonomic key*, that differentiates the genera by the mouthparts (*Roftrum*), the legs (*Pedes*), abdomen (*Abdomen*) and Wings (*Alae*). What do you think the differences are between Linnaeus' key and Kinsey's key to the cynipid wasp genus *Andricus* (see below)? [Hint: take into account the format of each key]

Woody Galls, Mexican Species of *Andricus*

1. Head or thorax or both with some reddish coloring, abdomen without any black; parapsidal grooves complete.....2.
   Head and thorax entirely black, ovipositor sheaths black; parapsidal grooves obliterated posteriorly.....*A. championi* Ashmead.

2. Cubitus continuous to the basal vein.....3.
   Cubitus not reaching the basal vein.....5.
3. Length 3.5 mm. or more; foveæ at base of the scutellum are very large, broad, and rugose.....4.
   Length 3.0 mm. or less; two small foveæ at the base of the scutellum; antennæ entirely rufous.....*A. montezumus* Beutenmüller.
4. Antennæ dark rufous; median groove continuous; areolet very large; length 4.0–5.0 mm.....*A. dugesi* Beutenmüller.
   Antennæ with first two and last few joints bright rufous, the other joints almost black; median groove scarcely perceptible in the rugosities of the thorax; areolet only moderately large; length 3.5–4.2 mm. . .*A. peredurus*, new species.
5. Length under 3.2 mm.; thorax entirely hairy; no median groove, but deep and continuous parapsidal grooves; abdomen with hairs on the sides at the base. *A. furnaceus*, new species.
   Length 4.0 mm.; median and parapsidal grooves indistinct; abdomen smooth (head and thorax not described as pubescent). .*A. durangensis* Beutenmüller.

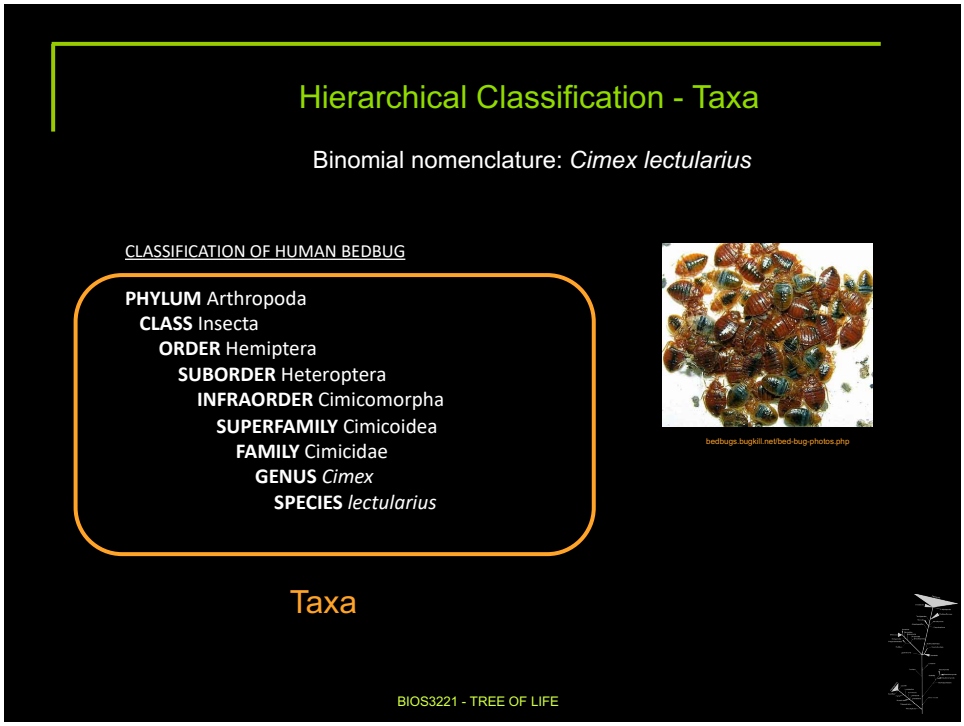
Exercise 4 – Linnean Classification – taxonomic hierarchy

Biological classifications are hierarchical – you could envisage them as nested like a set of Russian dolls. The fundamental unit in classifications is the *species*. Species names are subordinate to generic names, and those in turn to family, order, class, phylum kingdom categories respectively. All of these formal categories are referred to as *taxa*. Their hierarchical arrangement is called *Linnean classification.*

Read the biological classification explainer in the conversation:

https://theconversation.com/explainer-what-is-biological-classification-10691

Let's look at a realworld example, the Human Bed Bug – *Cimex lectularius*



The species *Cimex lectularius* is a worldwide pest species in human dwellings, feeding on our blood, with the blood meal necessary for the development of eggs. *Cimex* is a genus composed of 17 species, with *C. lectularius* and *C. hemipterus* as pests of humans. The other *Cimex* species feed on bats and birds. All species of *Cimex* are ectoparasites and are blood feeders.

Cimex belongs to the family Cimicidae which has a worldwide distribution, which comprises about 110 species. Most species feed on bats and birds. All genera and species of Cimicidae are ectoparasites and are blood feeders.

Watch the video outlining aspects of *Cimex lectularius*:

https://www.youtube.com/watch?v=PfuJ8iIMSKY

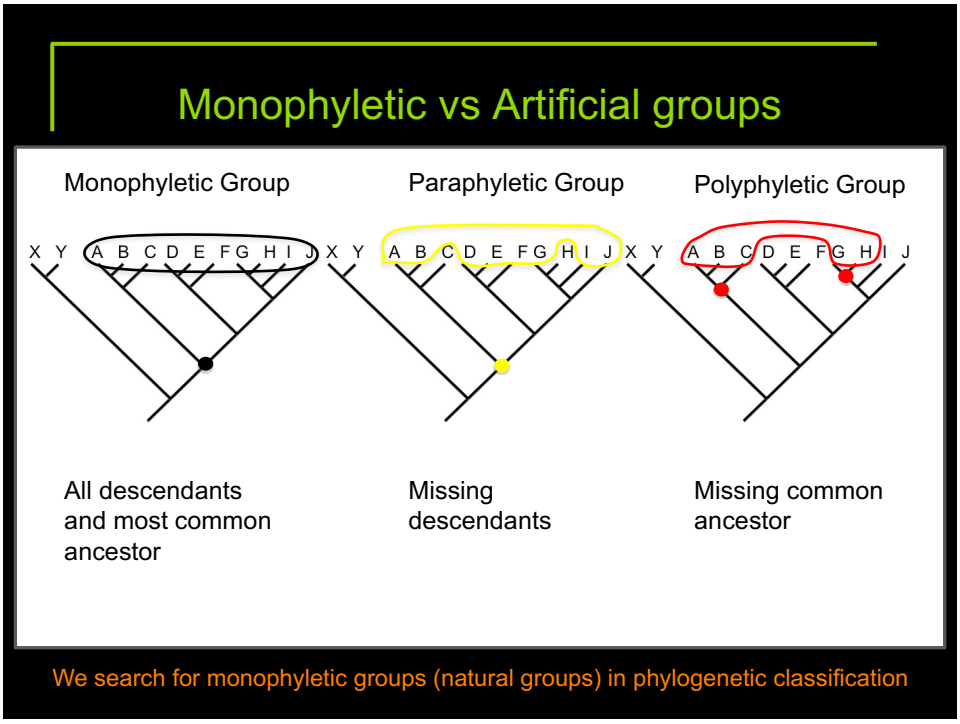
QUESTION 4: What are the advantages and information content of an hierarchical system of classification? Use *Cimex lectularius* as your example.

Exercise 5 – Classifications and relationships

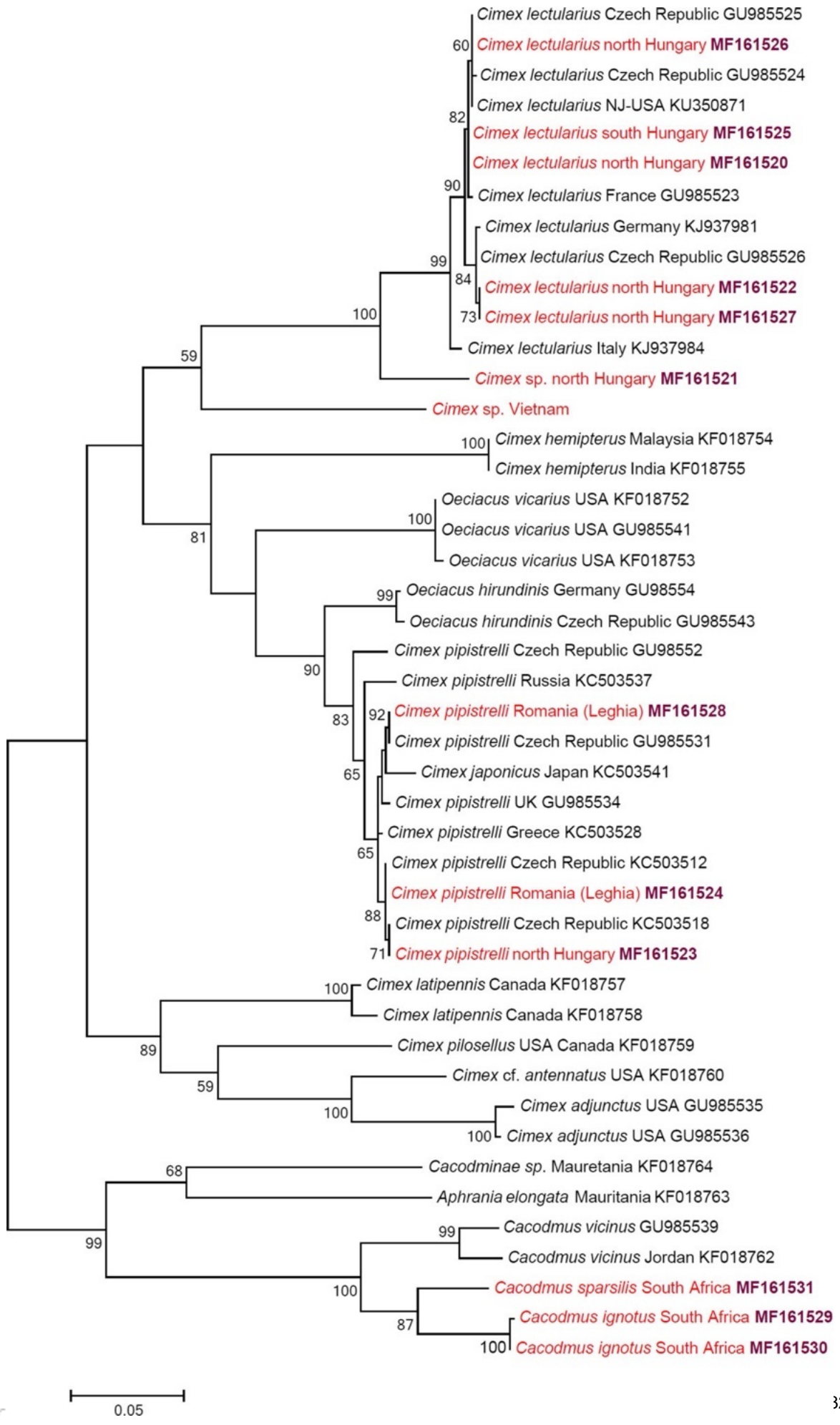
You will notice in the Kinsey paper that there are numerous genera and species that are described. On Page 293 of the paper (see below), he states that many of the genera are not well defined and indicates that the recognition of natural genera will require phylogenetic analysis

> Certain genera of the Cynipidæ are founded upon definite morphological characters which are clearly paralleled by biological considerations. But many of the species of oak gall producing Cynipidæ have long been held in groups which are based on the most meager of indefinite morphological characters, and the "genera" thus made are not confirmed by a more careful examination of the morphology and a study of the biology of the species concerned. And, moreover, until both of the alternate generations of dimorphic species can be included by a generic definition, the group remains an artificial creation. In another paper, on the phylogeny of the Cynipidæ, I am discussing this question in more detail and offering data which may be used to draw lines for natural genera. I

What is implied in Kinsey's statement is that genus-groups are determined by phylogenetic analysis. In Lecture 1, I refer to *natural* and *artificial* classifications. I also refer to Willi Hennig who wrote that classifications need to reflect the evolutionary process. Hennig recognised that natural taxa need to be *monophyletic* – that is a *clade*, which is defined by all the *descendants* and the *most common ancestor*.



QUESTION 5: The following phylogenetic is a recent analysis of the Cimicidae (Hornok et al. 2017) based on molecular data. Name a genus that is monophyletic and why? What is the issue with *Oecacius* in this phylogeny?


Cimex lectularius Czech Republic GU985525
60
Cimex lectularius north Hungary MF161526
Cimex lectularius Czech Republic GU985524
Cimex lectularius NJ-USA KU350871
82
Cimex lectularius south Hungary MF161525
Cimex lectularius north Hungary MF161520
90
Cimex lectularius France GU985523
Cimex lectularius Germany KJ937981
99
Cimex lectularius Czech Republic GU985526
84
Cimex lectularius north Hungary MF161522
100
73
Cimex lectularius north Hungary MF161527
Cimex lectularius Italy KJ937984
59
Cimex sp. north Hungary MF161521
Cimex sp. Vietnam
100
Cimex hemipterus Malaysia KF018754
Cimex hemipterus India KF018755
Oeciacus vicarius USA KF018752
100
Oeciacus vicarius USA GU985541
81
Oeciacus vicarius USA KF018753
99
Oeciacus hirundinis Germany GU98554
Oeciacus hirundinis Czech Republic GU985543
Cimex pipistrelli Czech Republic GU98552
90
Cimex pipistrelli Russia KC503537
83
92
Cimex pipistrelli Romania (Leghia) MF161528
Cimex pipistrelli Czech Republic GU985531
65
Cimex japonicus Japan KC503541
Cimex pipistrelli UK GU985534
Cimex pipistrelli Greece KC503528
65
Cimex pipistrelli Czech Republic KC503512
Cimex pipistrelli Romania (Leghia) MF161524
88
Cimex pipistrelli Czech Republic KC503518
71
Cimex pipistrelli north Hungary MF161523
100
Cimex latipennis Canada KF018757
Cimex latipennis Canada KF018758
89
Cimex pilosellus USA Canada KF018759
59
Cimex cf. antennatus USA KF018760
100
Cimex adjunctus USA GU985535
100
Cimex adjunctus USA GU985536
68
Cacodminae sp. Mauretania KF018764
Aphrania elongata Mauritania KF018763
99
99
Cacodmus vicinus GU985539
Cacodmus vicinus Jordan KF018762
100
Cacodmus sparsilis South Africa MF161531
87
Cacodmus ignotus South Africa MF161529
100
Cacodmus ignotus South Africa MF161530
0.05

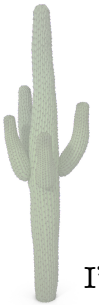
Exercise 6 – Characters and classifications

In this next exercise we will introduce you to four then five species, and a series of characters, and how the combination of taxa and characters are used to determine the phylogenetic relationships of organisms.

Character 1

I see "spines".

Character 2

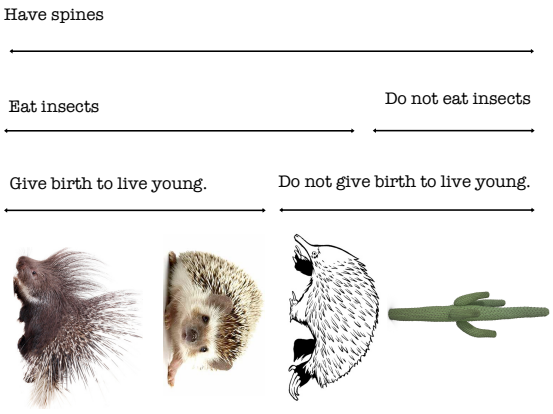


I've seen them eat insects.

Character 3

I've seen them give birth to live young.

Classification 1



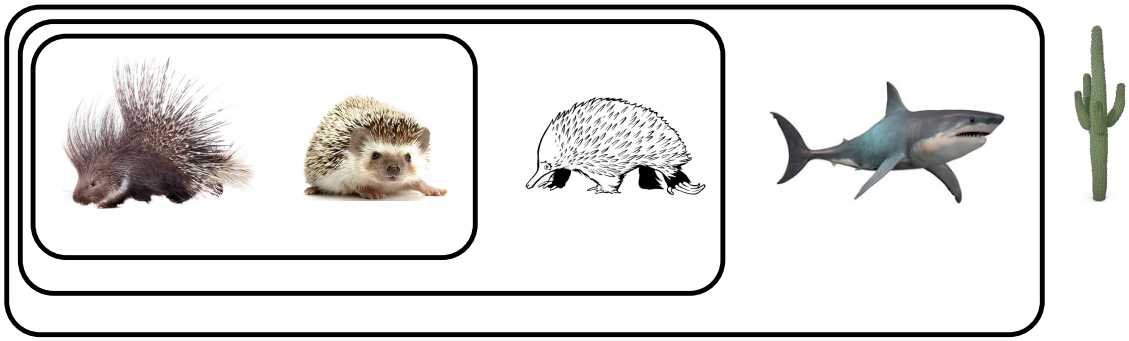
Introduce a new taxon – SHARK!!!

The Shark has the following characters gives birth to live young but does not have spines or eat insects

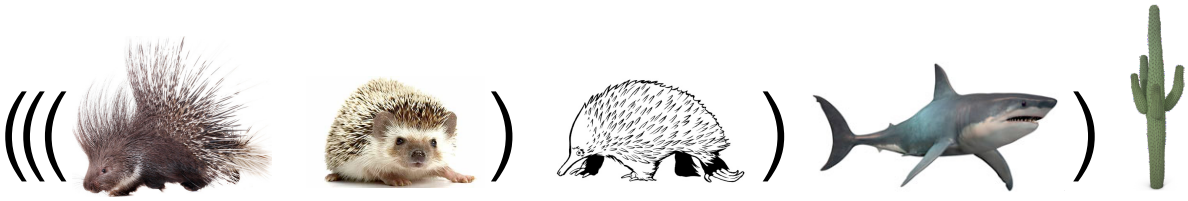
QUESTION 6: Prepare a solution on the board for the give taxa

Exercise 7 – Relationships

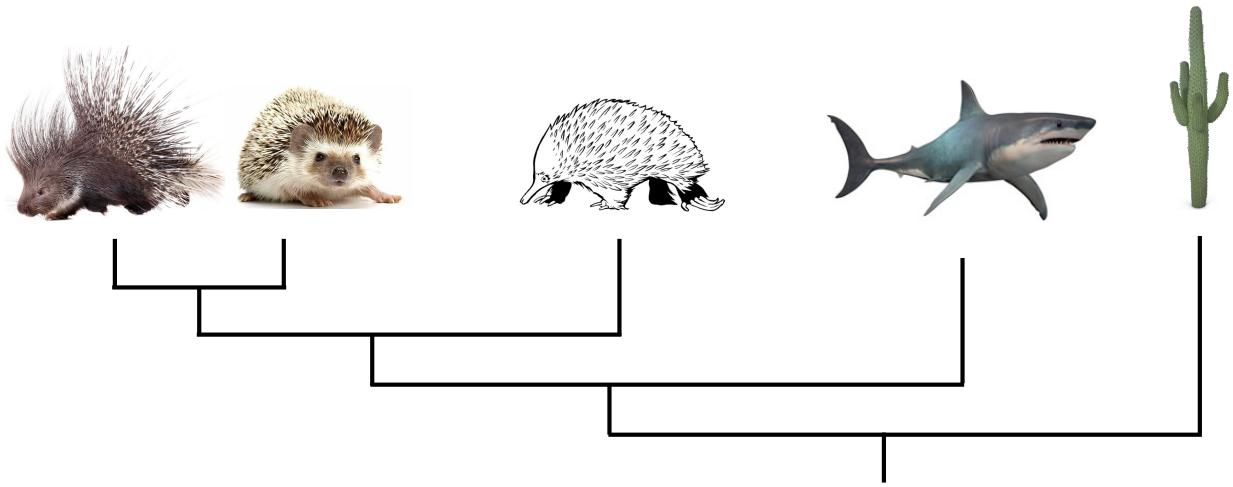
The relationships of the taxa can be expressed as a *Venn diagram* as in the photograph below. The way to read these relationships, is starting with the two organisms (Porcupine and Hedgehog) in the most inside box – these two taxa are most closely related. The Echidna is then most closely related to the Porcupine+Hedgehog; the shark is then most related to the other three taxa (the mammals); and finally, the Cactus is then the most distant relative, and the sister taxon to all of them..



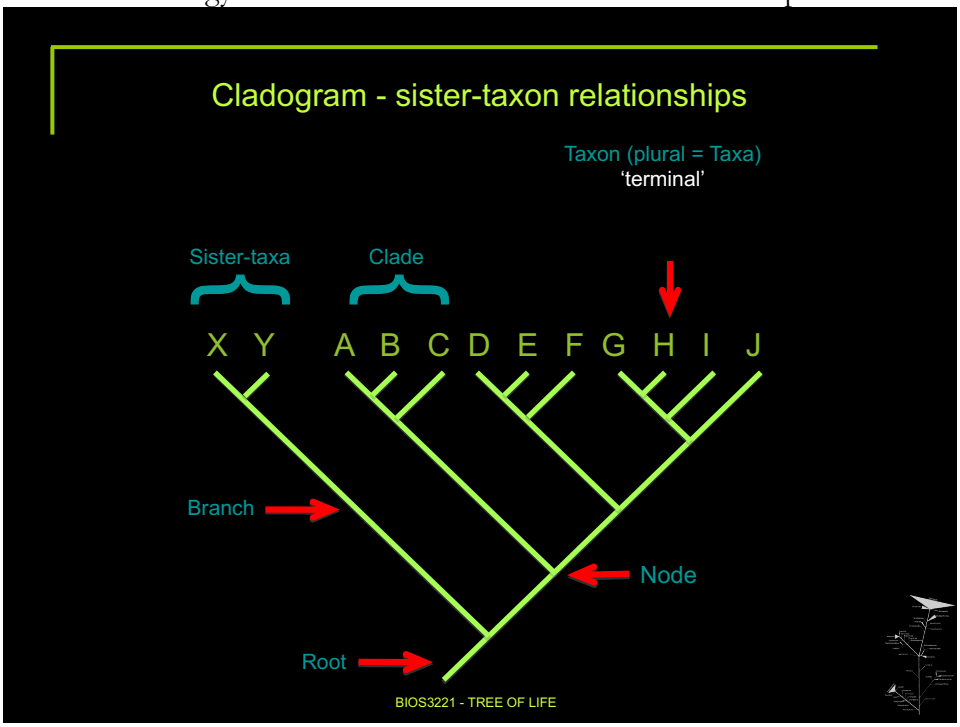
You can also depict the relationships in what is called *Parenthetical Form*, as follows:



Or finally, in *Phylogenetic Tree Form*.



In this course we will most communicate about relationships based on trees. It is therefore important to understand the terminology associated of trees. Here are some of the important features:



The terms are as follows: (a) clade or monophyletic group (ie all descendants and common ancestor); (b) terminal is a taxon at the tip of the tree; (c) sister taxa refers to most closely related taxa; (d) node refers to the meeting of two or more branches, referring to a hypothetical ancestor; (e) branch refers to the line connecting nodes or a node to a terminal; (f) root refers to the base of the tree.

QUESTION 7: Apply all of these terms to our example of the Cactus and four other taxa.

Exercise 8 – Caminalcule Phylogenetics

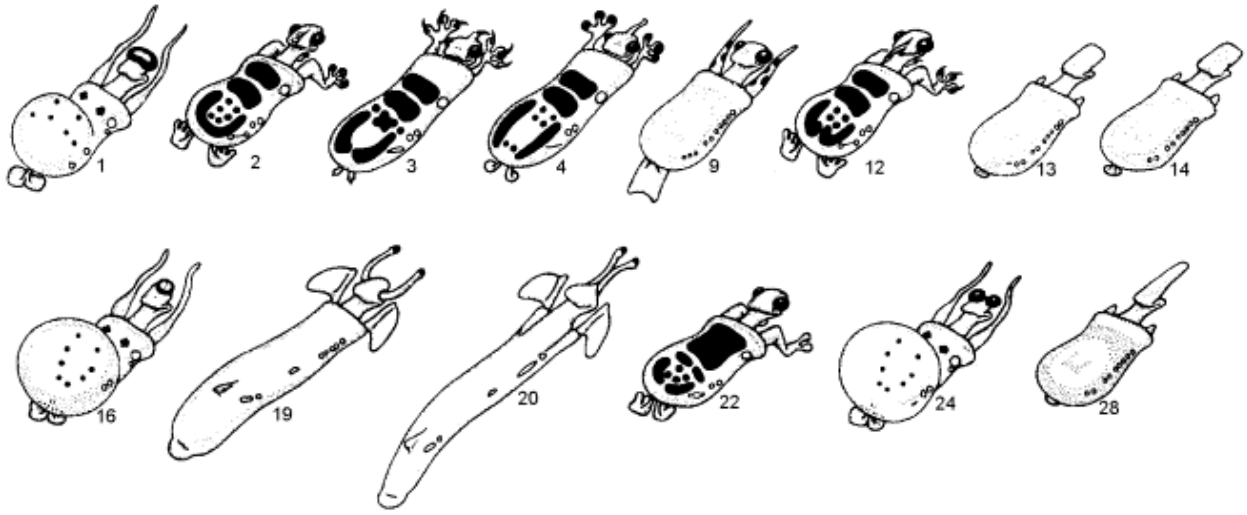
Phylogenetics has been used in lots of disciplines – including in the classical example of imaginary animals called Caminalcules.

QUESTION 8: Build a phylogenetic tree of these 14 Caminalcules by hand, using the following characters and character states in this order:

(a) body with or without front legs or tentacles; (b) body with or without large black markings; (c) head with or without ornaments or processes; (d) head with or without a spike; (e) head with or without a nose fin; (f) head with a narrow or fat body; (g) eyes fused or separated; (h) body with small spots or without small spots ; (i) rear legs fused or separated; (j) tentacles with or without spots; (k) eyes absent or present; ((l) head with or without tentacles; (m) front legs greatly reduced or elongate, with fingers or flaps.

r

Start by scoring each of your 14 Caminalcules by each of the characters in an excel spreadsheet.



QUESTION 9: Construct Venn diagrams of the sister-taxa based on your spreadsheet codification. Do not try to connect all the Venn diagrams at once. Give your Venn diagrams accounting for all 14 Caminalcules.

QUESTION 10: Connect your Venn diagrams, into a **Super Venn** diagram. From this Venn Diagram give the result in **parenthetical form** and as a **phylogenetic tree**. In the tree illustrate with the terms given in Question 7. On the tree write the characters at each node.

There is no wrong or right answer in terms of the relationships – there are literally millions of possibilities given 14 taxa. The point of this exercise is to understand key concepts and how characters and taxa are used to construct a phylogenetic tree. Have fun with this!!

PRACTICAL 2 (Week 2) – Proteaceae lab 1

Dr Richard Jobson – Royal Botanic Gardens NSW

E26 Teaching Lab 3

Plant morphology, character and character state formulation and data matrix

Practical 2 (Week 2) and Practical 3 (Week 3) are a pair of practicals that focus on the flowering plant family Proteaceae. This family includes the African genus *Protea*, the Australian native genus *Banksia* and macadamias (genus *Macadamia*) as well as almost 80 other genera, with a total of about 1500 species.

In practicals 2 and 3 you will test aspects of this classification by assembling and phylogenetically analysing a morphological data set for a small sample of species of Proteaceae.

In this practical you will key out each of the specimens to genus using the dichotomous key in Weston (2007). In the process you will familiarise yourself with some of the morphological characters that have been used in the taxonomy of the Proteaceae. You will also use Weston (2007) to produce a list of characters and character states that are likely to be informative in reconstructing phylogenetic relationships between these species.

Materials for the pracs

For both practicals you will be using Herbarium sheets. On the herbarium sheets you will have the following codes: NSW860364, NSW860365, NSW860366, NSW863986, NSW863987, NSW863988, NSW863989, NSW863990, and NSW863991. We will refer to the specimens using these numbers.

The first three of these collections have already been identified to species: *Persoonia pinifolia*, *Persoonia lanceolata*, *Hakea teretifolia* but the others, each of which belongs to a different species, are still unidentified. Use the herbarium sheets.

(2) Weston (2007) publication on the Proteaceae.

(3) Botanical Glossary

(4) A Powerpoint file on flower symmetry

EXERCISES

Exercise 1 – Shoot morphology of the Proteaceae, starting with *Persoonia lanceolata* (30 minutes)

In order to describe characters and character states so that other people can understand what you are talking about, you need to learn about some morphological structures of plants and the terminology that is used for them.

Examine your specimen of *Persoonia lanceolata* (NSW860365) and compare it with the illustration of this species in Weston (2007: page 384, figure 134).

Find as many of the following structures on the specimen as you can:

- Shoot (= above ground part of the plant from which the organs such as leaves, buds, stems, flowers arise)
- Shoot apex (= the growing point of the shoot)
- Stem (= the axis of the plant that bears buds and shoots with leaves, and is connected to the roots)
- axillary buds (= in the leaf axil, the "niche" between the leaf base and the stem)
- scale leaves (= small scale-like leaves that enclose and protect dormant buds)
- leaves, each with:
    - base (= where it connects to the stem)
    - entire margins (= edges of leaves smooth, not dissected or serrated)
    - apex (= top of the leaf)
    - internal veins (= seen as lines on the leaf surface, usually with a prominent midvein)
- trichomes (= hairs; most dense on the youngest part of a stem, just below a shoot apex, but also found on the outsides of flowers and on immature leaves)
- inflorescence (= flower-bearing part of the shoot)
- floral bracts (= scale leaves and leaves in the axils from which flowers develop)
- flower bud (= small lateral or terminal protuberances that give rise to s flower)
- open flower, each flower has:
    - 1 pedicel (= the stem that attaches the flower to the inflorescence axis),
    - 4 tepals (= outer floral leaves– four in each flower),
    - stamens (= **the male reproductive part of the plant** that produces pollen, that has a slender filament ('stalk') that supports each anther.
    - 4 anthers (= the apex of each stamen where the pollen is; it is attached about 1/3 of the way up the inside of each tepal), each with pollen grains.
    - 4 hypogynous nectary glands (= small blobs surrounding the base of the pistil and inside and alternating with the tepals)
    - Pistil (= gynoecium or the **the female reproductive part of a flower**), composed of one carpel in Proteaceae), with
        - gynophore (= a short "stalk" forming the base of the carpel)
        - ovary (= swollen part of the carpel, above the gynophore and below the style), containing:
        - style (= elongated, narrow, apical part of the pistil), with:
            - stigma (= terminal region of tissue on which pollen grains can germinate)

Question 1. List some of the characters that you observed. You can illustrate what you have seen and upload a digital file to Moodle.

Exercise 2 –Shoot morphology of the Proteaceae continued, with *Persoonia pinifolia* (10 minutes)

Examine your specimen of *Persoonia pinifolia* (duplicate of NSW860364) and compare it with your specimen of *Persoonia lanceolata*. Find the homologues of the structures that you examined in *Persoonia lanceolata*.

Question 2. Describe how the two species of *Persoonia* differ from one another. You can illustrate what you have observed in the herbarium specimen.

In particular, note the following ways in which *P. pinifolia* differs from *P. lanceolata*:

- leaves linear and terete (cylindrical), with a crease running longitudinally along the lower surface
- floral bracts are leaves that are about ¼ the length of leaves and do not extend under flowers
- inflorescences anauxotelic (= they terminate in an aborted terminal bud rather than growing on into a leafy shoot)

Exercise 3 – Shoot morphology of the Proteaceae continued, with *Hakea teretifolia* (20 minutes)

Examine your specimen of *Hakea teretifolia* (duplicate of NSW860366) and compare it with your specimens of *Persoonia lanceolata* and *P. pinifolia*. Find the homologues (= characters that are similar in different species and arise from a common ancestor) of the structures that you examined in *Persoonia*.

Question 3. Describe how the two genera differ from one another. Illustrate examples of what you observed and submit a digital file to Moodle.

Note the following ways in which *Hakea teretifolia* differs from the two persoonias:

- a basal involucre (= whorl) of large leaves that enclose the developing inflorescence
- the leaf is terete (= cylindrical) but without the linear crease seen in *Persoonia pinifolia* and has an acicular (= needle shaped) and pungent (=sharp) apex
- each trichome (= hair) has two arms, not one as in *Persoonia* (most easily seen on immature leaves)
- the inflorescence is a cluster of flower pairs that does not grow on into a leafy shoot
- the floral bracts have been lost
- the flower bud is actinomorphic (radially symmetrical) early in its development (only visible by dissecting an inflorescence enclosed within involucral bracts) but soon becomes zygomorphic (bilaterally symmetrical)
- the open flower is strongly zygomorphic with asymmetrical tepals and a strongly curved pistil
- the anther is attached inside the tip of its subtending tepal and is much shorter relative to its width than the anther of *Persoonia*
- pollen grains are much larger than those in *Persoonia* and have prominent, bulging pore caps at the angles of the triangular grain
- a solitary shell-like hypogynous nectary gland is located on one side of the base of the pistil
- the pistil contains 2 hemitropous ovules ("hemitropous" means that the funicle (ovule attachment "stalk") is at the side of the ovule)
- the style:
    - is prominently curved
    - has a prominent, flanged-conical swelling at the tip, to which some pollen grains may be adhering and at the apex of which is a patch of papillose cells - the stigma. This swollen structure (minus the stigma) is called the pollen-presenter because in this species the anthers dehisce before the flower opens, shedding their pollen contents onto the pollen presenter. The pollen grains adhere to the pollen presenter until removed by a pollinator (a species of bee in this species).
- persistent fruits are found on several specimens. These are woody, toothed structures that have developed from the pollinated pistils of previous years' flowers. They were closed when

collected alive but have dehisced (split open) during pressing and drying, revealing two black, apically winged seeds.

Exercise 4 – Identification of six specimens to genus (1 hour)

Using the key to genera in Weston (2007) key out the six unidentified specimens to genus.

Question 4. List the generic names against the specimen codes and write out the succession of key leads that brought you to each identification (e.g., NSW860364 = *Persoonia*: key leads 1, 12, 13, 21, 23, 24, 40, 42, 44, 45, 55, 56, 57, 58).

Exercise 5 – Formulation of characters and character states that distinguish the specimens (1 hour)

By observing and comparing the specimens, as well as using the descriptions of subfamilies, tribes, subtribes and genera in Weston (2007) and any other relevant scientific literature, formulate a list of characters and their character states that are phylogenetically informative for the sample of taxa represented by the specimens. That is, find characters that differ between the specimens and distinguish groups of them.

For example, the two species of *Persoonia* differ from the other species in the sample in lacking proteoid roots. Although you cannot observe this character in the specimens provided, you can find this information from relevant literature (see Weston 2007: 364-365). This difference can be formulated as a character with two states:

Character	State 0	State 1
1. Proteoid roots (presence)	absent	present

The two species of *Persoonia* also differ from the other species in the sample in having a style tip that is not modified as a pollen presenter. This difference can be formulated as a character with two states, which can be added to our character list:

Character	State 0	State 1
1. Proteoid roots (presence)	absent	present
2. Style tip (pollen presenter)	not modified as a pollen presenter	modified as a pollen presenter

Another obvious character is the number of hypogynous nectary glands. There are four of these in the two *Persoonia* species and in NSW863988 and NSW 863989 but only one in *Hakea teretifolia* and in our other specimens. This difference can be formulated as a character with two states, which can be added to our character list:

Character	State 0	State 1
1. Proteoid roots (presence)	absent	present
2. Style tip (pollen presenter)	not modified as a pollen presenter	Modified as a pollen presenter
3. Hypogynous nectary glands (number)	four	one

Question 5. Each character must have at least two states and, in principle, could have thousands of states. In this exercise, the maximum number of states that a phylogenetically informative character could theoretically have is seven. Why this limit?

Reference

Weston, P.H. (2006) Proteaceae. Pp. 364-404 in K. Kubitzki (ed.) *Families and Genera of Vascular Plants* Volume IX (Springer Verlag: Berlin).

Practical 3 (Week 3) – Proteaceae lab 2

Dr Richard Jobson – Royal Botanic Gardens NSW

E26 Teaching Lab 3

Scoring, polarity, and tree by hand

In this laboratory you will examine nine sampled species of Proteaceae using the Herbarium specimens; this includes specimens examined for the characters you formulated in the laboratory last week. Give a revised data matrix, and then phylogenetically analyse your datamatrix by hand.

Exercise 1 – Completing the Proteaceae datamatrix

Given that there are many characters that cannot be observed remotely, we have provided you with a datamatrix spreadsheet for the nine Proteaceae species that you identified in Practical 2 (in the Moodle Prac 3 tab). Download the spreadsheet from Practical 3 TAB in Moodle.

In the spreadsheet you will see that nearly all the characters have been filled out for you. Please fill out the leaf characters for the nine Proteaceae species for Characters 5 and 6

Character 5: Leaf in transverse section: Character states: flat (0); terete (1)
Character 6: Leaf margins: Character states: entire (0); toothed (1)

The nine species of Proteaceae are:

NSW860364 – *Persoonia pinifolia*
NSW860365 – *Persoonia lanceolata*
NSW860366 – *Hakea teretifolia*
NSW863986 – *Grevillea oleoides*
NSW863987 – *Hakea globosa*
NSW863988 – *Banksia marginata*
NSW863989 – *Banksia ericifolia*
NSW863990 – *Grevillea diffusa*
NSW863991 - *Grevillea sphaercelata*

Question 1. Revise the given datamatrix spreadsheet for the two leaf characters (5 and 6) and upload the spreadsheet in Moodle.

Exercise 2 – Constructing an unrooted tree by hand

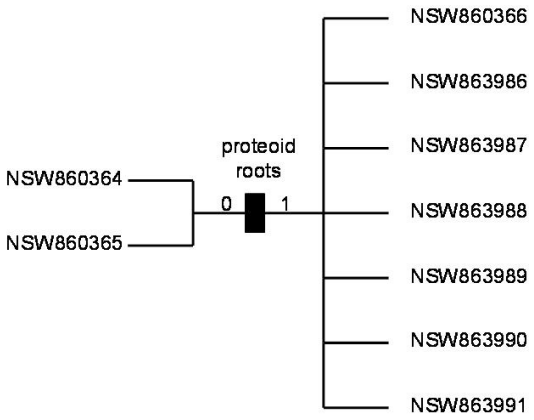
These days, systematists routinely use computer programs to find the optimal set of trees for a phyogenetic data set and you will get to do this in a later lab exercise. However, prior to the ready availability of computers with phylogenetic software, systematists routinely constructed phylogenetic trees by hand and this is what you will do today.

The optimality criterion that we will use is parsimony – trying to find the tree(s) that requires the minimal number of character state transformations ("steps") to explain the data in your matrices. We know that at least some of the characters will be completely congruent with each of the most parsimonious tree(s) so we will build several trees, starting each time with a different character that

will be allowed to define the first split between subgroups and then progressively split the subgroups using other characters. The procedure is as follows:

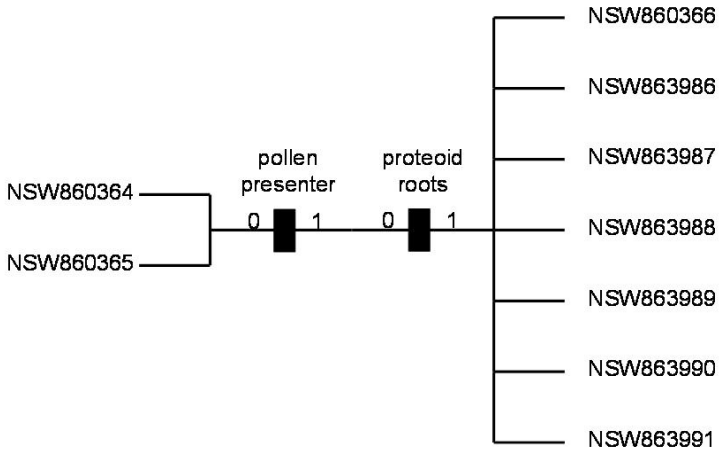
1. Choose a binary character (one with only two states) for your first split, or, if you choose a multistate character, lump states so that you end up with only two states. These choices can be arbitrary or you can try to use your insight to make more judicious choices.

2. Draw an unrooted tree with all of the species possessing character state 0 on one side of a central split and all of the species possessing character state 1 on the other side. For example, if we choose presence of proteoid roots as our first character, we get the following result:

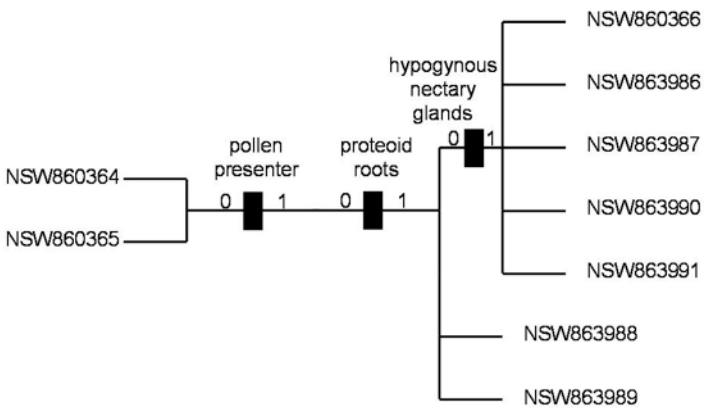


Question 2. Draw your unrooted tree and submit with Moodle.

3. Examine your data matrix for other characters that have the same distribution as your first character. Whether or not the style tip is modified as a pollen presenter does. Adding this character gives the following tree:



4. Choose another character as in step 1. Let us add the number of hypogynous nectary glands as our third character. Adding this character gives the following tree:



5. Continue adding the rest of your character data to the tree. Some characters will be incongruent with the topology that you have resolved so far and will need to be reconstructed as homoplasies (they will require parallel or reversed transformations). **NOTE:** Homoplasious characters are those that change more than once on a tree.

Question 3. Illustrate your unrooted tree with all your character states represented in the format shown above. How many character state transformations does your tree require (i.e. how long is the tree in "steps")?

6. Build another tree starting with a character that was reconstructed as homoplasious on your first tree. **NOTE:** Homoplasious characters are those that change more than once on a tree.

7. Keep building trees, starting with different characters, until you have exhausted all of the possibilities or run out of time.

Question 4. Which of your trees is most parsimonious (shortest)? Have you found more than one equally parsimonious tree? If so, illustrate at least two tree and outline their differences.

Exercise 3 – Rooting your best tree

The trees that you built in the previous exercise are unrooted, so they have no time dimension. In order to give one of these trees a relative time dimension, we need to identify part of the tree as its root. One way to do this is an "indirect method" called outgroup comparison, in which we use previous phylogenetic research to identify an outgroup, and locate the root on the internode connecting this outgroup with the rest of the tree. Figure 131 in Weston (2007) is the result of such research.

Question 5. Compare your shortest tree with this figure. Where does this suggest your best tree should be rooted?

In the absence of information about possible outgroups, we can use a "direct method", based on ontogenetic information, to polarise one or more characters and thus restrict the root to part of the tree. We have ontogenetic information for at least two characters. Firstly, we know that proteoid root clusters develop from lateral roots. Species that do not produce proteoid root clusters also have lateral roots. Lateral roots are thus observed to have a more general distribution than proteoid roots.

Question 6. Where does this suggest your best tree should be rooted?

Another character for which we have ontogenetic information is floral symmetry. Examine the very young flower buds of specimens with zygomorphic mature flowers (e.g. NSW 863990).

Question 7. How would you describe their symmetry, compared with the mature flowers on the same specimen?

Question 8. How would you describe them compared with mature actinomorphic flowers of other specimens?

Question 9. Does this suggest any conclusions regarding the placement of the root of your best tree?

Exercise 4 – Building trees

Question 10. Give three points about what you have learnt about building trees from Practicals 2 and 3.

Practical 4 (Week 4) – TNT and Mesquite lab

Prof, Gerry Cassis (UNSW)

E26 Teaching Lab 3

The purpose of this lab is to teach students how to construct matrices, run analyses, optimise characters, and manipulate trees using computer software.

You will learn how to produce your own phylogenies, from data matrix construction through to tree searching, tree selection, generation of support statistics and character analysis.

In this laboratory you will work with pre-existing datasets. These can be downloaded from the folder in Practical 4 TAB.

We will be using two software programs: 1) **Mesquite** – which you will use to build data matrices and view the results of analyses; and, 2) **TNT** – which you will use to construct phylogenies using parsimony based algorithms. These programs will be used interchangeably throughout the laboratory, and you will need to be familiarise yourself with the menus, icons and command lines.

Please note that all the programs can be accessed from your own computer via MyAccess.

PART 1: THE MATRIX

Most phylogenetics programs require matrices to be written in 'Nexus' format (a standardised data format used in phylogenetics). Other formats can be used which are tailored to specific software. The software program for generating cladograms in this laboratory is TNT (https://cladistics.org/tnt/) which can open basic Nexus files but works best with its own TNT format (similar to Nexus). If you have a Nexus file that won't open in TNT, the easiest way to convert it into a workable TNT format is to first open it in Mesquite software (which may ask you to convert the file), and then export it in TNT format (**File> Export> TNT**).

The TNT file format is simple once you get used to it, then datasets can be edited in Notepad, Wordpad or other text editors. In general though, it's easiest to build and manipulate your matrix in Mesquite, TNT, or some other phylogeny program.

TNT files generally begin like this:

```
xread
'My little matrix'
nchar ntax
Taxon0   1011110000
Taxon1   1111111000
Taxon2   1011110000
Taxon3   1111111000
;
```

xread is the command which tells TNT to read in the dataset.

Often you will also have a title or some comments placed within single quotes, like 'My little matrix'. Anything can be written within the single quotes as they inform the program to skip ahead and ignore the quotation contents.

Next we have a line listing the number of characters and the number of taxa. For this particular dataset it should read:
4 10

Lastly we have the matrix itself.

Taxon names must not have any symbols in them, nor any spaces, and can be up to 32 characters in length. If you wanted to use a genus and species name, the best way to do so is to use an underscore, e.g. Tyto_alba. You'll notice in the matrix above that the first taxon is numbered 0. The default numbering regime for TNT is to start at zero, both for taxa and for characters. For example, the 4 taxa are numbered 0-3, and the 10 characters are numbered 0-9.

Each matrix ends in a semicolon (;), which informs the program of the data endpoint. Matrices will often have other commands, comments or notes after the matrix (e.g. character names and state names, instructions which automatically change settings for different characters). These are best added in either the matrix editor (i.e. Mesquite for the character and state names) and in TNT itself (specific instructions for character polarity, character type, etc.).

EXERCISES

Exercise 1: Data

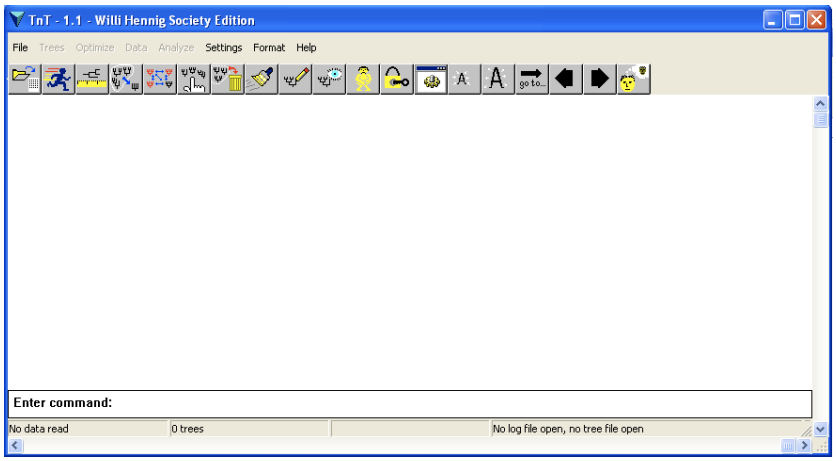
Open the file "**example.tnt**" in Notepad. Notice how there are the basic elements described above, (xread, a comment line, the number of characters and taxa, and a bunch of 1's and 0's) plus additional information (ignore this for now). You'll see that some characters aren't 1's or 0's, but are instead question marks. This tells TNT that we either don't know what this character is in this particular taxon, or that this character is inapplicable for the given taxon.

PART 2: TNT BASICS

TNT is a powerful phylogenetics program that can analyse very large amounts of data, and allows the user to manipulate many parameters during tree searching and analysis. Additionally, TNT can run user-generated scripts, which allow one to customise tree searches involving a series of commands and instructions. For now we'll keep it simple, and use only the basic TNT functions.

Exercise 2:

To begin, open TNT. You'll see a window that looks like this:

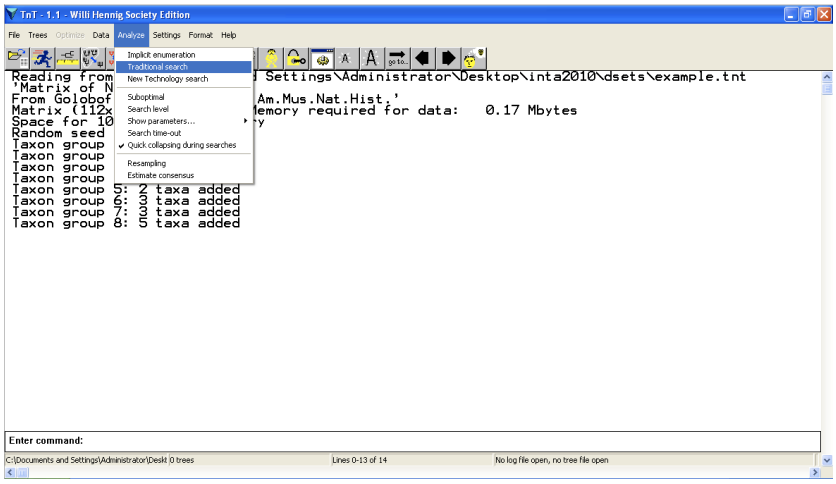


The operations indicated by the square icons can also be accessed from the various drop-down menus along the top, or if you are a TNT expert you can input commands through the command line.

To open a dataset for analysis, you can either click on the folder image (upper leftmost icon) or select **File> Open Input File>** and then scroll over to the file you want to open. The default settings search only for TNT files but you can also search for other file types if need be.

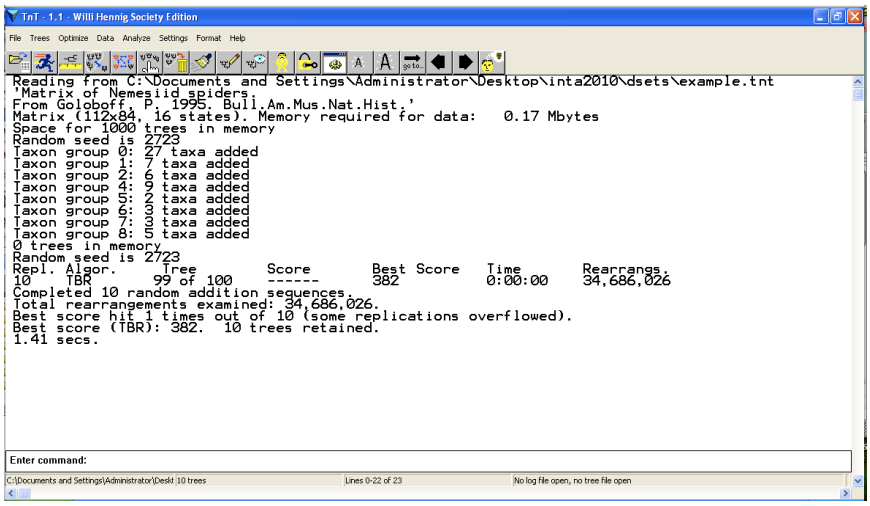
Go ahead and open **example.tnt,** which is in the folder labelled 'TNT files' on the desktop. As soon as you open the file TNT returns some basic information, which will vary slightly depending on what other instructions may have been included with the matrix file. Basically you need to make sure that the file has opened successfully and that no error messages are displayed.

TNT can run a search to find an exact solution (i.e. **Analyse> Implicit enumeration**) but this is seldom possible, particularly with datasets of over 18 taxa. Instead we use heuristic (approximate) methods of searching, whereby short-cut techniques are applied within specified parameters. In TNT this is achieved by clicking **Analyse>Traditional search**. Do this now.



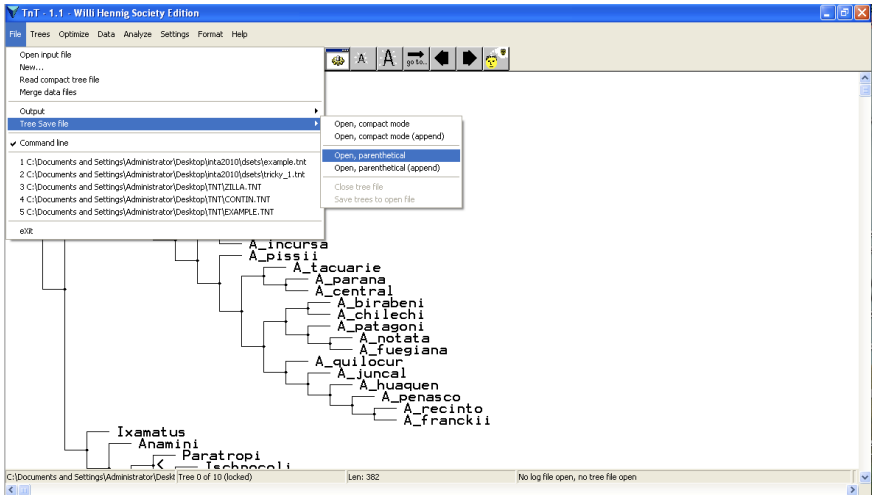
When you do this you will see a smaller drop down box which has various editable parameters. For now we'll leave the default settings as is. Hit **Search** and TNT will run the analysis.

Very quickly TNT will obtain a result. In this instance, it returns 10 trees, all of 382 steps in length.



To view these trees, click on the **view icon** (the icon with the eye floating in front of the phylogeny). You can scroll through the trees using the big arrow icons. The trees are numbered 0 through 9, which is indicated at the bottom of the screen.

At the moment the trees are loaded in the computer buffer, so if you want to save them (to open in another program for example) you need to first transfer them to a tree file. To do this, select **File> Tree save file> Open parenthetica**l and then create a tree file named **example.tre**.



To confirm that the trees have been sent to the file go back to the main TNT screen (if you're still in the tree viewing mode then click on the **view icon** again) and it should say "10 trees saved to c:\...". In some cases the trees might not automatically export (e.g. if you generate more trees after you open the tree save file), in which case you have to click on **File> Tree save file> Save trees to open file** to ensure the trees are saved. We'll do this later when we generate a **consensus tree**.

Now that we have 10 trees in our (still open) tree file, we want to generate a consensus tree, which is a summary tree (i.e. a tree that only includes the branches that are common to all the trees found.) A consensus tree isn't a true phylogeny and shouldn't be used for tracing character evolution, but it is a good way to summarise alternative solutions.

To generate a consensus, click on **Trees> Consensus**. A second window will pop open with several options. Here we can choose what sort of consensus tree to use, whether to include all the trees we've found, whether to include all taxa, and whether to just show the tree, or save it into the buffer (**RAM**).

In this case we want to use the default settings (strict consensus) and save the tree to **RAM**, so click on that option. A message now pops up saying you've generated a consensus of your trees, and that it is 'tree 10'. You can view this tree by clicking on the **view icon** and using the arrows to scroll to 'tree 10'.

We now want to append this consensus tree to the end of our tree file. Go back to **File> Tree save file> Save trees to file**. Since we already have the initial 10 trees saved to the tree file, this time we only want to add 'tree 10'. Click the **Select trees** option, and another window opens up where you can pick what trees to include and what trees to exclude. Set it up so that we only have 'tree 10' (which is the eleventh tree, our consensus) selected to be included (the invert button might help speed things up). Click **OK**, and **OK** again. You are now informed that one tree has been added to the open file.

We are now finished adding trees, so we can close our tree file: **File> Tree save file> Close tree file**. A message pops up to tell us the file is closed, with 11 trees saved.

We now have a set of trees, but what do we do with them? Although TNT does have the capability to display trees and examine how characters change states across the tree, this is much more easily done in Mesquite.

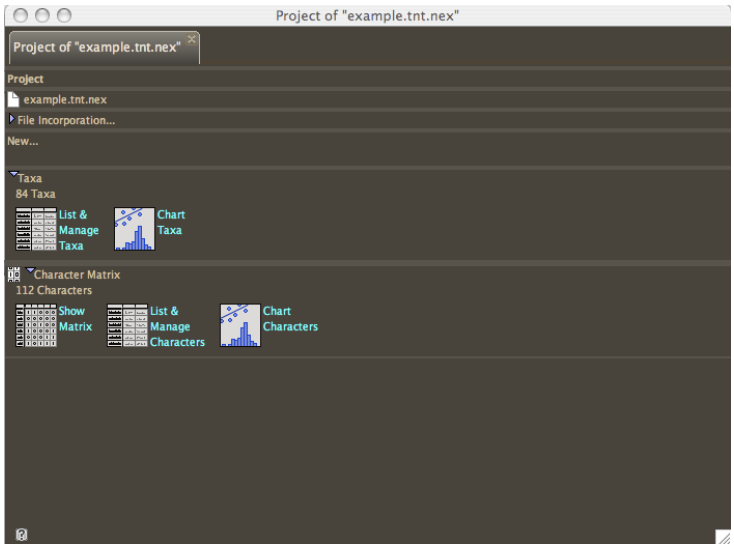
PART 3: VIEWING TREES IN MESQUITE

As with TNT, Mesquite is a very powerful tool which can do many things. Additionally, the creators encourage others to write their own modules for Mesquite, which are eventually supplied on the program website. If you are interested in learning more about what this program does, go to http://mesquiteproject.org/mesquite/mesquite.html. You can also run through a series of examples (nexus files that run like a slideshow) which can be found in the Mesquite_folder in Program Files

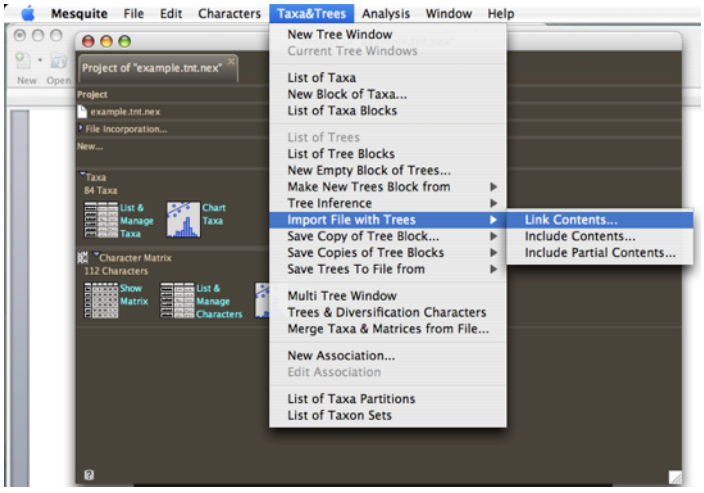
Exercise 3:

Open Mesquite and wait as it loads up all the modules. Mesquite works with Nexus files, so the first thing you need to do is convert your TNT file. Open your dataset **example.tnt**. A window will pop open asking you what format this is. Scroll down and select TNT – Mesquite then generates a Nexus file of the same name but with '.nex' added to the end (i.e. **example.tnt.nex**). Click ok and we're ready to go. (An alternative method would be to select **Data> Export (Nexus format)** in TNT. The constant need to generate new files when going back and forth between programs can be confusing, so it's good to develop a file naming system that allows you to keep track of which file you're on.)

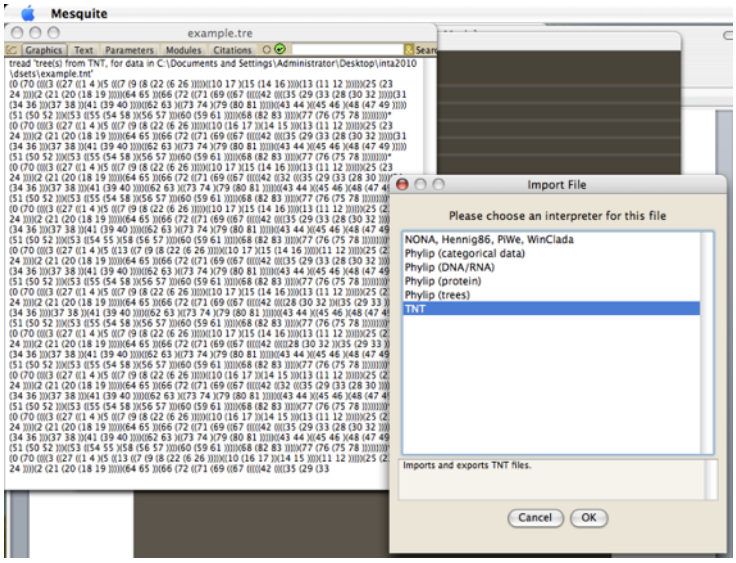
Our matrix is now loaded into Mesquite, and you will see the following:



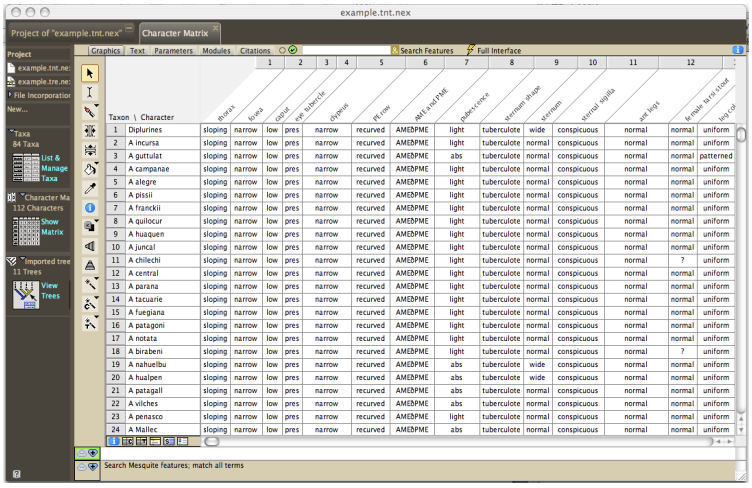
From here you have several options (including all the options in the tool bar at the top), but we'll mostly be working with the Matrix window (**Show Matrix**), the lists of characters and taxa (accessed from the Matrix window), and eventually the tree window. Before we proceed, we want to load up the tree file **example.tre**. To do this select **Taxa&Trees> Import File With Trees> Link Contents** and then select the appropriate file. (If you select **Include Contents** this will merge your matrix with your tree file, which sometimes makes it hard to open these in other programs.)



As before, TNT tree files aren't quite the same as tree files generated by other programs, and Mesquite will ask how to interpret your tree, then generate a new tree file ending with '.nex'.



We now have a matrix and multiple trees loaded. First take some time to see what the matrix looks like: you'll notice that the taxon names occupy the first row, and the characters are listed along the top, with their associated character states below in the cells adjacent to the respective taxa. (These state name labels are translated back into integers in the TNT file matrix block). As you move your cursor over the assorted tool icons you'll notice that their function is briefly described at the bottom of the window.



Let's move on to the tree window. Click on **View Trees** at the bottom left. A tree window appears with Tree 1 displayed (Note: while TNT numbers everything beginning with 0, Mesquite is much more logical and starts counting at 1, so tree 1 is the first tree!) Notice that as with the Matrix window, holding your cursor over the tools immediately left of the tree will generate a brief statement on the given tool's function. There are also many, many functions one can access from the toolbar along the top (most of which you'll never touch).

You can scroll through your trees with the blue-grey arrows at the upper left. With so many branches it's hard to see what's changing, but most of the variation across trees is tied up in the relationships of **St platen, St palmar, St crassi** etc. at the far right of the tree. Go to tree 11 and you'll see these taxa forming an unresolved polytomy.

Question 1. Why is there a polytomy here? What does it mean:

Tree statistics
There are a few tree statistics we commonly report when presenting trees.
These include:
Tree length: the total number of steps (character state changes) necessary to generate a given tree.
Consistency Index: Measure of relative homoplasy in a cladogram. The CI of a tree is a ratio of the sum of the minimum number of possible changes divided by the observed number of changes for each character. The CI is a rough assessment of the difficulty in fitting a dataset to a particular tree topology.
Retention Index: The fraction of potential synapomorphy retained as synapomorphy on a cladogram. In other words it is a measure of the proportion of similarities on a tree.

All of these tree statistics can be generated and displayed in Mesquite using the **Tree Legend**. **Analysis> Values for Current Tree>** will open another pop-up window, from which you can select from a list of tree legend options. Before you display the tree statistics, you should return to any one of the other equally most parsimonious trees.

First select **Tree Length**. Once you do this, a legend displaying the tree length appears at the upper left of the tree, and a **Legend** tab appears at the top of the screen. **Legend> Show> Tree Values Using Character Matrix** will allow you to add in the Retention Index and Consistency Index. (You can close the legend with **Legend> Close Values Legend**.)

Question 2. Are the statistics the same for all of the trees? Why or why not?

Mesquite allows you to move taxa around by using the **Move branch** tool (the arrow icon in the tree window). Moving branches around doesn't affect the original file, but it will change the tree statistics. Go ahead and move a few branches around while the legend is open.

Question 3. What does moving branches do to the tree length, CI and RI? Why do you think this is?

To return to the original tree click on the enter symbol (blue button at the upper left next to the tree).

PART 4: BUILDING A MATRIX IN MESQUITE

Matrices can be built in a number of programs, including TNT. However, as mentioned before, TNT is not terribly intuitive to use, nor is there much in the way of a user manual. Instead we will use Mesquite for our matrix building needs.

Exercise 4:

For this exercise we will produce a simple matrix for vertebrates, which we will then use to reconstruct their phylogenetic tree.

The following taxa will be coded:
Fish
Amphibians
Lizards
Birds
Mammals
Sponges (outgroup)

The following characters will be used to generate the matrix:

Vertebrae: Absent or Present
Limbs: None, Fins or Legs
Lungs: Absent or Present
Amniotic egg: Absent or Present (=tough sac enclosing embryo in egg, associated with animals that do not have an aquatic larval/tadpole stage)
Skull – temporal fenestrae: Absent, Diapsid (two openings – birds, lizards and most other reptiles) or Synapsid (one opening in skull behind the eye – mammals and extinct relatives).

To begin, open Mesquite and when prompted, choose to open a new file. Call this file **Vertebrates.nex**. When you do this, a new window pops open asking **"Do you want your new file to include taxa?"** – choose 6 new taxa and make sure to check the box that says **Make Character Matrix**. When you click ok yet another window pops open asking how many characters you want to include, and what type of characters these will be. We'll start with the 6 characters above, which will be standard categorical data.

To enter your taxon names, select the **Edit tool** from the buttons to the left of the matrix, adjacent to the arrow icon (it looks a bit like a square bracket). It is generally best to enter the outgroup taxon name first, as by default most programs set the first taxon named as such when tree searching.

Question 4. What is an outgroup and why do we include one?

Now that we have the taxon names entered we need to edit the characters and character states. To do this click on the **Show State Names Editor window** icon (the little icon with the S below the matrix). When you do this another tab is opened in which you can edit the character names and state names. Go ahead and enter in the names and states for the first 6 characters.

At the moment there is no character to differentiate lizards from birds. Let's add one more character: **feathers**. We can do this several ways: we can choose **Matrix> Add characters** (if in the Matrix window) or **List> Add Characters** (if in the **List of Characters** window, which is accessed from the icon with the letter C, located at the bottom of the Matrix window). Alternatively, from the Matrix window we can select the **Add characters icon** and click either at the end of the columns of characters or between character columns.(Additional taxa are added in a similar manner with the **Add taxa icon**.) We can delete characters or taxa by highlighting them and clicking **Edit> Clear**, which then gives you the option of clearing the cells or deleting the character/taxon.

Once we have our matrix ready we can start to fill in the values for each cell. As we enter 0's and 1's in the cells, Mesquite replaces these with the state names when you hit enter.

How should we deal with the character '**Skull – temporal fenestrae**'? For most taxa this is easy – temporal fenestrae are only found in Amniotes, so they are absent in fish and amphibians, but what about sponges? Since sponges don't even have a skull, in this case we have to treat this character as 'inapplicable' for this taxon. In other words, leave a '?'.

We are now ready to save our matrix. First save the file as a nexus file, but since we want to analyse it in TNT we should also export it as a '**.tnt**' file. Choose **File> Export** and select **TNT**.

Now open your file in TNT and run the analysis. If you didn't enter Sponges as your first taxon, you will have to tell TNT that this is the outgroup. See if you can figure out how to do this. '

Question 5. How many trees do you find? What is the tree length? Draw the tree in the prac sheet.

PART 5: COMBINING MORPHOLOGICAL AND MOLECULAR DATASETS

Often one has data from different sources e.g. both morphological and molecular data, sequences from different genes, etc. Often you'll want to combine all these datasets into a single analysis (**total evidence** or **simultaneous analysis**), but the different types of data cannot all be treated the same way. In TNT we can run such analyses by using interleaved data and a few simple lines of TNT commands. Firstly, we need to make our data interleaved – this means the data are separated into discrete blocks, each of which has rows of species names and columns of data (as opposed to non-interleaved datasets, where the taxon names appear only once and all of the characters are strung out to the right, scrolling as far as it takes).

To make TNT datasets interleaved we use the '&' symbol followed by a code designating the type of data in that particular block written in square brackets. e.g. '**& [dna]**' means that the following block of interleaved data is DNA sequence data, while '&[num]' refers to standard categorical (='numerical', with states coded as 0, 1, 2 etc) data. This line is written after the line detailing the number of characters and number of taxa in the file.

```
xread
'my file'
120 44
&[num]
Taxon 1 00101010101
Taxon 2 00011211001
etc.
```

When you get to the next block of data it should look something like this:

```
&[DNA]
Taxon 1 AACTGCTA
Taxon 2 AAGCGCTA
etc.
```

And of course, at the end of the whole dataset finish with a **semicolon** (;) after the data block, then a hard return, and lastly a line with nothing but '**proc/;**'

Exercise 5:

Datasets **part_a.tnt** and **part_b.tnt** are imaginary morphological and molecular datasets, respectively. Use Notepad to combine the two into a single file. If you get it right you'll know because it actually loads up.

Question 6. What is the minimum tree length of the most parsimonious reconstruction(s)?

PART 6: OPTIMISING CHARACTERS

Character optimisation, or character mapping, is extremely useful in exploring the evolution of specific traits across taxa. This might involve tracing the evolution of the characters which were used to build the phylogeny, characters which were not used in phylogenetic construction, or for comparing different characters. Mesquite can attempt to reconstruct character states at ancestral nodes through parsimony, likelihood or Bayesian methods.

In this exercise we will look at the evolution of male and female reproductive traits in the plant bug genus *Coridromius*, (using the dataset of Tatarnic and Cassis 2010). *Coridromius* is unusual in that males have hypodermic genitalia with which they stab females, and females have evolved various

'paragenital' structures to reduce mating damage and limit infection. In response to such costly mating, females of some species have essentially evolved a new copulatory tract!

Exercise 6:

Open TNT, and open the file **Coridromius.tnt**. This matrix includes two characters we do not want to use in tree building. We sometimes include characters which are interesting to look at on the finished tree, but for some reason or another might not be appropriate for building the tree (e.g. distribution data, food preference, behaviour). In this case, the last two characters, 'male complexity' and 'female complexity' are summaries of male and female reproductive traits which are already coded in the matrix – we do not want to use these in tree building but we do want to look at them in the second part of this exercise.

Question 7. What would be wrong with including these characters in the analysis?

To mask these characters, click **Data> Character settings**. This will open a smaller window – on the right side of the window select **CHARS** which will open a third window. From here you can select the two characters we wish to deactivate. Once selected, you then end up back at the second window. Check the '**inactive**' box then click **OK**. We are now ready to run the analysis.

Using the traditional search, find the most parsimonious reconstruction(s) for this dataset. When you have completed the tree search, save your tree(s) to a **treefile** and close TNT.

Question 8. How many trees did you find? What length?

In Mesquite, open the matrix **Coridromius.tnt**. As before, you will have to convert this into Nexus format. When this is done, we are ready to load up our trees. To do this select **Taxa&Trees> Import File With Trees> Link Contents**. (If you select **Include Contents** this will merge your matrix with your tree file, which sometimes makes it hard to open these in other programs.) As before, your tree file will need to be converted to Nexus format. Once all the files are loaded you can now open the **tree viewing** window and begin to examine character evolution.

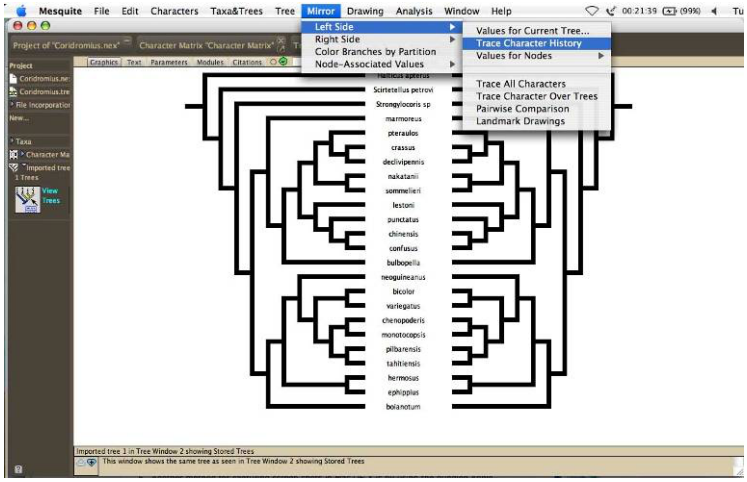
In the tree window, click on **Analysis** – there are many different analyses you can choose from, but for the time being we only want to trace character evolution. Select **Analysis> Trace Character History> Parsimony Ancestral States**. You should now see a small **Trace Character** legend at the lower left, and the first character mapped onto the tree. (The legend can be moved around if it's obscuring your view of the tree.) You will notice that Mesquite not only colours the terminals (i.e. branch tips) with the appropriate character state, but it also attempts to reconstruct the possible character state in hypothetical ancestors. Occasionally, the ancestral condition will be equivocal. This will be represented by branches coloured with multiple states. Scroll through a few characters to get an idea of how characters are mapped.

Now let's look at the last two characters: male complexity and female complexity (68 and 69). On the tree you may notice that for some taxa there is no little box below the taxon name for one or the other of the characters. This reflects that some species are known from only one sex, and the characters associated with the opposite sex are unknown. In such cases it is sometimes helpful to prune these taxa from the tree. This is done by selecting the scissor icon
and clicking on the branch leading to the taxon in question. This does not alter the tree file; it only changes the image on screen (unless you choose to save the changes when closing). Go ahead and snip off the taxa with missing data for characters 68 and 69.

When flipping back and forth between the two traits, you should notice that those taxa with higher male complexity also seem to have higher female complexity. An easy way to visually compare the evolution of two characters on the same tree is to use mirror trees. At the top of the screen, select **Tree> Mirror Tree Window**. This produces two mirror images of the tree, left and right. We can now trace characters independently on each mirror image.

From the **Mirror Tab**, select **Trace Character History** for both left and right sides. Display male complexity on one side and female complexity on the other. You should notice that male and female complexity appear strongly correlated. As we will show later in the course, such patterns can be tested empirically using methods such as phylogenetically independent contrasts, which take phylogeny into account while testing for correlation.

Question 9. Can you think of any other types of correlations that you could look at in this manner?



In parsimony ancestral state reconstruction, ancestral states are resolved as either a definite state or as equivocal. However, in the real world we can never really be certain what an ancestral state really was.

Question 10. Why can't we be 100% certain what an ancestral state was?

This uncertainty is best exhibited using **Maximum Likelihood** reconstruction. Likelihood character reconstruction is usually reserved for trees where relative branch lengths are known or have been estimated (e.g. trees derived from molecular data, trees with divergence times calibrated from fossils). For this demonstration we'll just assume that the branch lengths on this tree are correct. To change from parsimony to likelihood, select from the top of the screen **Trace> Reconstruction> Likelihood Ancestral States**. (There should be two **Trace** tabs along the top, one for each side.) For this exercise use the default **Current Probability Models**. Once you do this you will see that the reconstruction appears less certain. Holding your cursor over any branch or node will display the relative probabilities of each character state.

Another way to visualise this uncertainty is to change the form of the tree. **Select Drawing> Tree Form> Balls & Sticks**. This replaces each node with a pie chart, which is coloured in the relative probabilities of each state (you can still get exact probabilities by pointing your cursor at the nodes).

Practical 5 (Week 5) – Molecular computer lab 1

Dr Frank Koehler – Australian Museum

E26 Teaching Lab 3

To investigate the phylogenetic relationships of the organisms of interest, we need to create a data matrix from the available DNA sequences. Each row of this data matrix contains a DNA sequence and each of its columns contains a nucleotide. Nucleotides contained in the same column are considered to be in a homologous position across all sequences in the matrix. In other words, all sequences in this matrix are aligned with each other. Therefore, a data matrix built from DNA sequences is often referred to as a 'multiple DNA sequence alignment' (MSA).

This practical contains exercises that exemplify all necessary steps to construct a DNA sequence matrix (MSA), from processing raw sequence reads produced by an automated sequencing machine to building a multiple sequence alignment.

The used sequences stem from a systematic study of endemic land snail species from Timor (genus Parachloritis).

Two programs will be used, **BioEdit** (a stand-alone software to view and edit single DNA sequences) and **MEGA7** (software to view, produce, edit and analyse sequence alignments).

Please, follow the verbal instructions given during the practical.

Exercise 1 – assembling electropherograms to retrieve a single, complete DNA sequence of the target fragment (gene: *cytochrome c oxidase subunit 1*)

Program used: **BioEdit**.

Automated sequencers read both DNA strands of a target fragment and produce two separate nucleotide sequences that represent each of the two anti-parallel DNA strands.
These two sequences are often called the 'forward' (F) and 'reverse' (R) strand, respectively. In the present example, the two strands are demarked with 'L' (for light = forward strand) and 'H' (for heavy = reverse strand).

Both reads need to be combined with each other to retrieve the complete target fragment sequenced (a 'contig' sequence is being formed, which is a consensus sequence resulting from two or more overlapping DNA sequences). To form a contig, the two overlapping, reverse-complementary sequence reads need to be aligned with each other. Where both reads overlap, any possible mismatches may indicate reading errors that need to be corrected.

Your task in this exercise is to assemble each one forward (L) and reverse (H) read to a single contig, to check for and correct any potential misreads, to complement any missing sequence parts, and to truncate the contig to its final length by removing sequence residues of the PCR primers, which have been built into the sequenced DNA fragments during the PCR reaction.

By convention, sequences are always given in 5' to 3' direction (which refers to the C-atoms in the pentose sugars that form the backbone of the DNA molecule).

Open the program **BioEdit** and find the sample files on your computer (1681_COI-H.ab1; 1681_COI-L.ab1). When you open these files, each is shown in two separate windows:

- a NOT EDITABLE **graphic window** displaying the electropherogram (i.e., showing the signal as read by the sequencer),
- an EDITABLE **sequence window** containing the sole letter coded sequence inferred from the electropherogram (Fig. 1).

All edits are performed in the **sequence window**, while the **graphic window** is used only to look at the chromatogram to assess the reliability / quality of portions of the sequence reads.

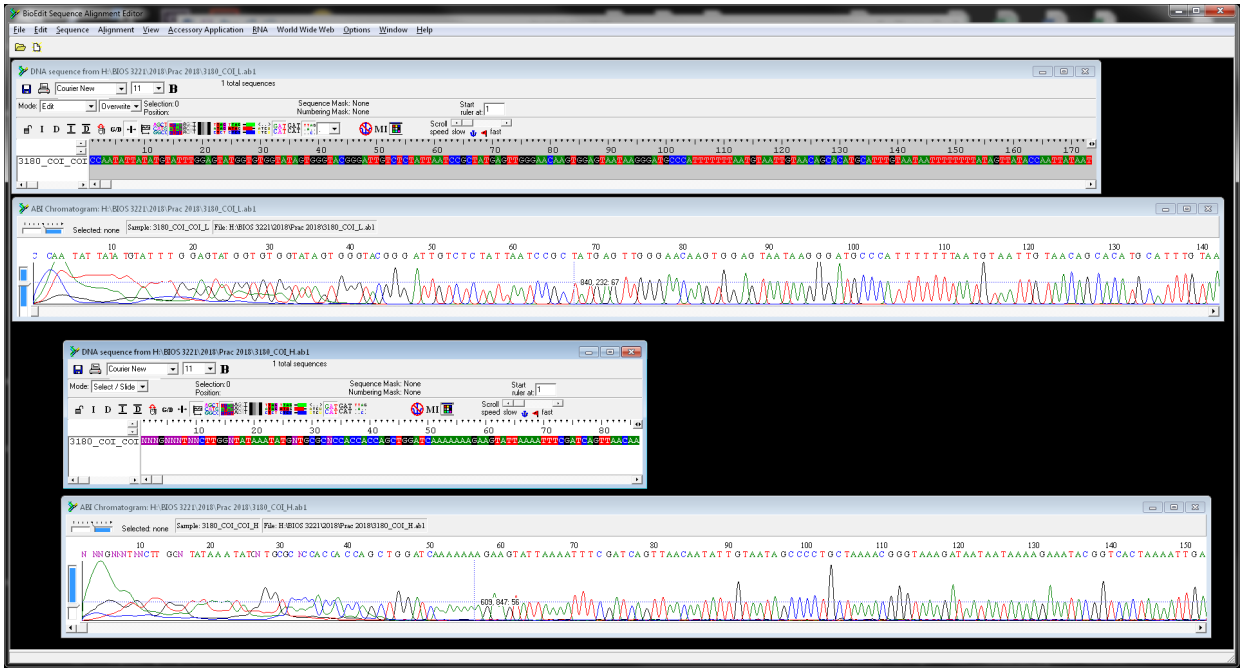


Fig. 1. The BioEdit window used to edit and concatenate sequence electropherograms. Each sequence file is shown in two windows, the sequence and graphic window. Commands are performed by activating the sequence you want to edit by a click and using the drop down menus from the line above.

To edit the nucleotide sequence in the sequence window, under **Mode** select the option "edit" (upper right corner of the window).

Keep in mind that the forward and reverse reads are two anti-directional versions of the same target fragment that are reverse-complementary to each other.

The first step is to identify and delete any unreliable or unreadable segments of reads (usually at either end of sequence). To find such regions, inspect the electropherograms by eye in the graphic window. Start from the end of each sequence and work your way back to the beginning. Not clearly readable portions of the sequence read (i.e., parts that lack well-defined, regularly spaced and non-overlapping peaks) should be deleted.

Correct or delete these unsuitable sequences segments in each the L and H sequence (Note: edits can only be done in the **sequence window on a selected sequence (mouse click)**).

Secondly, after having eliminated all ambiguous reads, copy both cleaned sequence reads into one window (i.e., you can copy and paste either L or H over to the complementary sequence file using **Edit > Copy Sequence / Edit > Paste Sequence**).

Next, ensure that both sequences are in the conventional 5'-3' direction in order to be compatible with each other. This is done by creating a reverse-complement of the reverse sequence (H sequence) by using the command: **Sequence > Nucleic Acid > Reverse Complement**). Ensure the sequence to be edited is active (**mouse click** on the sequence).

You can also display the electropherograms as reverse complements in the graphic window if you need to (**View > Reverse Complement**).

If both sequence reads are free from errors and have the same directionality, than both strands can be aligned with each other using the command '**Sequence > Pairwise alignment > Align two sequences (allow ends to slide)**' (note: select both sequences, **mouse click**).

A new window should open in which you find the L and H sequences aligned with each other. Check for errors or discrepancies in the nucleotide sequence of both strands and correct any potential misread by inspecting the two electropherograms at the relevant position. If no errors occur, concatenate both single strands into one consensus sequence by **copying & pasting the overlapping ends and save the consensus sequence to the computer as a new file** ('1681_COI_contig.fas').

The target fragment to be used for subsequent analyses is delineated by the two **PCR primers**, which were also amplified during the PCR reaction (Fig. 2). Because these residues are not part of your template DNA, they should be removed before the sequences are used in downstream applications.

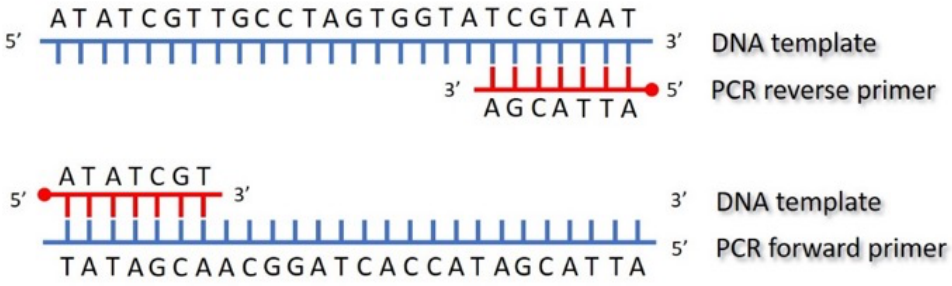


Figure 2. PCR primers become part of the PCR amplicon during the annealing step.

Knowing the sequences of the two PCR primers used in the PCR reaction, you should be able to prune any remainder of the PCR primers still contained in your complete DNA template from the consensus sequence ('1681_COI_contig.fas') and obtain the complete target fragment at its correct length (remark: the primer sequences are incomplete; find the end of each primer sequence).

The relevant information you need is given below in Table 1. Identify the primer binding regions at either end of 1681_COI_contig.fas and remove the primer sites from your fragment.

Table 1. Primers used to amplify the target fragment of the partial cytochrome c oxidase subunit 1 gene (COI).

Primer	Directionality	Primer sequence
forward	5' – 3'	GGTCAACAAATCATAAAGATATTGG
reverse	3' – 5'	TAAACTTCAGGGTGACCAAAAAATCA

Question 1. How many base pairs has your target fragment 1681_COI_contig.fas after the PCR primer residues are removed?

Exercise 2 – Pair-wise alignment of two short non-coding sequences by hand

The task is to align two non-coding DNA sequences with each other by hand. The sequences differ in the number of their nucleotides. To align the sequences it will be necessary to insert gaps. The question to be addressed is: **What is the optimal alignment**?

The two sequences are: ACCCCAGGCTTA
ACCCGGGCTTAG

Find three different ways to align these sequences with each other by inserting a variable number of gaps (0, 2, and 4) into the sequences in Table 2.

Scoring schemes are used to find the optimal alignment, which contains a maximum number of matches and minimum number of mismatches and gaps. Your task is to find the optimal alignment by employing two different scoring schemes:

Scoring scheme 1: Match: 2, mismatch: 0, gap: -5
Scoring scheme 2: Match: 2, mismatch: -1, gap: -2

Table 2. Scoring schemes to find the optimal alignment under each scoring scheme.

	Alignment	Matches	Mis-matches	Gaps	Score 1	Score 2
1	**No gaps:** ACCCCAGGCTTA ACCCGGGCTTAG					
2	**2 gaps:**					
3	**4 gaps:**					

Question 2. Which alignment is the optimal alignment under consideration of scoring scheme 1 and 2, respectively?

Exercise 3 – Transitions and transversions in a sequence data matrix of a protein-coding gene fragment (cytochrome c oxidase subunit 1 gene)

Program used: **MEGA**.

A data set containing two partial sequences of the protein-coding mitochondrial gene COI (*cytochrome c oxidase subunit 1*) is provided (**Two_sequences.fas**).

Open the file with the program MEGA (**File > Open File/session > path to your file**) when prompted choose ANALYSE – NUCLEOTIDE SEQUENCES – PROTEIN-CODING – INVERTEBRATE MITOCHONDRIAL).

To count pairwise nucleotide differences, on the main screen under the button Distance you can count the overall number of substitutions as well as transitions and transversions using the function '**Compute pairwise distances**' (options: 'nucleotides, p-distance', 'no of differences'; see Fig. 3).

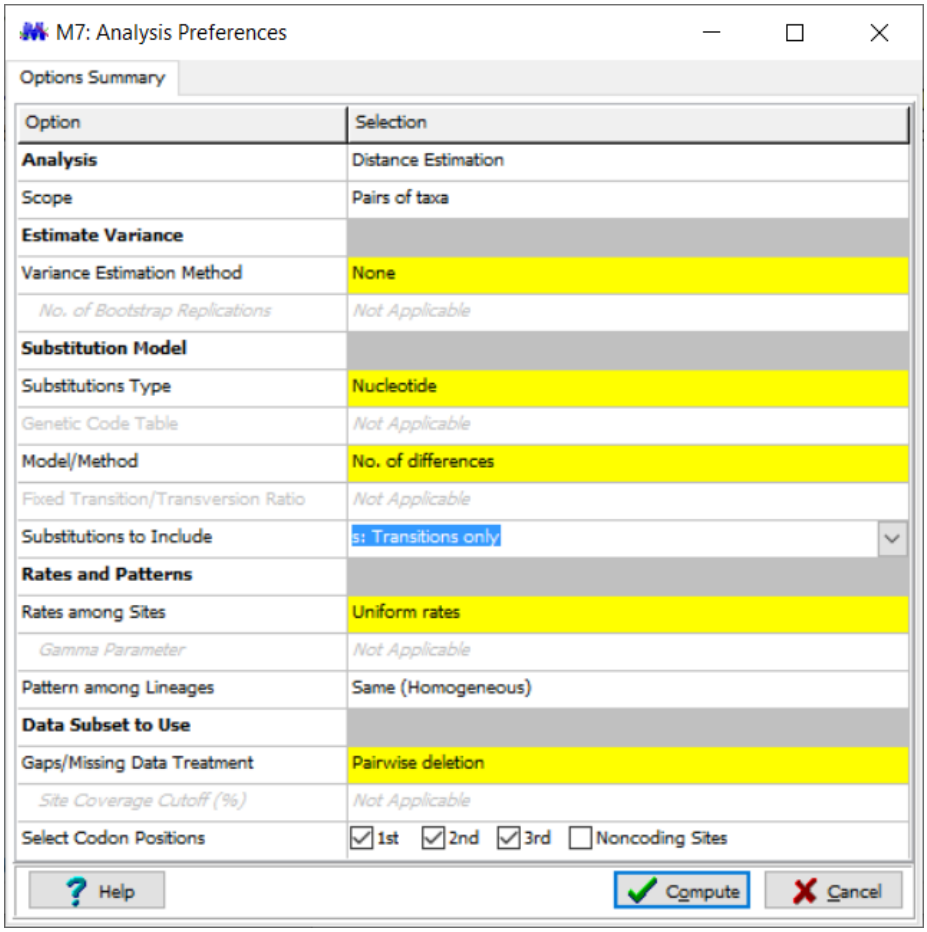


Figure 3. Pop-up window to count each, total number of pairwise nucleotide substitutions, numbers of transitions and transversions.

Question 3. What are transitions and transversions?
Generally, how many ways are there for nucleotide exchanges to result in transitions and transversions, respectively? Based on these numbers, what would you expect to observe more frequently, transitions or transversions?

Question 4. How many transitions and transversions to you observe in the pairwise sequence comparison (file: two_sequences.fas; Table 3). Do these numbers fit the expectation from Question 4?

Table 3. Numbers of transitions and transversions observed by comparison of different COI sequences of land snails.

	Total number of substitutions	Number of transitions	Number of transversions
Pairwise sequence comparison			

Exercise 4 – Silent and neutral nucleotide substitutions

Program used: **MEGA**.

A data set containing partial sequences of the protein-coding mitochondrial gene COI (*cytochrome c oxidase subunit 1*) is provided containing 149 DNA sequences of land snails endemic to the island of Timor (**Parachloritis_COI.fas**).

This sequence data set contains sequences that differ in length. So create a sequence alignment that can be analysed, you need to perform a multiple sequence alignment.
You can do this in MEGA by using the option:

File > Open A File/Session > Select file > Command: Align

I recommend to save the alignment before continuing with the exercise.

To translate the nucleotide sequences into amino acid sequences, it is necessary to select the correct translation table (**Data > Select Genetic Code Table**) and to select the correct reading frame (i.e., identify correct codon positions) (**Data > Select Genes and Domains**).

The genetic code is **invertebrate mitochondrial** (because the DNA sequences encode a mitochondrial gene from land snails).

The **reading frame is unknown** because we do not know at which position of the nucleotide alignment the first codon starts.

After opening the file in MEGA: The sequence alignment can be viewed by clicking the tab '**TA::**' in the top left corner.

Use the tab '**Select and Edit Genes/Domains**' to adjust the reading frame (start of the first complete codon). Translate the nucleotide sequence into amino acid sequence using the tab '**Phe > UUC**'

Once the reading frame has been identified and specified, in the Data window (click: **TA::**) you can count the numbers of conserved (C), variable (V) and parsimony informative (Pi) sites in the nucleotide and amino acid alignment, respectively.

Question 5. The reading frame of the partial COI alignment is unknown. How do you find out at which position in the sequence the first codon begins (or alternatively, at which codon position the first nucleotide in the alignment is)?

Table 4. Comparison of the total numbers of nucleotide and amino acid substitutions observed by comparison of different COI sequences of land snails.

	Total no of sites	Conserved sites (C)	Variable sites (V)	Parsimony informative sites (Pi)
Nucleotide alignment				
Amino acid alignment				

In addition, MEGA also calculates p-distances for nucleotide and amino acid alignments (p-distances are equivalent to the proportions of observed nucleotide / amino acid exchanges in pair-wise sequence comparisons).

Calculate the overall mean p-distances in each the nucleotide and amino acid alignment (**from the Main Window: Distance > Compute overall mean distance> p-distances each for nucleotides and amino acid sequences**) using the option no of differences (see Fig. 4).

Table 5. Comparison of the overall mean p-distances of nucleotide and amino acid substitutions.

	Nucleotide alignment	Amino acid alignment
No of differences		
Total length of alignment		

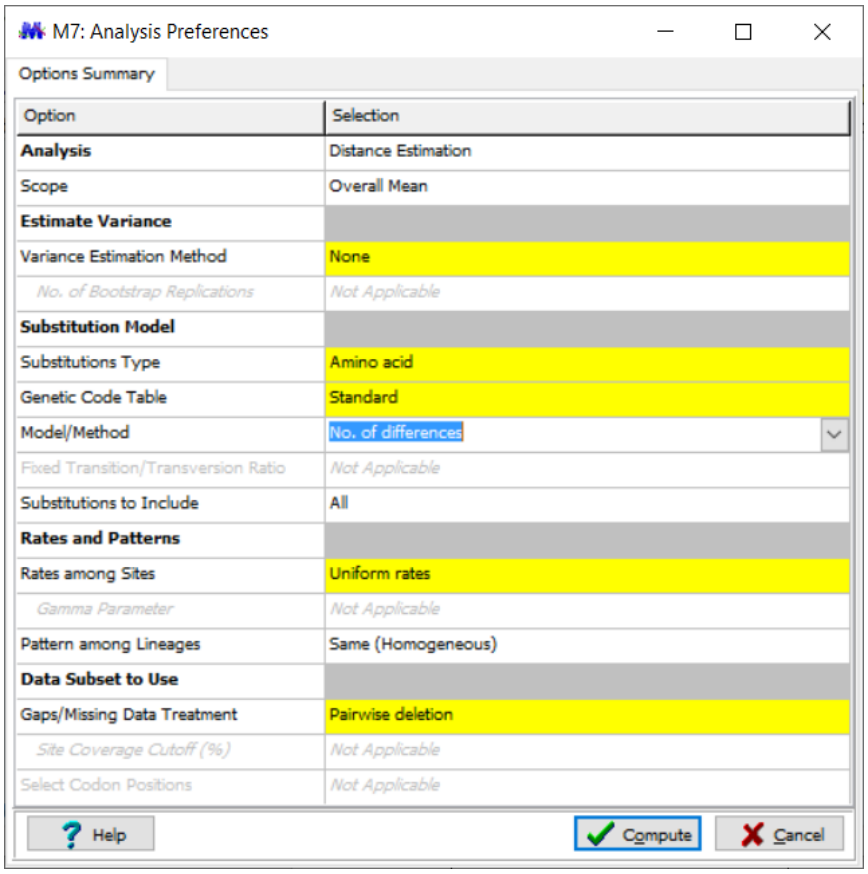


Figure 4. Pop-up window to compute overall Mean distance based on amino acid sequence.

Question 6. Compare the numbers of observed nucleotide and amino acid substitutions in the provided COI data set (Complete Table 4).

Explain the observed relationship between the proportions of conserved/un-conserved sites in the nucleotide and amino acid alignments?

Question 7. Given the mean numbers of substitutions in the nucleotide and amino acid alignment, respectively (Complete Table 5), what is the mean proportion of silent mutations of all nucleotide substitutions in the given nucleotide dataset.

Exercise 5 – Genetic distances in multiple DNA sequence alignments

While p-distances measure the observed proportion of nucleotide substitutions in the data set, MEGA also estimates corrected distances under the assumption of different models of DNA sequence evolution.

Calculate **overall mean p-distances** and compare them with **overall mean distances estimated under different substitution models** (Table 6).

Table 6. Uncorrected and corrected overall mean distances for the *Parachloritis* data set.

	p-distance	Jukes-Cantor	Kimura-2-Parameters	Tamura-Nei
Overall mean distance				

Question 8. Explain the differences between the uncorrected (p-distances) an corrected distances (JC, K2, TN) for the same sequence data set.

The sequence dataset of *Parachloritis* contains sequences of 21 species of *Parachloritis* and three sequences of species from different genera, which are used as outgroup in subsequent phylogenetic analyses.

Using the tab **Data > Setup Taxa and groups** you can group sequences by species (Fig. 5).

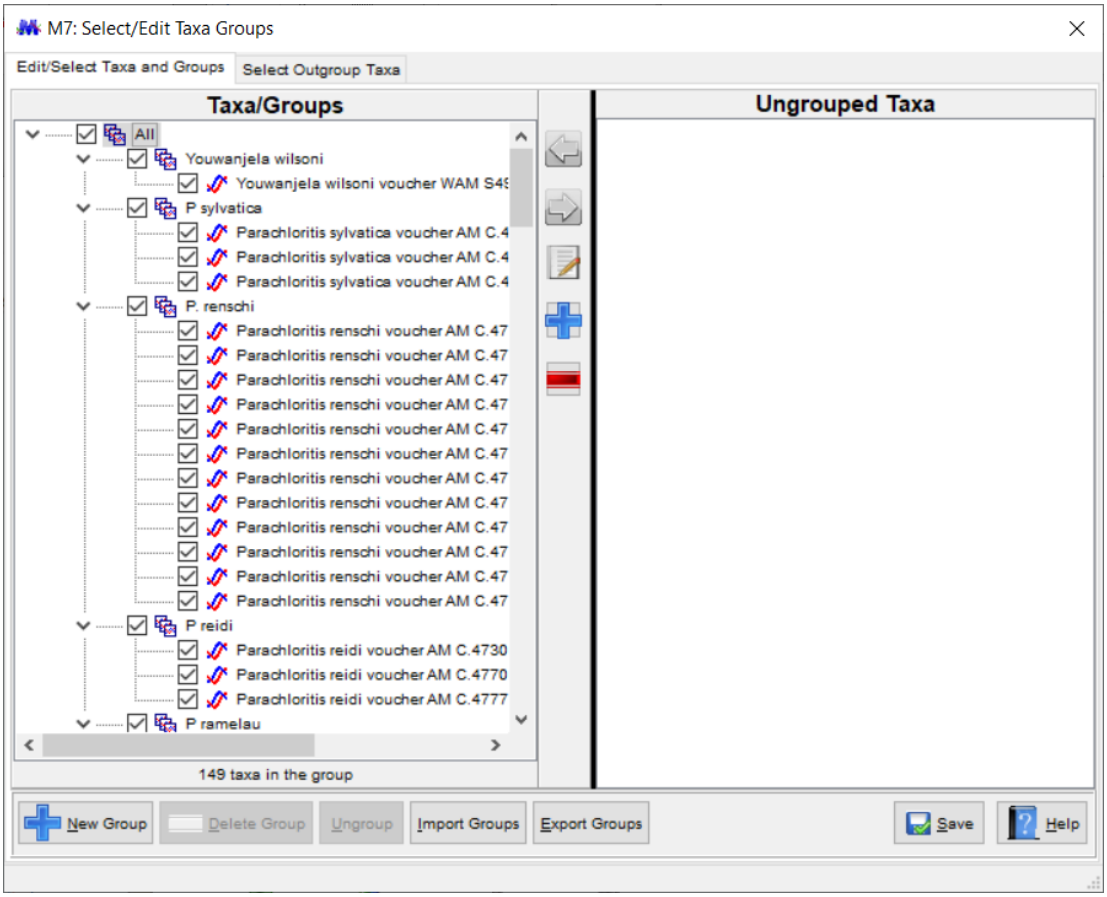


Figure 5. Pop-up window to setup groups ('Import Groups': file 'group_names.grp').

To assign sequences to different species, import the file '**group_names.grp**'.
Unselect the outgroup species (*Arnemelassa wilsoni*, *Chloritis sudestensis*, and *Youwanjela wilsoni*) in the **taxa tab**. Calculate mean intraspecific genetic p-distances under tab **Distance > Compute within group mean distances**. Calculate mean interspecific genetic p-distances under tab **Distance > Compute between group mean distances**.

Practical 6 (Week 7) – Molecular computer lab 2

Dr Frank Koehler – Australian Museum

E26 Teaching Lab 3

Exercise 1 – Multiple sequence alignment of non-protein coding sequences

For this exercise we use a dataset containing sequences of Australian land snails (family Helicarionidae). These sequences are of the mitochondrial 16S RNA, which is NOT a protein-coding gene.

The 16S sequence dataset used herein contains several unaligned, homologous sequences of varying length. Before these sequences can be analysed, they need to be aligned with each other (i.e., a sequence alignment needs to be built in which homologous nucleotides across all sequences in the dataset occupy the same column in the data matrix).

To produce a multiple sequence alignment, open MEGA7, and chose the option '**Align > Edit/Build Alignment > Retrieve sequences from a file**" to open the file **16S_sequences.fas**.

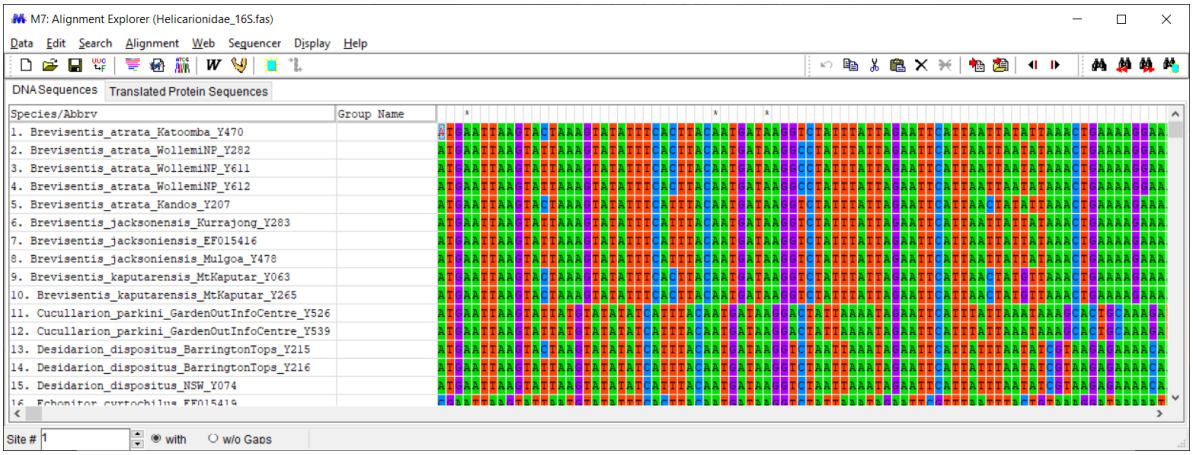
MEGA7 offers two different algorithms to align multiple DNA sequences with each other.

We will explore if the choice of alignment method may influence the outcome of our phylogenetic inference based on the available sequence data set.

First, use the **MUSCLE** algorithm ("Align DNA") with default settings to produce a multiple sequence alignment. Save the multiple sequence alignment as a new file ('16S_MUSCLE.fas') using the tab "**Data > Export alignment > Fasta format**".

The window should look like shown below.

The **Muscle** button is the nineth from the left, the **Clustal** button is the eight from the left.



Repeat the alignment procedure from scratch with the **CLUSTAL** tab, which uses a different alignment algorithm. Save this alignment as ('Helicarionidae_16S_CLUSTAL.fas').

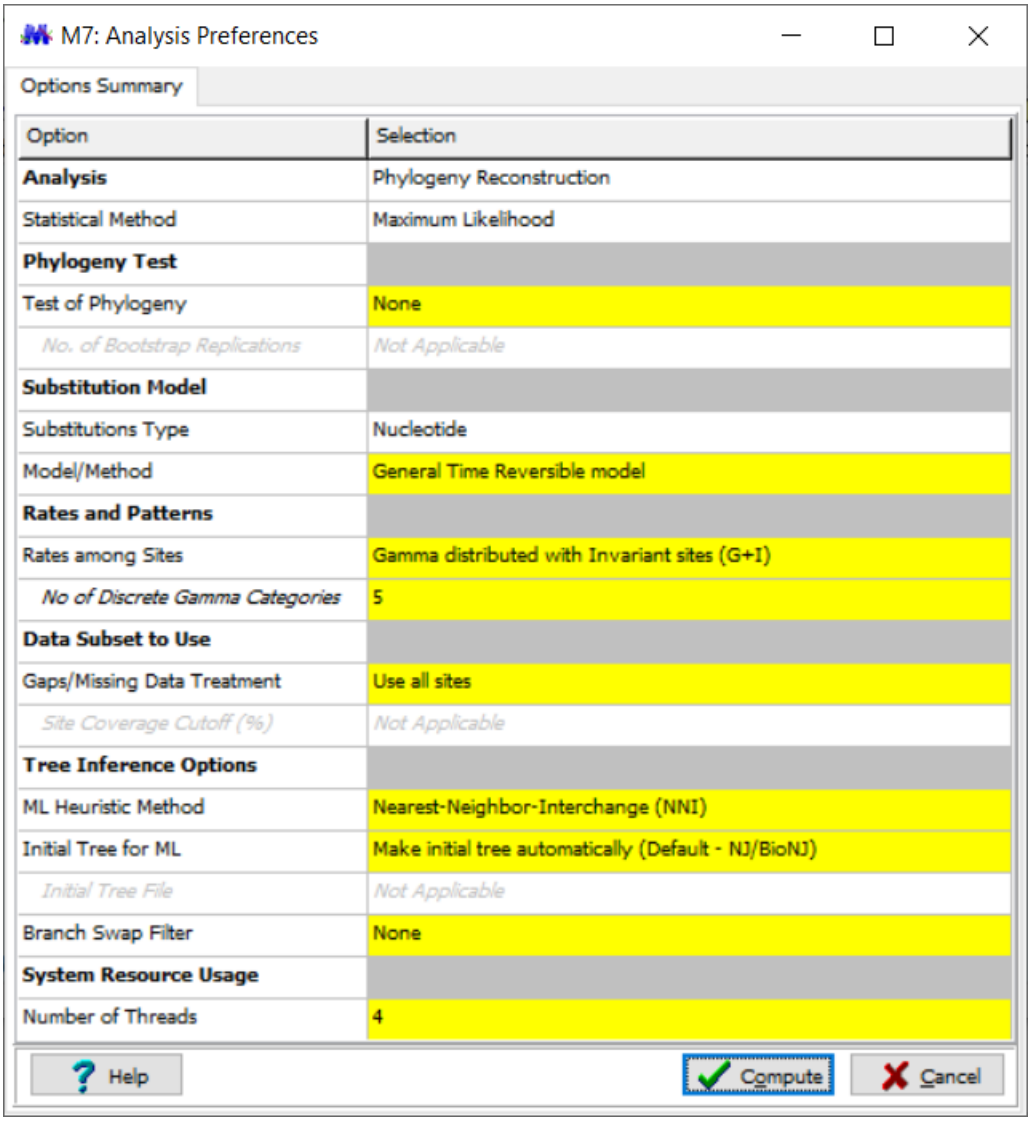
Open both alignment files in a viewer (**BioEdit**), sort the sequences in alphabetical order and save the files.

Question 1. Compare the alignments produced with Clustal and Muscle. Are they identical? (For example, do the alignments have the same total length?

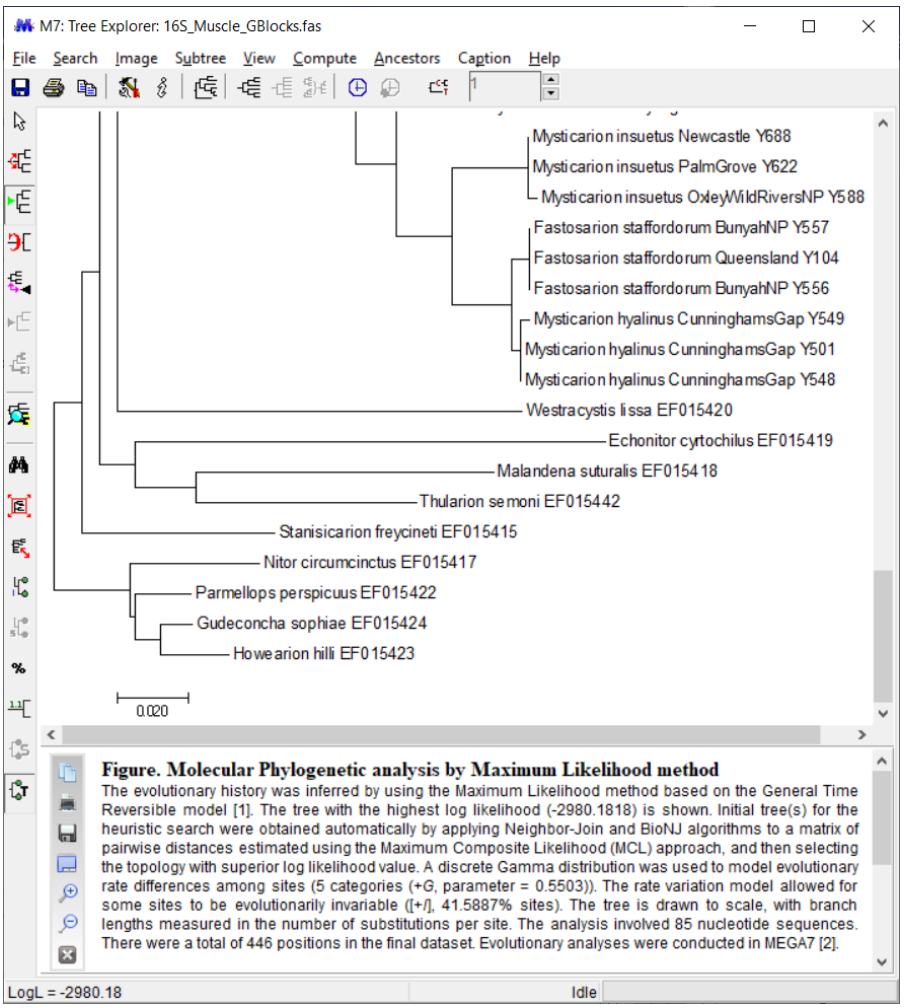
Open both alignment files in MEGA (not at the same time, one after another) and calculate a Neighbour Joining (NJ) tree using the tab "**Phylogeny > Construct/Test Maximum Likelihood Tree**" using the following settings:

Test of phylogeny > None
Substitutions type > Nucleotides
Model/method > General Time Reversible
Rate among Sites > Gamma distributed with Invariant Sites

See below.



Once the analysis has finished, select the clade containing the sequences *Gudeoconcha sophiae*, *Parmellops perspicuus*, and *Howearion hilli* (**Mouse click**) and set it as the **Root of the tree** (see below).



Question 2. Compare the topologies of the two trees produced for the Clustal and Muscle alignments, respectively. Are they identical?

Now, let's see if removal of ambiguously aligned positions can help producing more consistent results.

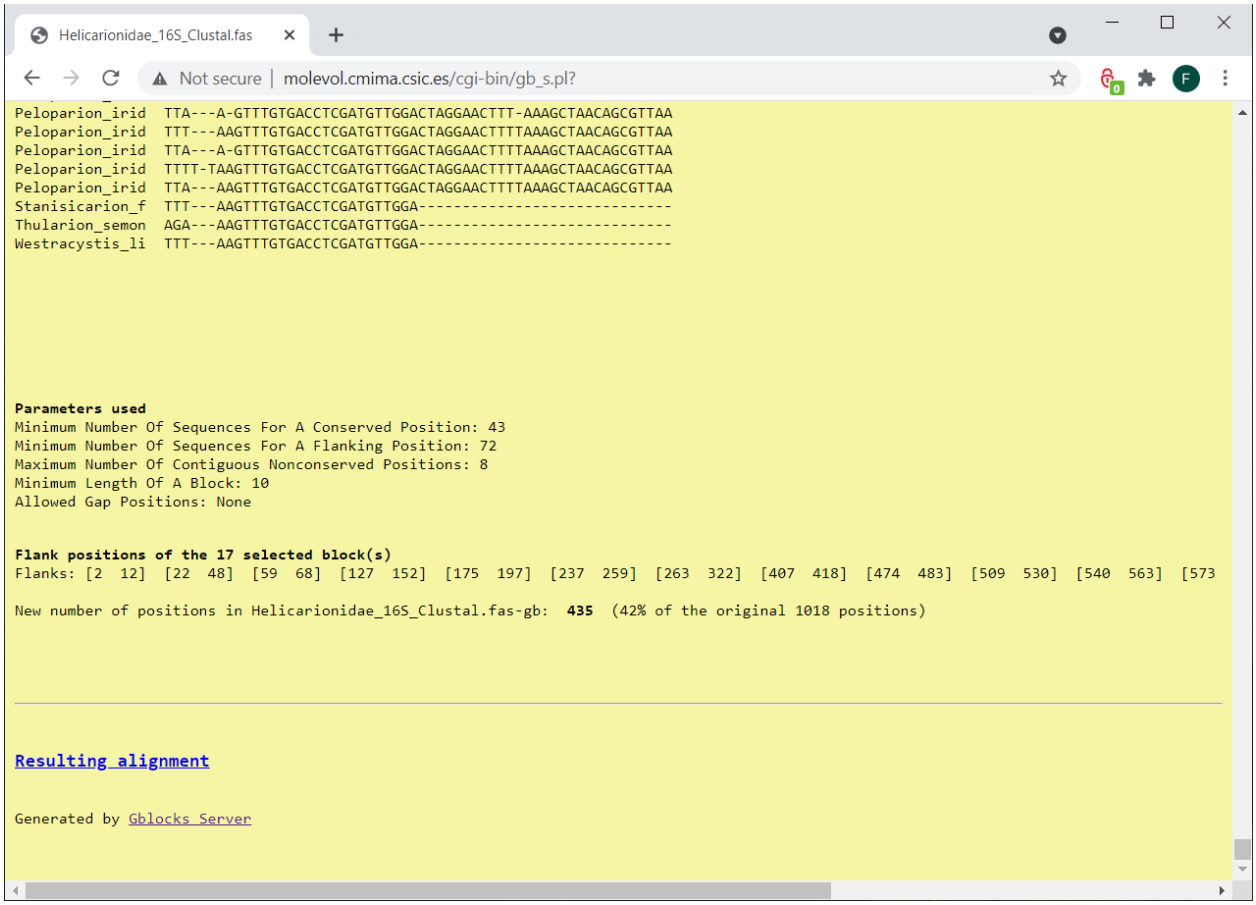
The program **Gblocks** is one of several software solutions that are available to eliminate ambiguously aligned parts of multiple sequence alignments.

The online interface of Gblocks is available from:

http://molevol.cmima.csic.es/castresana/Gblocks_server.html

Upload one alignment at the time (the ones produced using Clustal and Muscle), select '**DNA**' and '**Options for a more stringent selection**', and click '**Get Blocks'** to produce an alignment from which ambiguously aligned positions have been removed.
When finished, download the resulting sequence alignment using "**Resulting alignment**" link at the bottom of the new page (Safe as > e.g., '**16S_ Muscle_Gblocks.fas**').

Please also note down how many residues have remained after removing unreliably aligned parts (see below).



Repeat the NJ tree reconstruction using the alignments with ambiguous positions removed by Gblocks instead.

Question 3. How many alignment positions have been removed from the original sequence alignments by Gblocks for the Clustal and Muscle alignment, respectively?

Question 4. Compare the topologies NJ trees for the two different alignments with ambiguous positions removed by GBlocks with each other (make sure the trees are rooted with the same outgroup).

Exercise 2 – Substitution saturation

The genetic divergence between two sequences is estimated from the observed number of nucleotide substitutions (p-distances) based on models of sequence evolution that differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution.
For example, in protein-coding genes different substitution rates may be applied for transitions and transversions, for the three different codon positions, and for silent and non-silent substitutions.

As a result there is a difference between observed proportion of nucleotide substitutions (p-distances) and modelled distances (the latter being usually higher to compensate for unobserved substitutions.

Substitution saturation describes the phenomenon that in pairwise sequence comparisons a large, but unknown proportion of residues has been hit by more than one mutation since their divergence from the last common ancestor. Substitution saturation poses a problem to character-based methods of phylogenetic interference (especially Maximum Parsimony, but also Maximum Likelihood).

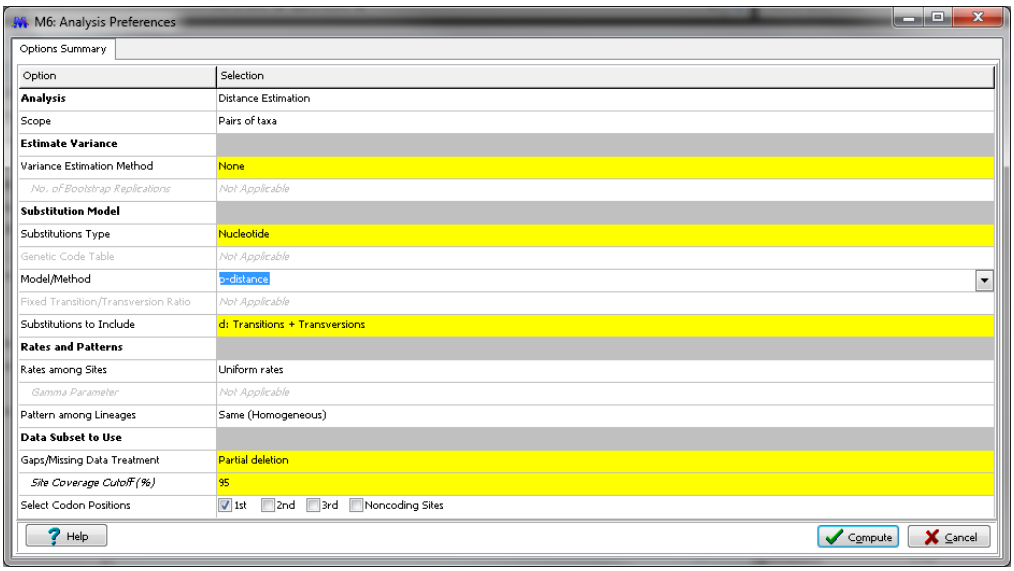
Sequence saturation can be detected by plotting p-distances and modelled distances against each other.

Open the dataset '**COI_data.fas**' with MEGA (**Data > Open a file/session**) and use the settings for **nucleotide sequences > protein-coding sequences > mitochondrial DNA**.

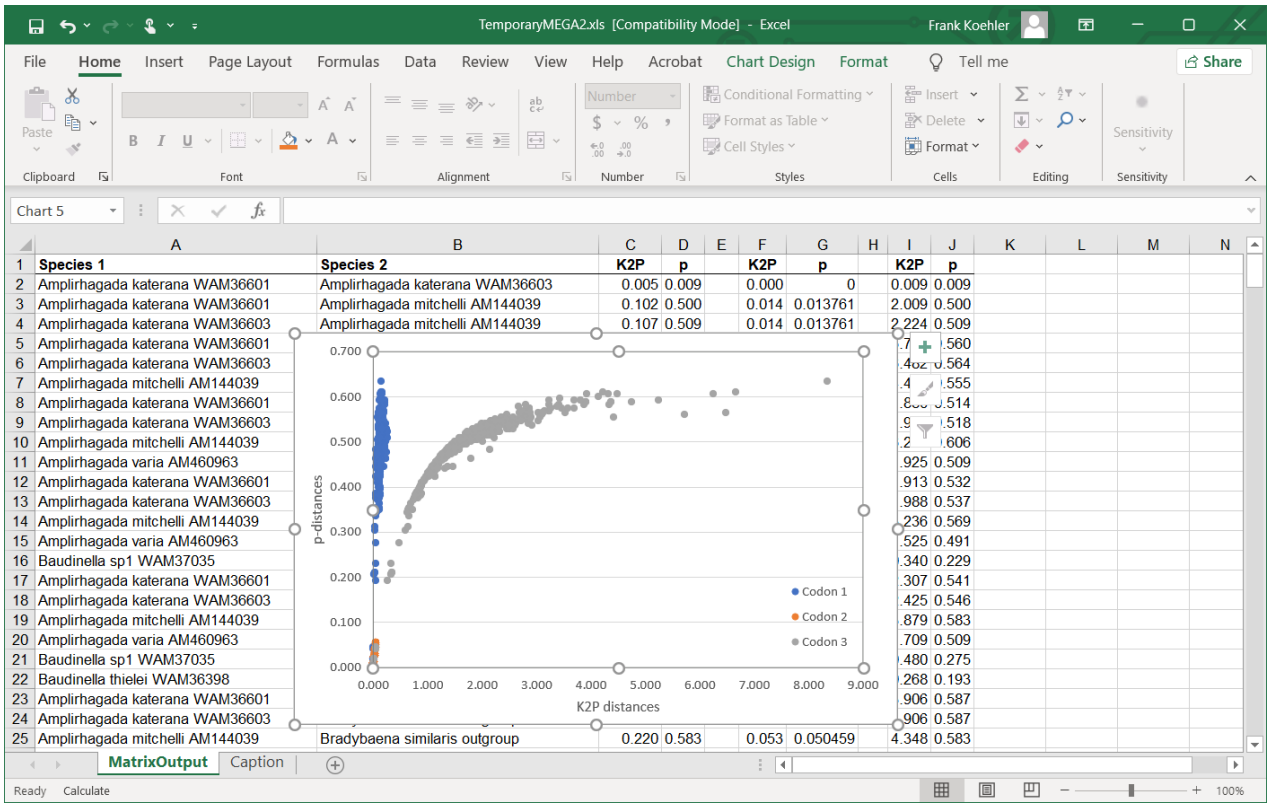
The alignment has already been trimmed so that the first alignment position is also the start of the first codon.

Use MEGA to calculate corrected pairwise Kimura-2-Parameter distances (Tab: **Distance > Compute Pairwise Distances**) for each codon position separately (select codon positions as shown below).

You can select the respective codon position in the last line of the pop-up window.



Export all distances from the results matrix to Excel using the 'XL' button (**as column; Print Save matrix**). Do this for all three codon positions and copy all columns into one Excel tab as 1st, 2nd, 3rd, respectively (e.g., columns C, F, and I in the example below).



	Species 1	Species 2	K2P	p	K2P	p	K2P	p
2	Amplirhagada katerana WAM36601	Amplirhagada katerana WAM36603	0.005	0.009	0.000	0	0.009	0.009
3	Amplirhagada katerana WAM36601	Amplirhagada mitchelli AM144039	0.102	0.500	0.014	0.013761	2.009	0.500
4	Amplirhagada katerana WAM36603	Amplirhagada mitchelli AM144039	0.107	0.509	0.014	0.013761	2.224	0.509
5	Amplirhagada katerana WAM36601							.560
6	Amplirhagada katerana WAM36603							.564
7	Amplirhagada mitchelli AM144039							.555
8	Amplirhagada katerana WAM36601							.514
9	Amplirhagada katerana WAM36603							.518
10	Amplirhagada mitchelli AM144039							.606
11	Amplirhagada varia AM460963							.925 0.509
12	Amplirhagada katerana WAM36601							.913 0.532
13	Amplirhagada katerana WAM36603							.988 0.537
14	Amplirhagada mitchelli AM144039							.236 0.569
15	Amplirhagada varia AM460963							.525 0.491
16	Baudinella sp1 WAM37035							.340 0.229
17	Amplirhagada katerana WAM36601							.307 0.541
18	Amplirhagada katerana WAM36603							.425 0.546
19	Amplirhagada mitchelli AM144039							.879 0.583
20	Amplirhagada varia AM460963							.709 0.509
21	Baudinella sp1 WAM37035							.480 0.275
22	Baudinella thielei WAM36398							.268 0.193
23	Amplirhagada katerana WAM36601							.906 0.587
24	Amplirhagada katerana WAM36603							.906 0.587
25	Amplirhagada mitchelli AM144039	Bradybaena similaris outgroup	0.220	0.583	0.053	0.050459	4.348	0.583

Repeat the whole procedure, but this time calculate uncorrected **p-distances** instead and paste these distances into the same Excel tab into the column following the Kimura-2-Parameter distances for the respective codon position (e.g., columns D, G, J as shown above).

Plot the Kimura-2 distances and p-distances against each other using the X-Y plotting function of excel for each codon position (**INSERT > Charts > Scatter Plot**).

After creating a X-Y plot for the distances in the first codon positions, you can right click on the plot area to activate the function "Select data" enabling you to rename your data columns as well as to add additional data series to the same plot.

The plot of all three codon positions combined should look as shown above.

Question 5. Describe what kind of relationships you observe between pairwise p-distances and K-2-P distances for each separate codon position. Which codon position reveals a clearly non-linear relationship and how can this be explained?
How else do the distances at the three different codon positions differ from each other?

Exercise 3 – Identifying the best-fit model of sequence evolution

For this exercise, please use the data set **COI_data.fas**.

Maximum Likelihood and Bayesian Inference-based methods of tree reconstruction are reliant on the choice of a model of sequence evolution that fits the present dataset.

Using the 'Models' Tab (**Find Best DNA/Protein Models ML**), determine the best-fit model of sequence evolution for the COI data set (make sure you are using all sites).

Question 6. What is the best-fit model of sequence evolution for the present COI dataset (by means of the Bayesian Information Criterion [BIC] and the corrected Akaike Information Criterion [AICc])?

Exercise 4 – Phylogenetic tree reconstruction using different methods

For this exercise use the data sets **COI_data.fas** and **28S_data.fas**, please, which contain sequences of two different genes from the same set of individuals of Australian land snails.

Reconstruct Maximum Parsimony [MP], Neighbour Joining [NJ] and Maximum Likelihood [ML] phylogenies based on analyses of the COI dataset (use: Test of Phylogeny: None, use all sites).

Use the appropriate model of sequence evolution as identified by the performed Modeltest when calculating the NJ and ML trees.

If necessary, re-root trees by selecting the outgroup taxon (*Cepaea nemoralis*).

Question 7. Do the topologies of the trees based on MP, NJ and ML analyses differ from each other (yes/no)? Complete Table 1 below (monophyly of genera supported/not supported)

Table 1. Analyses of the COI dataset ('**COI_data.fas**') – support for monophyly of genera. Please entder "**YES**" / "**NO**" in each cell for the support of monophyly of different genera (those represented by more than one species).

MONOPHYLETIC?	**MP**	**NJ**	**ML**
Amplirhagada			
Baudinella			
Setobaudinia			
Torresitrachia			
Xanthomelon			

Perform Modeltest runs and phylogenetic analyses (MP, ML, NJ, without Test of Phylogeny) for the 28S sequences (28S_data.fas).

Question 8. Complete Table 2 (monophyly of genera supported/not supported) by analyses of the 28S dataset.

Table 2. Analyses of the 28S dataset – support for monophyly of genera. Please enter "**YES**" / "**NO**" in each cell for the support of monophyly of different genera (those represented by more than one species).

MONOPHYLETIC?	**MP**	**NJ**	**ML**
Amplirhagada			
Baudinella			
Setobaudinia			
Torresitrachia			
Xanthomelon			

Question 9. Across the analyses of both sequence datasets, the monophyly of which genera is consistently supported / rejected? (possible answers: consistently rejected / consistently supported / ambiguous)

Exercise 5 – Inferring support of tree topology

Statistical support of the shown tree topology can be estimated by bootstrapping, which is a resampling test that provides an estimate of how statistically well-supported a phylogenetic tree is.

Bootstrapping is computationally expensive (requiring a lot of time) as effectively the same analyses is repeated for n times with a differently sampled data matrix. Therefore, we cannot infer the statistical support for all trees produced here in an appropriate manner.

Compute ML trees for the COI and 28S dataset under the use the option "Test of Phylogeny > 50 Bootstrap replicates".

From a statistical point of view any node receiving at least 95% of bootstrap support can be considered as well-supported.

Question 10. The monophyly of which genera is statistically well-supported by bootstrapping in the COI and 28S tree, respectively? Fill Table 4 ('yes' / 'no')

Table 4. The monophyly of genera consistently supported/rejected or ambiguous?

MONOPHYLETIC?	**COI**	**28S**
Amplirhagada		
Baudinella		
Setobaudinia		
Torresitrachia		
Xanthomelon		

Exercise 6 – Inferring a partitioned, multi-loci sequence dataset

Increasing the sampling of data (either number of taxa or amount of sequence data) should improve the reliability of phylogenetic analyses. Therefore, instead of analyzing each gene separately, one can build a combined dataset to analyze the different genes simultaneously.

Because different models have been suggested to best fit the individual sequence datasets, for ML analyses it is possible and recommended to treat each DNA fragment as a separate data partition in the ML analysis and to allow the analysis to estimate model parameters for each of these partitions separately while enforcing a single tree topology.

MEGA7 does not provide for the analysis of such a partitioned multi-loci dataset.

However, the program IQ-Tree offers a freely available Web-server that offers ML analyses with integrated modeltest.

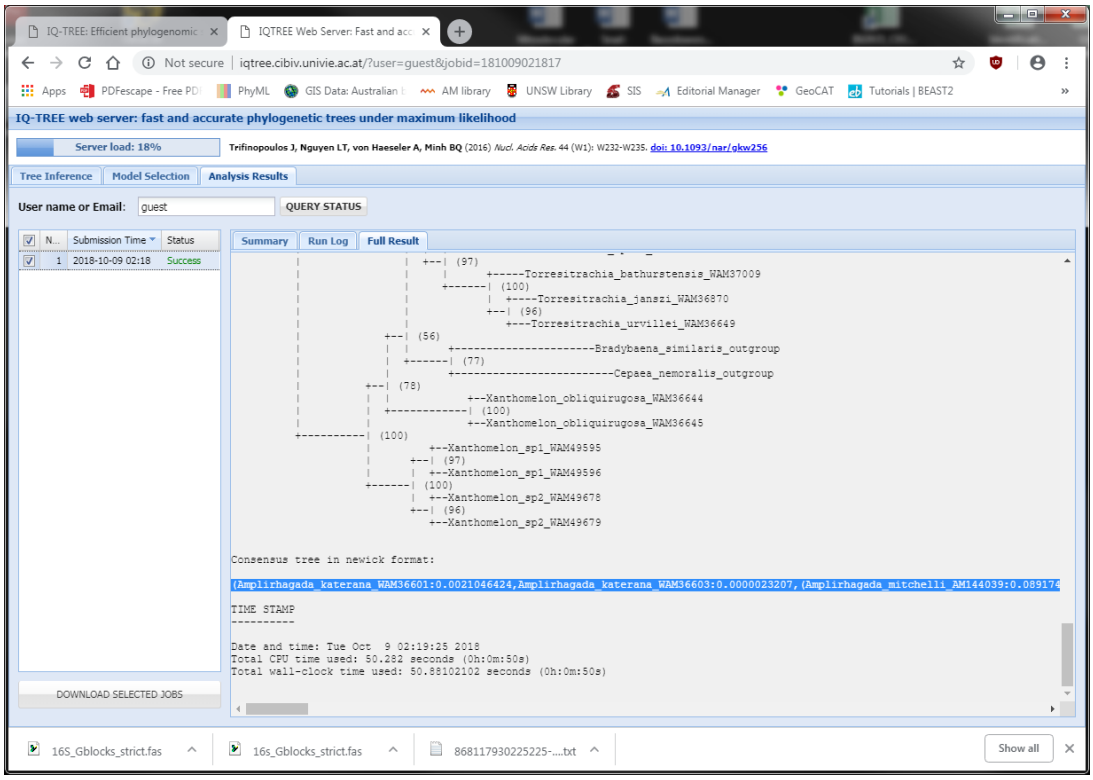
IQ-Tree can be found here: http://www.iqtree.org/ (choose the Web server from University of Vienna, Austria).

The user interface is easy to use: Upload the concatenated sequence data of three gene fragments (COI, 16S, 28S) named '**concatenated_data.fas'** under '**alignment file**', and upload the file specifying the data partition ('**partitions.txt**') under '**Partition file**'.

Choose a low number of ultrafast bootstraps (1000) as bootstrap method and run the analysis, which should not take long to run.

By default, IQ-Tree identifies the best-fit model of sequence evolution for each partition, which are then used to reconstruct the best Maximum Likelihood tree.

The result it looks like this:



Copy the Consensus file "(…..)" into a text document and save this document as "**IQ-Tree.tre**".

You can view this tree file in MEGA by using the tab "**User Tree**" (> "**Display Newick Trees**").
Table 5. The monophyly of genera consistently supported/rejected or ambiguous?

MONOPHYLETIC?	**IQ-Tree**
Amplirhagada	
Baudinella	
Setobaudinia	
Torresitrachia	
Xanthomelon	

ORAL PRSENTATION

Practical 7 (Week 9) – Student Powerpoint presentations

E26 Teaching Lab 3

EXTENSIONS WILL ONLY BE ALLOWED THROUGH SPECIAL CONSIDERATION APPLICATION.

You will give your oral presentation in Week 9 between 1-5 pm in E26 Teaching Lab 3. Also upload your presentation to Moodle. After each presentation questions will be asked for you to answer.

ORAL PRSENTATION TOPICS

SELECT ONE OF THESE TOPICS FOR YOUR ESSAY AND PRESENTATION

1. The whole of the Tree of Life is now being put together (http://phys.org/news/2015-09-tree-life-million-species.html). Describe what this involved, where we are at in this exercise, what are the challenges and what we can do with it?
2. It was traditionally thought that the close relatives of the insects were centipedes and relatives. Now insects are considered to be a group nested within Crustacea? This is an outcome of the implementation of molecular sequence data. Review this topic.
3. Molecular phylogenetics has impacted our understanding of the relationships of insect orders. Pick one of the Big 5 insect orders (Hemiptera, Coleoptera, Diptera, Lepidoptera and Hymenoptera) and discuss the molecular impact on classification and how it compares with traditional morphologically-based classifications.
4. About 60% of all organisms are insects. What do we known about the insect Tree of Life, what are the latest controversies, and what are the evolutionary novelties that are given that explain their success?
5. We only know about 20% of all species on the planet. Some have argued that taxonomy is an old fashioned science (http://www.independent.co.uk/environment/nature/taxonomy-the-naming-crisis-2240872.html). Review this topic, including its relevance today and the value in documenting life on Earth.
6. Most of the modern orders of birds first appear in the fossil record of the early Paleogene, suggesting that ordinal diversification occurred after the dinosaurs went extinct at the Cretaceous-Tertiary boundary. In contrast, genetic estimates usually point to a Cretaceous diversification of bird orders. Which of these scenarios has stronger support?

CRITERIA FOR GRADING OF STUDENT PRESENTATION

Assessment Criteria	% of mark	Your mark
Content • Present the subject matter in an engaging fashion • Explain the scope of your project • Demonstrate that you understand any controversy or alternative theories • Indicate how you undertook your project (methods and materials) • Show your results in a concise fashion • If you have drawn any conclusions make sure that you put them into your own words • In your conclusions make sure that you demonstrate an understanding of phylogenetics • Indicate any future directions	**60**	
Speech • Make sure that you speak clearly (practice cures a lot of nervousness) • Speak to your audience (not your shoes) • Demonstrate your interest in your project	**20**	
Slide quality • Make sure that your graphics are of the correct resolution and are relevant (please run through them beforehand) • Do not overdo animations • Try to be innovative in your slides – make them standout if you can (but do not overdo the colour schemes)	**20**	
TOTAL	**100**	

BIOS3221 Smiths Lake Field Trip

15-19 October 2022

AIMS

The Smiths Lake exercise focuses on methods in insect taxonomy and systematics, with emphasis on insects living on plants. You will learn the following:

1) techniques in sampling insects;
2) preparing a field note book;
3) preserving and curating specimens;
4) identifying and diagnosing insects to orders;
5) phylogenetic analysis of insect orders.

MATERIALS

You will be provided with a detailed Smiths Lake Field Trip document on Moodle that includes all the materials for the field trip.

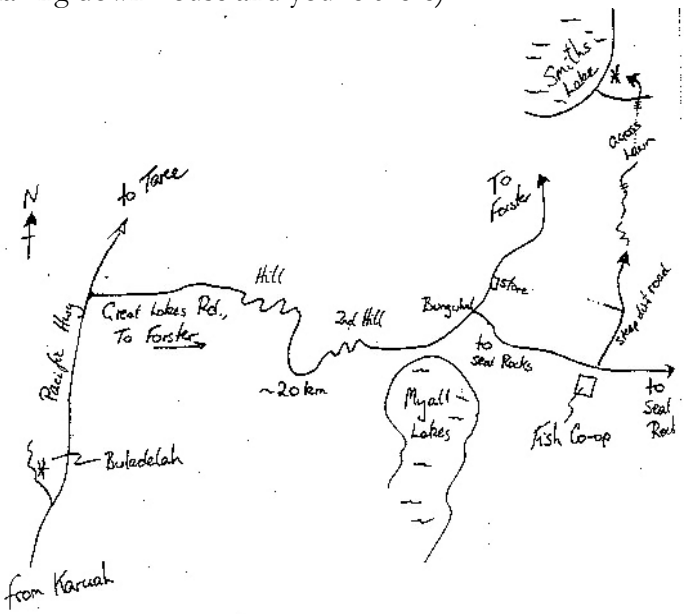
SMITHS LAKE ASSESSMENT

The Smiths Lake assessment will involve the following components on which you will be graded:

1) 30 curated insects identified to order per person (20% of total course grade)
2) Phylogenetics of insects (10% of total course grade)
3) Group presentation (5% of total course grade)

WHERE IS SMITHS LAKE FIELD STATION?

Smiths Lake is approximately 320km from UNSW, 30mins north of Bulahdelah. The following directions will get you there: take the F3 freeway north from Sydney, at the end of that follow the Pacific Highway until Bulahdelah. Just north of Bulahdelah it turns into a Freeway again. Take the right turn to Forster via the Great Lakes Way just north of Bulahdelah. If you have gone more than 10km out from Bulahdelah you have gone too far. The turn is very well signposted. Travel along the Lakes Way for about 26km and turn right onto Seal Rocks Rd, just as the houses appear at Bungwahl. If you pass the Bungwahl Store, you have gone too far. Travel along this road for about 3km. Past the Bungwahl School, and off to your right you should see a big ugly blue and white fish co-operative. Turn left (the opposite direction to the fish co-op) and travel straight along the dirt road for 1.6 km (don't veer left), past the falling down house and you're there).



SMITH LAKE FACILITIES AND INFRASTRUCTURE

Smiths Lake Field Station has open fires and gas cookers for cooking, electric lights, hot showers and toilets, dormitory-style accommodation with bunks/mattresses, lots of cold open air, no nearby shops, no phone (public phone is 5 km away - in an emergency you may be contacted via the school office (02 9385 2126). Mobile phone coverage is still minimal!

WHAT TO BRING

1) Assume that it will **rain**: you should bring wet-weather gear to work in.
2) You will sample at night, so bring a flashlight/headlamp.
3) Please bring your course manual and writing materials.
4) All students should bring sleeping bags/sheets + blankets, pillow/pillowcase, towels, soap toothbrush etc.
5) Plenty of warm clothes, swimming costume, old shoes that can get wet/boots, raincoat, hat, sunscreen, football, musical instruments, torch (flashlight).
6) **No hair dryers allowed (they blow the circuits).**

YOUR OTHER DUTIES

1) As you are such a small group you will be preparing your own meals. Please keep all communal areas clean and tidy and pack away any leftover food so as to not attract rodents etc.
2) Inform BEES technical officer of any special dietary requirements (e.g vegetarians or religious requirements)

BEHAVIOUR

Participants are responsible for:

- Adopting a responsible attitude whilst on the fieldtrip.
- Reading and following any instructions or notices produced relating to the field activity, including attending any briefing sessions and returning any forms to the staff members in charge.
- Seeking instructions if they are unsure of what is required.
- Not operating any equipment that they are unfamiliar with.
- Complying with instructions and directions issued by instructors. Please note that Prof. Cassis is responsible for all staff and students. Joanne Wilde is the person you should listen to most regarding operational issues concerning daily activities and how Smiths Lake works.
- Taking action to avoid, eliminate or minimize risks.
- Avoiding, as far as possible, exposure to venomous animals and plants likely to cause allergic reactions. If there is risk of exposure, steps should be taken to minimize risk (e.g., wear appropriate clothing, apply insect repellant, carry appropriate treatment for hay fever etc).
- **Please inform Guy Taseki or Gerry Cassis if you have any special medical issues, such as allergies etc, before the commencement of the trip.**
- Ensuring that adequate protection from sun and cold weather is carried and used. This includes hat, sunglasses, lip balm, sunscreen, warm clothing, and rainproof or windproof jacket.
- Carrying sufficient water, minor medical necessities (headache tablets, bandaids) and minor emergency food (e.g Chocolate bar). Try to drink at least 2 litres of water per day.
- Making use of all safety devices and personal protective equipment.
- Reporting any unsafe conditions or hazards.
- Do not place any other person at risk the health and safety on the fieldtrip or any member of the public by your own actions or omissions. **Act responsibly and do not engage any irresponsible actions**.

- Seeking information or advice on hazards and procedures where necessary before carrying out new or unfamiliar work.
- Being familiar with the emergency evacuation procedures and the location of first aid kits, personnel, and emergency equipment, and if appropriately trained, using emergency equipment. Joanne Wilde will advise.
- **Alcohol consumption is not banned and will be allowed after work is completed for the day. However, you are expected to drink responsibly and in moderation, and no one should drink to excess – intoxication will not be tolerated. There will be zero tolerance for drink and driving. You can arrange for purchase of alcohol as a group. You will be expected to be ready for work by 8 am each morning. We are allowing alcohol consumption so please respect our decision and act responsibly**.
- There will be no tolerance for use of non-prescription/recreational drugs during the field trip. Such incidences will be reported to the Head of School.
- Treating all other field participants and members of the public with courtesy and respect.

Insect Health and Safety

There are a range of Health and Safety issues that you must familarise yourself with and comply with. Please read carefully the Risk Assessments and Safe Work Practice documents provided to you at the Smiths Lake field station. There are risks associated with the field station and the field and lab work, and you will be instructed about those on your arrival at Smiths Lake. You will be notified before the field trips of Risk Assessments and Safe Work Practice that you will need to read on Safesys and declare as red.

You will be handling two types of hazardous compounds:

1) **Ethanol**. Ethanol is a preservation agent. You will be issued with and trained in the use of ethanol and you will need to comply with the **safe work practice** procedure. Ethanol is flammable.
2) **Ethyl acetate**. Ethyl acetate is a killing agent for insects. Use the ethanol safe work practice for ethyl acetate. Ethanol is flammable.

ACTIVITIES

DAY 1 – Saturday 15 October

10 AM:	Departure from UNSW. Meet 10am to assist with packing
2-3 PM:	Arrival, Health and Safety instructions.
4 PM	Accommodation and lab setup.
6 PM:	Dinner and clean up.
7.30 PM:	Outline of the field trip activities and exercises.

DAY 2 – Sunday 16 October

8 AM:	Breakfast and clean up.
8.45 AM :	Instruction on following: 1) How we will work together. 2) Insect collection. 2) Host plant collection. 3) Minimum data requirements for field notes. 4) Allocation of field gear.
9.30 AM – 1 PM:	Insect sampling.
1 PM – 2 PM:	Lunch and clean up.

2 PM – 6 PM: Continuation of insect and plant sampling
6 PM – 7.30 PM: Dinner and clean up.
8 PM – 9 PM: Insect curation.

DAY 3 – Monday 17 October
8 AM: Breakfast and clean up.
9 AM – 12.30 PM: Identification of specimens to insect order and families.
1 PM – 2 PM: Lunch and clean up.
2 PM – 5 PM: Continuation of insect identification.
6 PM – 7.30 PM: Dinner and clean up.
7.30 PM –10 PM: After dinner discussion

DAY 4 – Tuesday 18 October
8 AM: Breakfast.
9 AM – 1 PM: Construct insect phylogeny based on insect collection.
1 PM – 2 PM: Lunch.
2 – 4 PM: Complete individual insect collections and identification and present for marking.
4 – 5.45 PM Prepare presentations for evening.
6 PM – 7 PM: Dinner and clean up.
7 PM – 8.15 PM: Presentations.

DAY 5 – Wednesday 19 October
8 AM – 8.45 AM: Breakfast.
9 AM – 10.00 AM: Clean up and pack up.
10.30 AM: Departure.