Developing and applying supertree methods in Phylogenomics and Macroevolution

A thesis submitted to the National University of Ireland for the Degree of
Doctor of Philosophy

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Presented by:
Wasiu A. Akanni
Department of Biology,
NUI Maynooth,
Maynooth,
Co. Kidare, Ireland.

April 2014

Head of Department: Professor Paul Moynagh, BA (mod), Ph.D. (Dublin)
Supervisor: Dr. Davide Pisani, B.Sc., Ph.D. (Bristol)

Table of Contents

Index of Figures I

Index of Tables II

Index of Equations III

Abbreviations IV

Dedication VI

Acknowledgements VII

Declaration VIII

Abstract IX

Chapter 1: Introduction 1

1.1 Terms and Definitions 7

1.1.1 Trees 7

1.1.2 Tree resolution 8

1.1.3 Subtrees 10

1.1.4 Splits 11

1.1.5 Triplets 11

1.1.6 Nestings 12

1.1.7 Quartets 13

1.2 Building a Phylogeny 14

1.2.1 Getting the data 15

1.2.2 BLAST and Homology 16

1.2.3 Multiple sequence alignment (MSA) 17

1.2.4 Phylogenetic methods 19

1.3 Test of two trees 27

1.4 Tree merging and summarisation 29

1.4.1 Consensus of trees 29

1.4.2 Supertrees 33

Chapter 2: Implementation of a Maximum Likelihood and Bayesian (MCMC) supertree method 35

2.1 Overview 35

2.1.1 Supertree methods and their properties 36

2.1.2 Estimating support for supertree clades 40

2.2 Maximum LIkelihood (ML) supertree method 42

2.2.1 Tree representation 44

2.2.2 Searching the tree space for the elusive ML supertree 46

2.2.3 Extending test of two trees to supertrees 53

2.2.4 Likelihood Utility for Supertrees (L.U.St) Package 54

2.3 A Bayesian supertree method 56

Chapter 3: Testing on Case Studies 58

3.1 Introduction 58

3.2 Methods 60

3.2.1 Bias testing 60

3.2.2 Analysis of the *Drosophila* data set 61

3.3 Results ... **62**
3.3.1 Testing for biases ... 62
3.3.2 The *Drosophila* data set ... 66
3.4 Discussion ... **69**
3.5 Conclusions ... **70**

Chapter 4: Reanalysis of Real world Data sets ... **71**
4.1 Introduction ... **71**
4.1.1 The Metazoan dataset ... 72
4.1.2 The carnivore data set ... 73
4.2 Methods ... **75**
4.2.1 Supertree analysis of the Metazoa ... 75
4.2.2 Supertree analysis of the Carnivores ... 76
4.2.3 Statistical test of metazoan and carnivore supertrees ... 78
4.3 Results ... **79**
4.3.1 Bayesian (MCMC) metazoan phylogeny ... 79
4.3.2 Bayesian (MCMC) carnivore phylogeny ... 81
4.4 Discussion ... **91**
4.5 Conclusion ... **93**

Chapter 5: Tree of Life ... **94**
5.1 Introduction ... **94**
5.2 Methods ... **97**
5.2.1 Data acquisition ... 97
5.2.2 Cluster of orthologous proteins ... 97
5.2.3 Building gene trees ... 99
5.2.4 Supertree analysis ... 100
5.2.5 Testing previously proposed positions of the Eukaryotes ... 100
5.2.6 Identification of rogue taxa ... 101
5.3 Results ... **102**
5.3.1 The Prokaryote Supertree ... 102
5.3.2 The Archaeabacterial Supertree ... 108
5.3.3 The Eubacteria Supertree ... 110
5.3.4 The position of the Eukaryotes ... 114
5.4 Discussion ... **115**
5.5 Conclusion ... **116**

Chapter 6: General Discussion & Conclusions ... **118**

Chapter 7: Future prospective ... **123**

Chapter 8: Bibliography ... **124**

Appendices ... **152**
Appendix A ... **152**
Appendix B ... **159**
Appendix C ... **165**

Publications ... **175**

Index of Figures

Figure 1.1: Rooted phylogenetic trees with different levels of resolution. **9**
Figure 1.2: A tree and its subtree. A tree (a) and (b) a subtree of the tree in (a) **10**
Figure 1.3: Two rooted phylogenetic trees and their Adams consensus tree. **13**
Figure 2.1a: L.U.St Maximum Likelihood supertree search strategy..... **51**
Figure 3.1: Analyses of input tree shape bias. **64**
Figure 3.2: Analyses of input tree size bias. **65**
Figure 3.3: ML supertree analysis of *Drosophila* empirical dataset. **68**
Figure 3.4: A line graph of MRP parsimony scores and likelihood scores..... **69**
Figure 4.1: Bayesian (MCMC) phylogeny of the Metazoans. **84**
Figure 4.2: Phylogenomic supertrees of the Metazoan. **85**
Figure 4.3: Distribution of Metazoan supertrees likelihood scores. **86**
Figure 4.4: Bayesian (MCMC) phylogeny of the Carnivores. **88**
Figure 4.5: Bayesian (MCMC) Carnivore phylogeny excluding rogue taxa. **89**
Figure 4.6: Phylogenomic supertrees of the Carnivora with rogue taxa pruned. **90**
Figure 4.7: Distribution of Carnivore topologies likelihood scores. **91**
Figure 5.1: Bayesian (MCMC) phylogeny of the prokaryotes. **105**
Figure 5.2: Network visualisation of taxonomic equivalents in the Prok dataset..... **106**
Figure 5.3: Bayesian (MCMC) phylogeny of the prokaryotes after pruning the rogue taxa. **107**
Figure 5.4: Rooted Bayesian (MCMC) phylogeny of the Archaeabacteria. **109**
Figure 5.5: Bayesian (MCMC) phylogeny of the Eubacteria. **111**
Figure 6.6: Network visualisation of taxonomic equivalents in the Bac dataset. **112**
Figure 5.7: Bayesian (MCMC) phylogeny of the Eubacteria with the rogue taxa pruned. **113**

Index of Tables

TABLE 2.1: EFFICIENCY OF L.U.ST'S ML SEARCH STRATEGIES. **50**
TABLE 4.1: SUMMARY OF THE STATISTICAL TESTS OF THE METAZOAN SUPERTREES. **86**
TABLE 4.2: SUMMARY OF STATISTICAL TESTS OF THE CARNIVORE SUPERTREES.. ..... **87**
TABLE 5.1: SUMMARY OF THE STATISTICAL TEST OF THE EUKARYOTIC RELATIONSHIPS. **114**

Index of Equations

EQUATION 1: LIKELIHOOD RATIO THEOREM **23**
EQUATION 2: PROBABILITY THEOREM **24**
EQUATION 3: HYPOTHESIS RANKING THEOREM **25**
EQUATION 4: BAYES THEOREM **26**
EQUATION 5: INPUT TREE LIKELIHOOD THEOREM **43**
EQUATION 6: SUPERTREE LIKELIHOOD THEOREM **43**

Abbreviations

AA	Amino Acid
AU test	Approximately Unbiased test
BS	Bootstrap Support
DNA	Deoxyribonucleic Acid
E-value	Expectation value
EST	Expressed Sequence Tags
GTR	General Time Reversible
HGT	Horizontal Gene Transfer
ITSE	Input Tree Shape Effects
KH test	Kishino Hasegawa test
LBA	Long Branch Attraction
LGT	Lateral Gene Transfer
L.U.St	Likelihood Utility for Supertree
MCL	Markov Cluster algorithm
MCMC	Markov Chain Monte Carlo
MJCC	MJ-rule Component Consensus
ML	Maximum Likelihood
MP	Maximum Parsimony
MPTs	Multiple Parsimonious Trees
MRP	Matrix Representation with Parsimony
MSA	Multiple Sequence Alignment
MSS	Most Similar Supertree
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NJ	Neighbour Joining
NNI	Nearest Neighbour Interchange
PAUP	Phylogenetic Analysis Using Parsimony
PAUP	Phylogeny
PGM	Personal Genome Machine

Qs	Quality support
RAxMl	Randomized Axelerated Maximum Likelihood
RF	Robinson Foulds
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
SCC	Strict Component Consensus
SH test	Shimodaira Hasegawa test
SM	Supertree Method
SMRTs	Single Molecule Real Time sequencing
SNC	Strict Nestings Consensus
SPR	Subtree Pruning and Regrafting
SSU	Small Subunit

For my Family and Friends

Acknowledgements

As I sit here, as the clock reaches 12 midnight, and attempt to find words to turn into sentences to truly convey my deepest gratitude to all these people I will mention below, I find myself feeling truly blessed and I know there is truly no way I would be here at this penultimate moment of my PhD thesis without the immeasurable care, kindness, support, friendship and love of each and everyone of you.

To Dr. Davide Pisani, my supervisor, my mentor, my dear friend, I wish express my sincerest gratitude for your immense patience and understanding, opportunities, advice, and guidance on all things, science and beyond. I feel very lucky to have met you, got to know you, work with you, and learn from you over all these years. Thank you.

Next I would like to truly thank Dr. Mark Wilkinson, Dr. Christopher Creevey, Prof. James McInerney and Dr Peter Foster for all time, effort, and teachings during the course of this PhD.

My research would not possible without the funding and support of the Irish Research Council and the computing resources I have relied upon over the years, Darwin, Sioc, beagle, and doraemon.

I would also like to say thanks to my examiners, Prof. James McInerney and Dr. Marcello Ruta, for their time in reading this thesis. I am very grateful.

To every member of the Bioinformatics unit, Maynooth that was I lucky to meet, work and play with, I wish to express my heartfelt thanks for not only being my friends but also for all your various efforts to help get me to this point. A special thank you my one of kind friend Mrs Karen Siu-Ting for teasing me, Spanish lessons, caring for me when I was sick but mostly for pushing me to improve everyday.

To the gang at 141 St Michael's hill, it's been an absolute pleasure to be your housemate and friend. Thank you for all the memories.

To all my friends, both near and far, thank for your support, love, care. Special thank you to Femi Adeboye, Boye Karunwi, Kathryn Hennessy, Dayo Morakinyo, and all the Torpedo football team lads (CYOT).

To my Family, Mom and Dad, truly no words can express how much I appreciate all that you have done for me. Thank you for being the best role models any son could ask for. Thank you for all your constant encouragements, love, and care, and although I could never truly repay you, I am very lucky and grateful to be your son. To my wonderful siblings, thank you for never failing to be a source inspiration, motivation and happiness. Love you.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:_____

Wasiu Ajenifuja Akanni

Abstract

Supertrees can be used to combine partially overalapping trees and generate more inclusive phylogenies. It has been proposed that Maximum Likelihood (ML) supertrees method (SM) could be developed using an exponential probability distribution to model errors in the input trees (given a proposed supertree). When the tree-to-tree distances used in the ML computation are symmetric differences, the ML SM has been shown to be equivalent to a Majority-Rule consensus SM, and hence, exactly as the latter, it has the desirable property of being a median tree (with reference to the set of input trees).

The ability to estimate the likelihood of supertrees, allows implementing Bayesian (MCMC) approaches, which have the advantage to allow the support for the clades in a supertree to be properly estimated.

I present here the L.U.St software package; it contains the first implementation of a ML SM and allows for the first time statistical tests on supertrees. I also characterized the first implementation of the Bayesian (MCMC) SM. Both the ML and the Bayesian (MCMC) SMs have been tested for and found to be immune to biases. The Bayesian (MCMC) SM is applied to the reanalyses of a variety of datasets (i.e. the datasets for the *Metazoa* and the *Carnivora*), and I have also recovered the first Bayesian supertree-based phylogeny of the Eubacteria and the Archaebacteria. These new SMs are discussed, with reference to other, well-known SMs like Matrix Representation with Parsimony. Both the ML and Bayesian SM offer multiple attractive advantages over current alternatives.

Chapter 1: Introduction

"The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species… The limps divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups" – Charles Darwin, 1859.

It is uncontroversial that the biologist's interest in recovering relationships of common ancestry among organisms dates back, at the very least, to the work of Darwin (see above), and the use of trees to depict evolutionary trends (not necessarily based on the Darwinian concept of common ancestry) predates the work of Darwin himself, dating back at the very least to the work of Lamarck (1809). Rightfully, phylogenetics still plays a central role in evolutionary biology. Relationships among many living organisms are still poorly understood, and the development of methods to recover such relationships and test phylogenetic hypotheses is still a central aim of theoretical biology. The goals of this PhD thesis are to develop new methodological approaches to reconstruct phylogenetic trees, to test pre-existing phylogenetic hypotheses, and to apply such methods to real data

sets. More broadly, this thesis is about bioinformatics. Biologists need informatics not only to assemble sequenced genes into full genomes but also to store and analyse the genomic data that is now readily available. Phylogenetics, the reconstruction of trees depicting the relationships among a set of objects, is a particularly important branch of bioinformatics. Firstly, phylogenies are used as part of other bioinformatic tools. For example, standard multiple sequence alignment methods exploit phylogenetic trees to decide the order in which sequences are to be added to growing alignments (Feng and Doolittle, 1987). Secondly, phylogenies are important *per se* because they allow the relationships among several objects (e.g. species or genes) to be defined, and this is prerequisite to understand several aspects of their evolutionary history. Notable examples are represented by the "comparative approach" (e.g. (Harvey and Purvis, 1991; Lamarck, 1809; Rohlf, 2001; Rihoux and Ragin, 2008), whereby phylogenetic trees are used to correct for the effect of common ancestry when correlating biological variables (van Hooff, 1972; Rowe and Arnqvist, 2002; Losos and Glor, 2003), and by application in macroecology, e.g. to understand patterns of biodiversity through time (Ruta *et al.*, 2003) and (Ruta *et al.*, 2007)). It is thus unsurprising that theoretical phylogenetics has become a vibrant area of research at the interface between informatics and evolutionary biology, and it is equally unsurprising that *Systematic Biology* (the journal that publishes original research in evolutionary biology with the highest impact factor) specialises in theoretical phylogenetics and method development. Indeed, there is a constant need for new analytical tools that can efficiently deal with the ever-increasing amount of data that are currently being generated, and for new methods that can improve on the accuracy of currently available methods.

Phylogenomics, the use of genome scale data sets in phylogenetics, has now virtually replaced standard (using single or a few genes) phylogenetic analyses (Fitzpatrick *et al.*, 2006; Pisani *et al.*, 2007; Holton and Pisani, 2010; Pisani *et al.*, 2007). Advances in molecular biology and next generation sequencing (NGS) techniques have led to an explosion in the amount of full genome data that is available for analysis. NGS methods are discussed in some detail in section 1.2.1. but for an in-depth review see Metzker (2009) and Ansorge (2009). Phylogenetic analyses based on a handful of genes are now generally considered to be of low significance, at best providing preliminary results that ought to be validated in light of phylogenomic analysis.

The consequence of the above is that there is now a growing need for sophisticated new methods that can deal with the reconstruction of phylogenies based on entire genomes (or at least based on large data sets composed of hundreds to thousands of genes) (Gordon, 1986; Baum, 1992; Ragan, 1992; Ranwez *et al.*, 2007; de Queiroz and Gatesy, 2007; Steel and Rodrigo, 2008; Smith *et al.*, 2009; Bansal *et al.*, 2010). Such techniques can be divided into two broad classes: 1) those based on a gene concatenation approach (and inspired by the total evidence approach (Kluge, 1989, 2004; Farias *et al.*, 2000) and 2) those based on some form of reconciliation of the gene trees into single, species trees. Approaches that fall into this second class are essentially grounded on the theoretical underpinnings defined by the Taxonomic Congruence approach (Farris, 1971; Mickevich, 1978; Miyamoto and Fitch, 1995). Gene concatenation approaches are generally referred to as supermatrix-based approaches. These approaches involve the generation of sequence alignments or rectangular phylogenetic matrices through the

concatenation of all the character data followed by simultaneous analyses. The supermatrix approach also known as 'combined analysis or simultaneous analysis' is a "total evidence approach" because of its direct and simultaneous use of all of the taxa included in a given study (de Queiroz and Gatesy, 2007). Taxonomic congruence approaches include a more heterogeneous set of tools including the consensus and supertree methods, and the gene-tree/species-trees approaches (Page, 1998; Liu and Pearl, 2007; Drummond and Rambaut, 2007; Kubatko *et al.*, 2009). This thesis will focus on the supertree methods and any mention of supermatrix, unless stated otherwise, will be in relation to the rectangular matrix of pseudocharacters representing nodes on trees (e.g. the Matrix Representation with Parsimony (MRP) supertree method developed independently by Baum (1992) and Ragan (1992)).

In addition, one should keep in mind that there is another important data set that bears on our understanding of evolution: Morphology. The latter is key because for some types of biological data (e.g. fossils) genomic data will never be available. Yet it is well known that fossil information is key to our understanding of evolution (Gauthier *et al.*, 1988). For example no study of the evolution of birds will ever lead to any solid conclusions if researchers were to limit their comparisons to the extant vertebrate lineage (Chiappe, 2002). Several approaches that can integrate morphological and molecular data have been developed within both of the taxonomic congruence and the total evidence frameworks. The former uses supertree approaches to integrate phylogenies derived from the analyses of fossil data. The latter, on the other hand, exploits the availability of models that can accommodate morphological characters and data partitioning (Lewis, 2001; de

Queiroz and Gatesy, 2007; Geisler *et al.*, 2011; Ronquist *et al.*, 2012b; Ronquist *et al.*, 2012a).

According to Steel and Rodrigo (2008) mathematicians involved in the development of methods for phyologenetic inference have often complained that biologists are not always sure what is it that they want when they build a phylogeny. As a biologist I think it is fair to assume that what we want is the best possible interpretation for the data that is available to them. Hence, the principal aim of this thesis is to provide a solid framework for data interpretation. In particular I shall focus on supertree approaches inspired by Taxonomic Congruence and investigate the developments and the applications of Maximum likelihood (ML) and Bayesian (MCMC) supertree methods. These new methods will be shown to represent improvements over currently available supertree methods (Bininda-Emonds, 2004a) and are implemented in the L.U.St software package.

In the second chapter of this thesis, **Implementation of Maximum Likelihood (ML) and Bayesian (MCMC) supertree methods**, I discuss briefly concepts introduced by Steel and Rodrigo (2008). These provide the foundation for the calculation of a maximum likelihood supertree given a set of input trees on partially overlapping taxa. This chapter will focus on theoretical issues and present software that has been developed as part of this project in order to implement ML and Bayesian supertree reconstruction. Topics include representation of trees as data structures that can be manipulated by computer software, the creation of a tree-class in the context of object oriented programming, and the application of this class to the development of a ML supertree software in which I implement a subtree pruning and regrafting (SPR) heuristic search strategy (Swofford *et al.*, 1990).

Chapter three, **Testing case studies**, discusses the different tests that have been performed on each of the supertree methods in order to ensure that it is fit for its purpose. The ML and Bayesian (MCMC) supertree methods are tested for both input tree size- and input tree shape- related biases. In this chapter, I also take a look at how both of these supertree methods perform when used to analyse an empirical data set for which I know the expected result a priori.

Chapter four, **Reanalyses of published data sets**, examines the performance of the ML and the Bayesian (MCMC) supertree methods when used to analyse real world data sets. Two previously analysed data sets, the metazoan dataset of Holton and Pisani (2010) and the carnivore data set of Nyakatura and Bininda-Emonds (2012) were used in this chapter. The first data set represents a phylogenomic data set (with a high level of taxonomic overlap) while the second is a more traditional data set composed of trees sampled from the literature (with a low level of taxonomic overlap).

Chapter five, **Reconstructing the Bayesian Tree of Life**, explores a key question in phylogenetics, the nature of the tree of life, and evaluates whether it will be possible to improve the current understanding of the evolution of life by using the new tools introduced in this thesis. This chapter is essentially an application of the various tools I have developed over the course of my PhD to a novel genomic-scale data set. In chapters three, four and five I compare and contrast the performance of both the ML and Bayesian supertree methods against other available and widely used supertree methods. For this, I decided to compare the ML and Bayesian methods against three well established methods, i.e.: matrix representation with parsimony (Baum, 1992; Ragan, 1992), most similar supertree

(Creevey and McInerney, 2005), and the Robinson Foulds supertree (Bansal *et al.*, 2010). The results obtained from these chapters are used to explain the advantages offered by the novel supertree methods developed and characterised in this thesis over existing supertree methods.

Chapter six, **General Discussion and Conclusions**, will address the results I have obtained, discuss the questions answered by this thesis and the new questions posed by these results, and I attempt to philosophize on the impact that this thesis will have on phylogenetics.

Chapter seven, **General conclusion**, is a short concluding chapter where I will evaluate the extent of the future work I need and want to perform to improve on what I have achieved here.

1.1 Terms and Definitions

1.1.1 Trees

Most of this thesis focuses on the development of new supertrees methods. These methods are used to construct more inclusive and larger phylogenies using the information in smaller trees (input or source trees). Accordingly I start this thesis by describing and formally introducing the concept of trees. As this is not a mathematical thesis, this introduction is mostly written from the standpoint of practising biologists. For a more mathematically in-depth definition of trees, see (Harary and Palmer, 1973; Bryant, 1997; Thorley, 2000).

Trees are acyclic, connected graphs (Harary and Palmer, 1973; Bryant, 1997). In particular, phylogenetic trees differ from standard trees because, aside from consisting of a set of nodes (vertices) that are connected by a set of branches

(edges), they further have labelled terminal nodes. Each node in a tree has a degree, representing the degree of a node is the number of branches incident to it. In a phylogenetic tree, nodes can be terminal or internal. Terminal nodes are those with a degree of 1. Internal nodes have a degree greater than 1. Phylogenetic trees can be either **rooted** or **unrooted**. An **unrooted phylogenetic tree** is a tree with no nodes (vertices) of degree of 2. This corresponds to the phylogenetic tree in Steel (1992) and Dress and Steel (1992). A **rooted phylogenetic tree** is described in the same way, except that the internal node called the **root** is distinguished by having a degree of 2.

Given any tree *T*, the leaf set or taxon set of such a tree, denoted by *L(T)*, is the ensemble of terminal nodes. However, in the case where *T* is a set of trees, *L(T)* represents the union of the leaf sets of the trees in *T.* A node *a* in a rooted tree is a descendant of a node *b* if we have to go through node *b* to get to the root from node *a.* In this situation, node *b* is considered an ancestor of node *a*. The nodes that are both adjacent to and descendant of node *a* are considered the children of *a*. The adjacent node that is also an ancestor of node *a,* is known as the parent of node *a*.

1.1.2 Tree resolution

The resolution of a tree is the amount of structure (or information) it contains. A tree is **bifurcating** or fully **resolved** if all of its internal nodes (except the root) have a degree of three (see figure 1.1c). A polytomous tree is a tree with one or more internal nodes of degree greater than three. A tree containing a single polytomous internal node (a tree with no internal branches) is known as a **bush** (see figure 1.1a).

Polytomies in phylogenetic trees can be interpreted in two different ways (Maddison, 1989). They can be either hard, indicating that more than two lineages diverged from the same speciation event (i.e. simultaneously), or soft, indicating ignorance of the true cause. Polytomies throughout this thesis will be treated as soft. This means that resolved trees are considered to be maximally informative while bushes are considered to be totally uninformative.

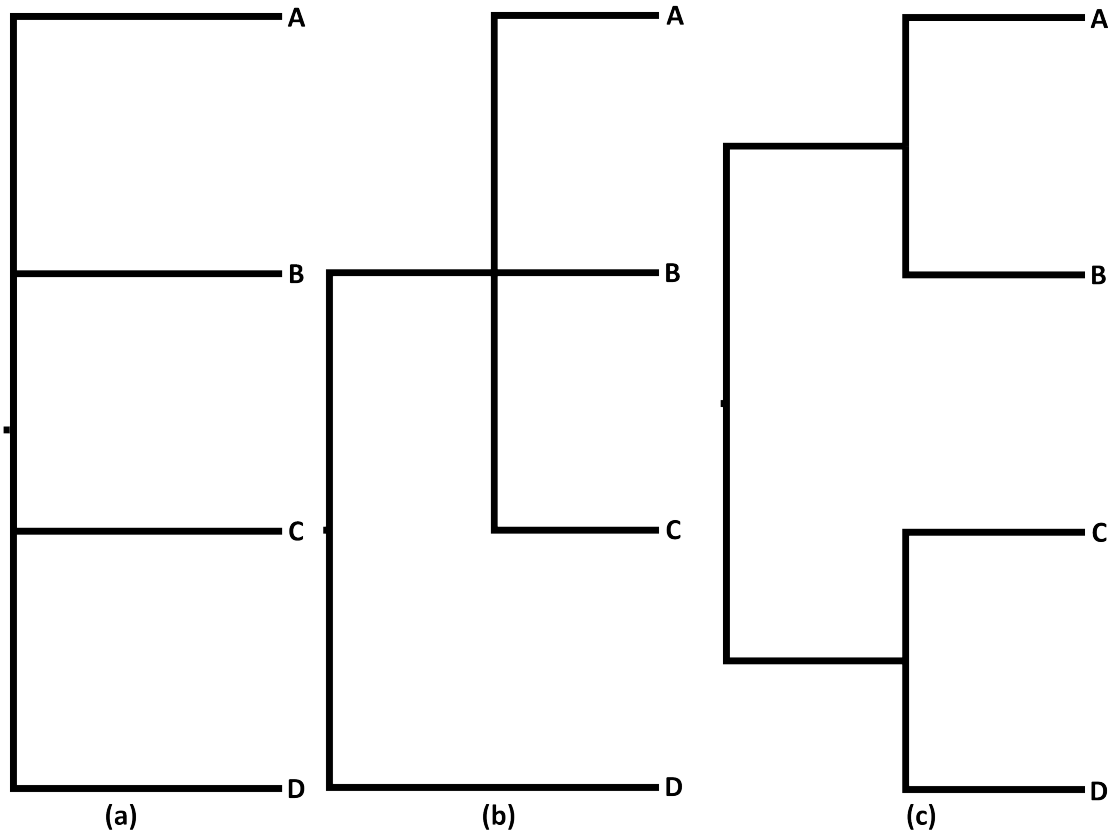


Figure 1.1: Rooted phylogenetic trees with different levels of resolution.
(a) A rooted bush, (b) A rooted partially resolved tree and (c) A rooted bifurcating tree.

1.1.3 Subtrees

Given any tree *T,* and a set of leaves *F* where $F \subseteq L(T)$, the **subtree** of *T* induced by *F*, denoted $T_{|F}$, is the minimal subgraph of *T* when only the node labels from *F* are connected, with all nodes with a degree of two suppressed (see figure 1.2) (Buneman, 1974).

l_1 is an internal node in a rooted tree *t*; by removing the branch between l_1 and its parent node we are left with two connected subgraphs. Rooting the subgraph containing l_1 at the node l_1 leaves us with the subtree of *t* rooted at l_1.

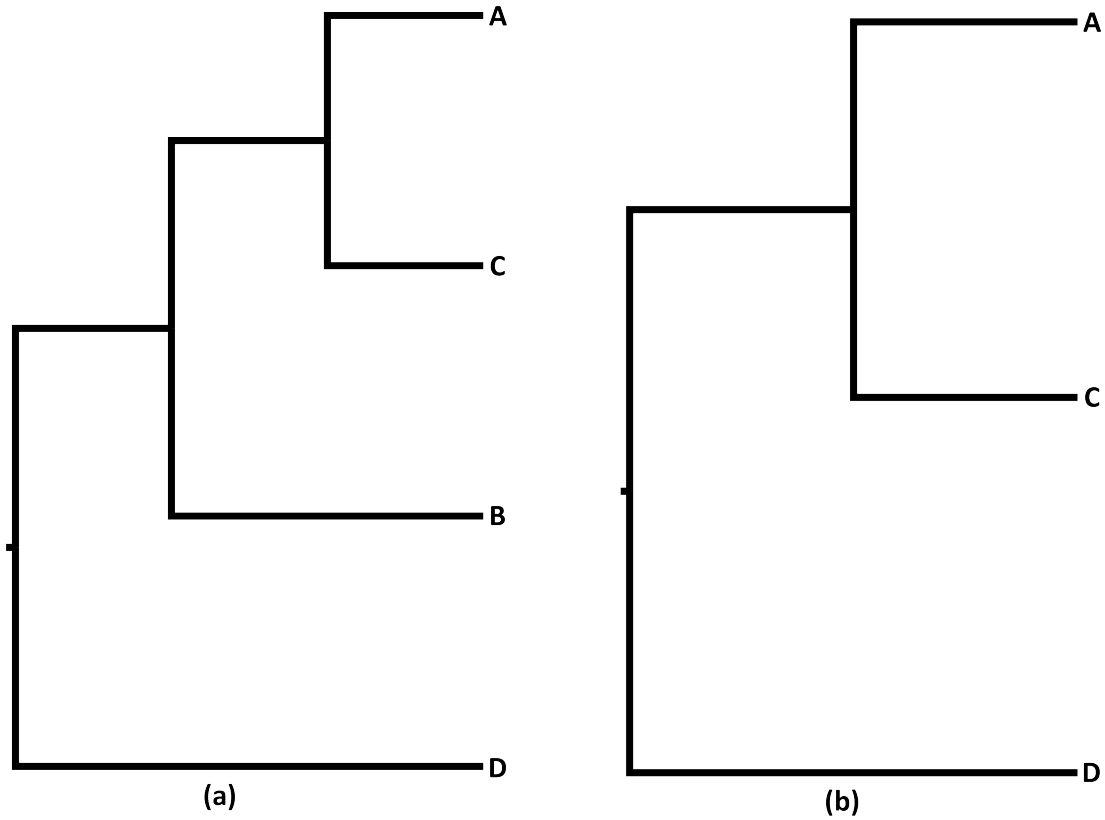


Figure 1.2: A tree and its subtree. A tree (a) and (b) a subtree of the tree in (a)

1.1.4 Splits

The set of the **splits** of a tree *T* is the set of all bipartitions corresponding to edges in *T* (see figure 1.3). If we assume *b* to be an edge in tree *T*, then by removing *b* we partition the leaves of *T* into two subsets each corresponding to one of the parts composing the bipartition defined by b. One of the parts of this bipartition will represent a monophyletic group (i.e. a group of leaves with a common ancestor). The other could be either mono or paraphyletic. The first element of the bipartition can be defined a **clade** or a **component**. The remaining partition is a clade or a component only if it is monophyletic. If it is paraphyletic, then it does not represent a clade or a component.

A trivial split is the split corresponding to an external branch of *T*. This split is characterised by one of the partitioned subsets having a cardinality of one (the cardinality of a set being the number of elements, in this case taxa, it contains). Trivial splits are phylogenetically uninformative as they are always true (irrespective of the data analysed).

Non-trivial splits correspond to splits on the internal branches of *T* and are regarded as representing cladistics information. In a rooted tree the splits are denoted the inner set and the outer set respectively to convey the direction of evolution specified in tree *T* (Gusfield, 1991; Bryant, 1997)

1.1.5 Triplets

A tree *T* is a **triplet** if its leaf set *L(T)* has cardinality $|L(T)| = 3$. For every three leaves there is only one unrooted tree and three possible rooted trees. The unrooted three-taxon tree is cladistically uninformative. A **rooted triplet** is considered to be

equivalent to the smallest possible non-trivial rooted split (see figures 1.2b, 1.3) (Thorley, 2000).

1.1.6 Nestings

If *a* and *b* are two nodes in a rooted tree, *a* is an ancestor of *b* if we have to go through *a* to get to the root of the tree from *b*. Node *a* is the last common ancestor to all nodes that are descendant of *a*. The common ancestor of a set of leaves, which is also a descendant of the set of common ancestors for that leaf set, is said to be the most recent common ancestor for that leaf set. Two groups are said to **nest** together if the last common ancestor of group 1 is an also an ancestor of the most recent common ancestor of group 2 (Adams, 1986) (figure 1.3). This implies that while a clade is always a nesting a nesting is not necessarily a clade.

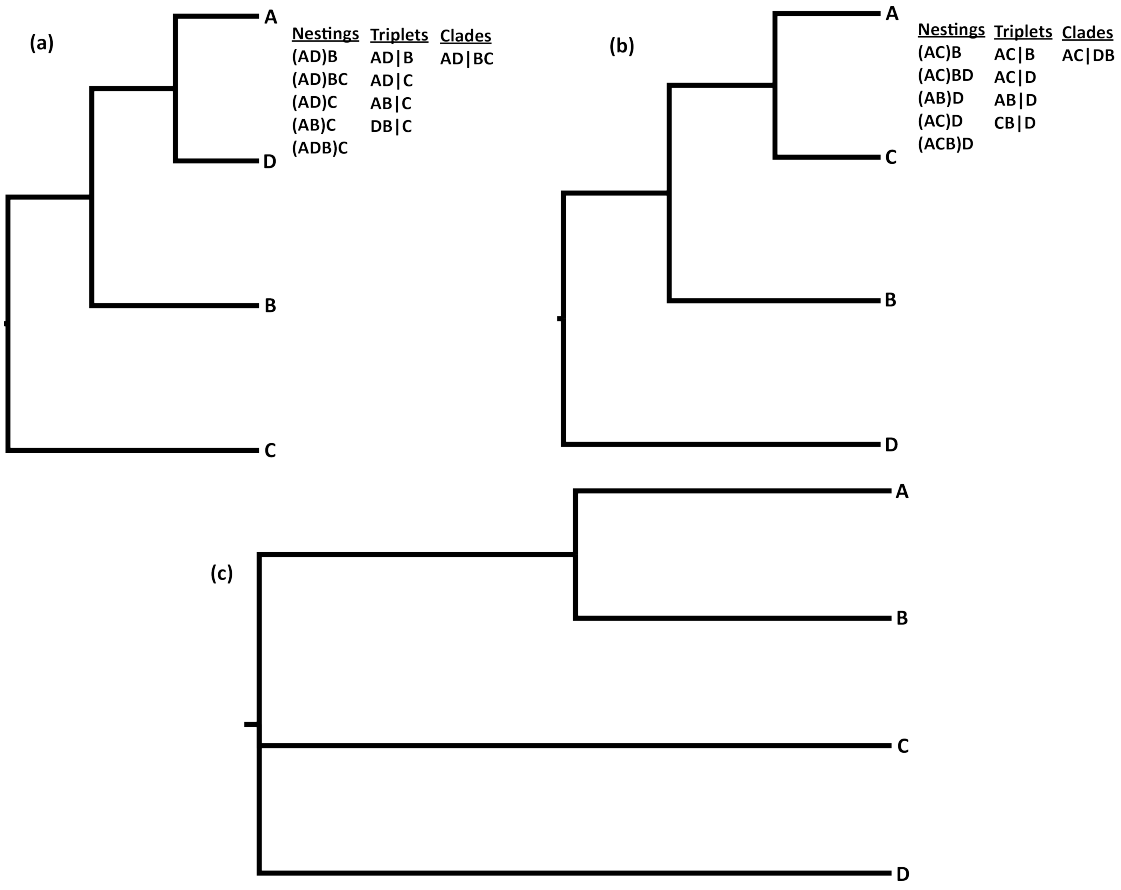


Figure 1.3: Two rooted phylogenetic trees and their Adams consensus tree.

(a) and (b) The two rooted input trees with a list of all their respective nestings, clades and triplets. (c) The strict nesting consensus tree (Adams consensus tree) of the trees in (a) and (b).

1.1.7 Quartets

A tree constructed on four leaves is known as a **quartet** (figure 1.3c is an example of a quartet). The cladistics information presented in a tree summarizes the inferred evolutionary histories of the taxa on the tree based on their phylogenetic relationships. There are three possible resolved unrooted trees that can be inferred on a quartet of leaves and each of these is regarded as having cladistic information. There are fifteen possible resolved rooted trees that can be inferred from a quartet

of leaves. Any tree *T* is made up of, and can be reconstructed from, its set of rooted triplets or quartets (Bandelt and Dress, 1986).

1.2 Building a Phylogeny

Phylogenies (aka evolutionary trees) are the basic tools that we employ to understand the evolutionary history of a group of objects (e.g. a group of species) and to analyse their relationships statistically (Felsenstein, 2004). So far phylogenies have been built for organisms using whole genomes (Snel *et al.*, 1999), ribosomal RNA (Woese, 1977), microbacterial strains (Werren *et al.*, 1995), metabolic pathways (Forst and Schulten, 2001), human languages (Pagel, 2009) to name but a few. Phylogenies can be built from morphological or genomic (DNA or protein sequences) data.

DNA is the four-letter genetic code responsible for the development and functioning of all organisms. It is in this form that genetic information is passed from one generation to the next through evolutionary time. DNA corresponding to protein coding genes is transcribed into RNA and translated into amino acid (AA) sequences. Phylogenetic analyses can be carried out at the DNA, or AA level, and analyses performed using sequences representing DNAs or proteins generally have different aims. This is because DNA and AA sequences evolve differently, in the sense that DNA tends to accumulate mutations faster than AA sequences. This is because of the degeneracy of the genetic code and the existence of silent/synonymous mutations (i.e. mutations in the DNA sequence that do not cause the AA sequence to change) (Rota-Stabelli *et al.*, 2013). As a consequence DNA data

are not ideal to identify divergent homologs and for deep-time phylogenies while AA data sets are inadequate to resolve shallow level relationships (e.g. at the species level) (Rota-Stabelli *et al.*, 2013).

Proteins (amino acid) sequences have been used extensively in the reconstruction of phylogenies ((Hashimoto *et al.*, 1994; Adachi and Hasegawa, 1995; Baldauf *et al.*, 2000; Harper *et al.*, 2005) and there is an abundance of methods available (Kishino *et al.*, 1990; Adachi and Hasegawa, 1992; Hasegawa and Fujiwara, 1993; Posada and Crandall, 1998; Castresana, 2000). All analyses performed in this thesis will use AA sequences.

1.2.1 Getting the data

The first step in building a phylogeny is getting the data (molecular or morphological characters). The major challenge in using morphological data for the reconstruction of a phylogenetic tree is deciding on the phenotypic characteristics to use as characters among the organisms in question (Swiderski *et al.*, 1998).

The continuous improvements in the field of Next Generation Sequencing methods (NGS) hides the fact that although the structure of the DNA was established in 1953 (Watson and Crick, 1953), the first DNA sequence was not acquired till more than 20 years later using techniques based on two dimensional chromatography (Summers *et al.*, 1973). However, the field of sequencing has never looked back, with the first full gene being obtained only a few years after (Fiers *et al.*, 1976). The road to genomics was paved by important discoveries such as the Maxam-Gilbert sequencing method (Maxam and Gilbert, 1977), the chain termination method of Sanger *et al.* (1977) and the whole-genome shotgun sequencing techniques of Smith

et al. (1995). The development of high throughput NGS techniques (such as the 454 pyrosequencing of Margulies *et al.* (2005), the illumina sequencing of Metzker (2009) and the ABI solid system of McKernan *et al.* (2009)), which adapted the sequencing process for running on a parallel process, means that DNA sequencing has become easier, faster, more reliable and most importantly cheaper.

However it is the development of revolutionary sequencing machines such as Ion Torrent's Personal Genome Machine (PGM) (Rothberg *et al.*, 2011), Pac Bio's Single Molecule Real Time sequencing (SMRTs) (Eid *et al.*, 2009) and illumina miSeq (Bentley *et al.*, 2008) that are really bringing sequencing into more labs than ever was thought possible.

1.2.2 BLAST and Homology

Introduced by Richard Owen (1843), the concept of homology is fundamental in modern biology (Fitch, 2000). Homology from the Greek word *Homologia* (meaning agreement) was defined at the time of its birth as "the same organ under all varieties of form and function". Homology, defined in an evolutionary context as the same structure in two species that has been inherited from a common ancestor, has become the foundation of any comparative analysis. After obtaining the data (nucleotide or AA) for the phylogenetic analyses the next step is the identification of homologous sequences.

Homology identification is complicated by homoplasy (independently derived similarity). Homologous sequences in this thesis were identified using the Basic Local Alignment Search Tool (BLAST - (Altschul *et al.*, 1990)). The BLAST approach uses sequence similarity to identify homologous sequences. Given a seed sequence

(a query sequence) and a database of potentially homologous sequences BLAST uses a significance score (also known as the Expectation-value (E-value)) to represent how likely it is that the compared sequences have the observed level of similarity (given the dimension of the analysed data set) by chance alone. The smaller the E-value of two compared sequences the higher their likelihood of being homologous. As a rule of thumb, sequences with E-values < 10E-50 are considered to be close homologs while an E-value of 10E-20 < E-value < 10E-8 would indicate distant to very distant homologs and may represent false positives (i.e. sequence similarity as a result of analogy and/or homoplasy). Any sequences with an E-value above 10E-5 are considered not to represent homologous relationships.

Homology among sequences can be divided into three types that cannot be distinguished by BLAST. The three types of molecular homology are as follows: paralogy (homology as a result of gene duplication), orthology (homology as a result of a speciation event) and xenology (homology as a product of the lateral transfer of genetic material) (Fitch, 2000). In this thesis, only orthologous sequences are combined to build species trees from gene trees.

1.2.3 Multiple sequence alignment (MSA)

The next step in the reconstruction of phylogenies using molecular data is to build a multiple sequence alignment. A MSA arranges three or more sequences (nucleotides or amino acids) into a rectangular array to refine further the hypothesis of homology among them by identifying homologous sites. To construct a MSA for a given number of sequences one should generate an n-dimensional matrix expanding the dynamic programming technique introduced originally by Needleman and

Wunsch (1970). However, such a strategy has been shown to represent a NP-complete problem (Elias, 2006).

The progressive alignment method was introduced to circumvent the unsolvable complexity of this NP-complete problem. Progressive alignment uses sequential addition and a phylogenetic tree (the decision tree) to produce the final MSA, which is a heuristic approximation of the true optimal alignment. The key step in progressive alignment is the use of a phylogenetic tree reconstruction method to define a decision tree. Essentially, a low quality tree (from a set of all the pairwise distances estimated after the generation of all independent pairwise alignments between the considered sequences) is generated. This tree is then used as a guide tree to decide the order in which the sequences will be added to the alignment, starting from the two most similar and finishing by adding the most dissimilar ones. At each step insertions and deletions are dealt with by introducing gaps that are kept fixed in the growing alignment. To deal with point mutations, MSA methods use a weighting scheme. MSA is not only useful for phylogenetic inference but also for protein structure prediction and many other tasks in sequence analysis (Edgar and Batzoglou, 2006).

Several computational algorithms and software for producing an MSA have been developed over the years. An example of a series of MSA tools based on the progressive alignment method is the Clustal programs family (Higgins and Sharp, 1988), e.g. ClustalW. T-Coffee (Notredame *et al.*, 2000) is another example of MSA software that is based on progressive alignment. T-Coffee offers an improvement in accuracy (especially for distantly related sequences) over Clustal but at the expense of speed. Muscle (Edgar, 2004), another common progressive alignment software,

is an example of an MSA tool that uses an iterative approach. Muscle improves over Clustal and T-Coffee by updating the distance measures among sequences between iterations and by using a distance method that is more accurate to assess the relatedness of two sequences when building the guide tree.

Choosing a MSA tool is not straightforward and choice is often problem dependant. In this thesis we have generally used the **PRANK** alignment software (Löytynoja and Goldman, 2008). PRANK is a phylogeny-aware MSA method (uses the evolutionary distance between sequences in the alignment process) and although it is slow in comparison to some other MSA methods it produces alignments that are likely to be more accurate than those generated by other methods.

1.2.4 Phylogenetic methods

There are several algorithms available today for phylogenetic inference. Tree reconstruction methods are evaluated based on their speed, accuracy, efficient use of data and other factors. The available tree computation methods can be divided into three categories.

1.2.4.1 Distance based methods:

Distance methods require a distance measure between pairs of sequences in a dataset to be calculated. This means that distance-based methods create a phylogeny that represents a certain distribution of distances on the set of sequences. Distances between sequences can be defined as the number of differing alignment positions, weighted differences, edit distances, Poisson corrected distances etc. See (Felsenstein, 1984).

The neighbour-joining method (Saitou and Nei, 1987) is an example of a distance-based phylogenetic method. It generates unrooted trees and assumes that distances are additive. Distance-based methods are very fast. However, they are not very accurate, and are unable to use efficiently the information in MSAs (Felsenstein, 2004), as they convert all mutations into a single value representing the distance between two sequences.

1.2.4.2 Character-based methods

Unlike distance-based methods Character state-based methods require a matrix of discrete characters as input. They use information more efficiently than distance-based methods because information is not lost during the transformation of the alignment into a distance matrix. Character-based methods include maximum parsimony (MP), maximum likelihood (ML) and Bayesian methods.

1.2.4.2.1 Maximum Parsimony (MP)

MP is a character-based method. It selects the tree that explains the observed data using the minimum number of evolutionary events (character substitutions along the branches). The task of identifying the most parsimonious tree is a NP-hard problem, and becomes more and more difficult as the number of taxa increases. This is because the tree space grows exponentially with the number of taxa in the dataset (Felsenstein, 2004). As a consequence, several heuristic methods have been developed to search the tree space and find the MP tree or trees (the MP tree is not necessarily unique). Although faster than maximum likelihood, maximum parsimony is easily swayed by systematic biases that can affect the data (such as long branch attraction). This is because this method is “naïve” in the sense that it does not

assume a model of DNA or amino acid substitution, and it considers each character-state change as representing an evolutionary event. Hence, parsimony is misled, when multiple substitutions and parallel substitutions happened in at heterogeneous rate in distantly related lineages.

1.2.4.2.2 Maximum Likelihood (ML)

Edwards and Cavalli-Sforza introduced the likelihood estimation method of Fisher (1922) into phylogenetics in 1964. Currently it is one of the most widely used methods for phylogenetic inference. Similarly to parsimony, the ML method uses character data rather than distances. However, instead of a parsimony criterion being used to select a tree, the tree selected by the ML method is the tree maximising the likelihood of having generated the observed alignment (i.e., the analysed data).

Essentially the likelihood of a hypothesis is the probability of the data given the model. The data here is what we have observed and does not change. For a phylogeneticist, this is usually a multiple DNA or amino acid (AA) sequence alignment, or a morphological data set.

What the model represents, however, is more ambiguous. In phylogenetics, this is composed of a tree with branch lengths (representing sequence relatedness) and the mechanism of molecular change (Felsenstein, 2004). The mechanism of molecular change (loosely referred to as the model) itself is how we think molecular sequences change over time. Phylogenetic models for molecular data are generally composed of two parts, the nucleotide or AA composition frequencies and the substitution rates (Foster, 2001).

This means that we evaluate the likelihood of the tree under the composition and substitution rates. The composition is the frequency of the four nucleotides or 20 AAs while the substitution rate is usually a matrix showing the probability of one nucleotide changing to another nucleotide or one AA changing to another AA. Substitution models exist for both DNA and proteins. However, their level of complexity varies with the Jukes and Cantor (1969) model being the simplest and the General Time Reversible (GTR- (Tavaré, 1986)) model the most complex among the site-homogenous models. The Jukes and Cantor model assumes one substitution rate only for all possible character state substitutions and equal frequencies for all character states. The GTR model assumes that each individual character state substitution can have its own rate and that the frequencies at which the characters appear in the data are character specific. Generally, amino acid substitution models are empirical GTR models (i.e. they represent a GTR matrix frozen in time) (see (Jones *et al.*, 1992; Jones *et al.*, 1994a; Jones *et al.*, 1994b; Koshi and Goldstein, 1995; Koshi and Goldstein, 1997; Koshi and Goldstein, 1998)). However it is now possible to also derive mechanistic (i.e. directly inferred from the data) GTR models also for amino acid data sets. Site heterogeneous models that are even more complex than the GTR model (e.g. the CAT based models of Lartillot and Philippe (2004)) also exist but are neither used nor discussed in this thesis.

The probability of any result (the data) can be estimated given a model under which we expect the result to be generated e.g. if in a coin tossing experiment 5 heads are obtained in 10 trials, we can calculate the probability of this result if we know that the coin is fair. The likelihood of a hypothesis is proportional to the probability of observing the data under that hypothesis (the constant of

proportionality being arbitrary – (Edwards, 1984)), and the likelihood ratio test (Equation 1) is a test that can be used to evaluate between two alternative hypotheses and decide which one fits the data better.

Equation 1: Likelihood ratio theorem

$$\boldsymbol{L\left(\frac{H1}{H2}\right) = \frac{k * \ Prob\ (D|H1)}{k * \ Prob\ (D|H2)}}$$

Equation (1) states that to calculate the likelihood ratio of the two hypotheses, their likelihoods have to be divided. As mentioned above, the likelihood of a hypothesis is proportional the probability of the data (given the hypothesis – with the constant of proportionality being arbitrary). Hence, from a practical perspective, it is generally assumed that the arbitrary constants of proportionality cancel out, and the likelihood ratio test is simply calculated as the probability of observing the data under the first hypothesis divided by the probability of observing the data under the second hypothesis (Equation 2).

Equation 2: Probability theorem

$$L\left(\frac{H1}{H2}\right) = \frac{Prob\ (D|H1)}{Prob\ (D|H2)}$$

The likelihood ratio test can be used to compare only two hypotheses at a time. However, given an alignment (the observed data - D) and a set of possible hypotheses (H_1, H_2, … H_n), the likelihood ratio test can be used to obtain a global ranking of the hypotheses. The ranking will be relative, as one of the hypotheses has to be taken as a reference, and is used as the fixed denominator in all likelihood ratio tests that need to be performed (one for each alternative – i.e. non-reference hypothesis). To make the ranking global, it is sufficient to select as the reference hypothesis the one under which the probability of observing the data is the maximal possible (i.e. P=1). Note that this hypothesis does not need to be known or exist (in the case of phylogenetics, the hypothesis with the probability of P=1 is the true tree that generated that generated the data). Accordingly, the global ranking of all available hypotheses (against the best –unknown– possible one), is simply obtained by dividing the probability of the data (under each hypothesis) by 1. That is, by calculating the probability of observing the data under each hypothesis and ranking the hypotheses according to these probabilities (see Equation 3).

Equation 3: Hypothesis ranking theorem

$$L\left(\frac{Hi}{Href}\right) = \frac{Prob\,(D|Hi)}{1}$$

Therefore, from a practical point of view, the likelihood of a phylogenetic tree equals the probability of observing the data under that tree.

1.2.4.2.3 Bayesian inference

Introduced to phylogenetic inference after a long stint in statistics, Bayesian inference is a relative of the ML inference method (Mau *et al.*, 1999). Bayesian inference of phylogeny is based on the estimation of the posterior probability of a hypothesis (a tree) given the alignment (the observed data) and a prior distribution over all possible hypotheses (Yang and Rannala, 1997). The major difference between the Bayesian and the ML approach is that the Bayesian approach uses (or should use) an informative prior distribution over all possible hypotheses (Felsenstein, 2004). The prior probability of a phylogeny, representing our beliefs on how likely particular parameter values are before the data have been observed, is combined with the probability of the data given the tree (i.e. its likelihood – Equation 3). The posterior probability of a tree, the probability that a tree is "true" (given a prior probability distribution), is calculated using the Bayes's theorem (Equation 4), which is used to estimate the relationship between prior and posterior probabilities.

Equation 4: Bayes theorem

$$P(H|D) = \frac{P(D|H)P(H)}{\Sigma_n P(D|H)}$$

The Bayes theorem (Equation 4) states that the posterior probability of the hypothesis (P(H|D)) is calculated as the likelihood of the hypothesis (given the observed data - P(D|H)) multiplied by the prior probability of the hypothesis (P(H)) and dividing this value by the likelihood of all the hypotheses (trees). Regardless of the general validity of the Bayes theorem, many statisticians disagree with the application of Bayesian methods to situations where there are an infinite number of alternative hypotheses, because in such cases proper prior distributions for the hypotheses cannot be defined. However, because the number of hypotheses in phylogenetics is always finite (i.e. the number of trees on *n*. taxa), it is always possible to use proper distributions for sets of trees. For example, one could simply use an uninformative prior that assigns a probability equal to 1/Bn (where Bn is equal to the number of binary trees on n. taxa) to each of the possible trees. Note that this is exactly what software like MrBayes (Ronquist *et al.*, 2012b) and Phylobayes (Lartillot *et al.*, 2009) do. Consequently the use of Bayesian statistics in phylogenetics is uncontroversial and Bayesian statistics has became a powerful tool for addressing many long-standing phylogenetic questions (Huelsenbeck and Ronquist, 2001).

An interesting aspect of modern Bayesian phylogenetics is that it is not strictly speaking based on the application of Equation 4. This is because analytically calculating the denominator of this equation is all but impossible for data sets with more than ~10 taxa (Yang and Rannala, 1997). However, the use of Markov chain Monte Carlo (MCMC) techniques coupled with the introduction of the Metropolis-Hasting algorithm (Metropolis *et al.*, 1953; Hastings, 1970) , algorithms that enable sampling from the posterior distribution, has revolutionized Bayesian inference by allowing the calculation of the denominator of Equation (4) to be avoided. Accordingly, all modern Bayesian phylogenetic approaches are based on the MCMC approach, and they have allowed the use of complex models on large data sets well above the limits of previous studies (Mau *et al.*, 1999).

1.3 Test of two trees

The test of two trees also known as paired site test allows two trees to be tested to access which 1 of the two fits the data better. Tests of two trees are based on the premises that a statistical test can be performed on the mean of the differences in the likelihood support of two trees at each of the sites of the alignment from which they have been derived, if we assume that evolution at each site in an alignment is independent. Alan Templeton (1983), developed the first test of two trees but this test proved too complex. However, a simplified version (a Winning site test) was developed by Allan Wilson (Prager and Wilson, 1988).

The Winning site test (also used in (Felsenstein, 1985b)) uses a binomial distribution to test if the fraction of the number of sites for which tree A fits better

than tree B (represented by a +) versus the number of sites for which tree B fits better than tree A (represented by a -) is significantly different from 50% of the sites at which the two trees have different fit. Several methods of calculating the test of two trees have been proposed such as the z test (Felsenstein and Kishino, 1993), the t test (Swofford *et al.*, 1996), the RELL test (Kishino *et al.*, 1990) etc. The above tests are influenced by the level of positive or negative signals provided by a small sample of the sites unlike the Winning sites test, which gives equal voting to every sites in the trees (Felsenstein, 2004).

The Kishino Hesegawa (KH) test (Kishino and Hasegawa, 1989) is a test of two trees that can be used to test the statistical significance of tree topologies. It was introduced as an appropriate test for maximum likelihood trees but it was noted by Goldman *et al.* (2000) that the KH test should only be applied to *a priori* selected trees that are inferred independently of the observed data. When used to create a confidence set from a set of trees that include the maximum likelihood tree, the KH test was noted to show a selection bias (Goldman *et al.*, 2000). This selection bias in the KH test led to the introduction of the Shimodaira and Hesegawa (SH) test (Shimodaira and Hasegawa, 1999). The SH test is based on multiple comparisons and it should automatically account for the selection bias in the KH test. However, the SH test was noted to suffer from a conservative bias that resulted in less trees being rejected as the number of trees to compare increased (Strimmer and Rambaut, 2002).

The approximately unbiased (AU) test (Shimodaira, 2002) is a test of two trees that is less conservative than the SH test and robust against the selection bias seen in the KH test. The AU test requires generating a number of bootstrap

replicates of different sample sizes to the original data and calculating the number of times the topology of the tree being tested is supported by the replicates. This provides bootstrap values that can be used to rank the trees being compared.

1.4 Tree merging and summarisation

Phylogenetic trees can be merged in one of two ways.

1.4.1 Consensus of trees

The aim of this thesis is the development of supertree methods. We can therefore not proceed without a short introduction to the concept of consensus.

A consensus phylogeny shows the agreement among a set of phylogenies on the same taxon set (Wilkinson, 1994). Techniques that have been developed to build consensus phylogenies are called consensus methods. There are several types of consensus methods from the Adams consensus method of Adams (Adams, 1972) to the strict consensus method of Sokal and Rohlf (Sokal and Rohlf, 1981). The inclusion of methods like the majority rule consensus method of Margush and McMorris (1981), in the group of consensus methods has courted some controversy in the literature. Nixon and Carpenter (1996), stated that the goal of a consensus method is to summarize the agreement in the phylogenetic relationships displayed by a set of phylogenies, hence only the phylogenies inferred by the strict consensus methods fulfils this goal. They suggested all other methods such as the majority rule consensus method be labelled compromise consensus methods.

Wilkinson and Thorley (2001), rightfully challenged the position taken by Nixon and Carpenter (1996), by calling such a restriction unreasonable and unhelpful

as one could make a case for the usefulness of being able to represent not just the clades but other cladistic relationships on a set of trees and their varying levels of agreement (see also (Kitching *et al.*, 1998)). Consensus methods are used in several contexts and the usefulness of any method is context dependant (Akanni *et al.*, In Prep. ; Omland *et al.*, 1999; Pisani *et al.*, 2007)).

1.4.1.1 Strict consensus methods

The strict consensus methods depict all the relationships that are unambiguously supported by a set of input trees. In other words, if all the trees in the set support a specific relationship, then that relationship is represented in the consensus tree. Relationships for which the input trees disagree are represented as unresolved polytomies. As a result, strict consensus methods tend to return less resolved trees and are more insensitive to topological differences among the input trees in comparison to more lenient consensus methods. However, there is no ambiguity in the interpretation of the relationships that they display (Swofford, 1991; Wilkinson, 1994; Adams, 1986).

The **strict component consensus method** (SCC) is the most widely implemented strict consensus method (figure 1.4, see also Swofford (2003). The SCC tree is normally used to represent sets of equally optimal phylogenies.

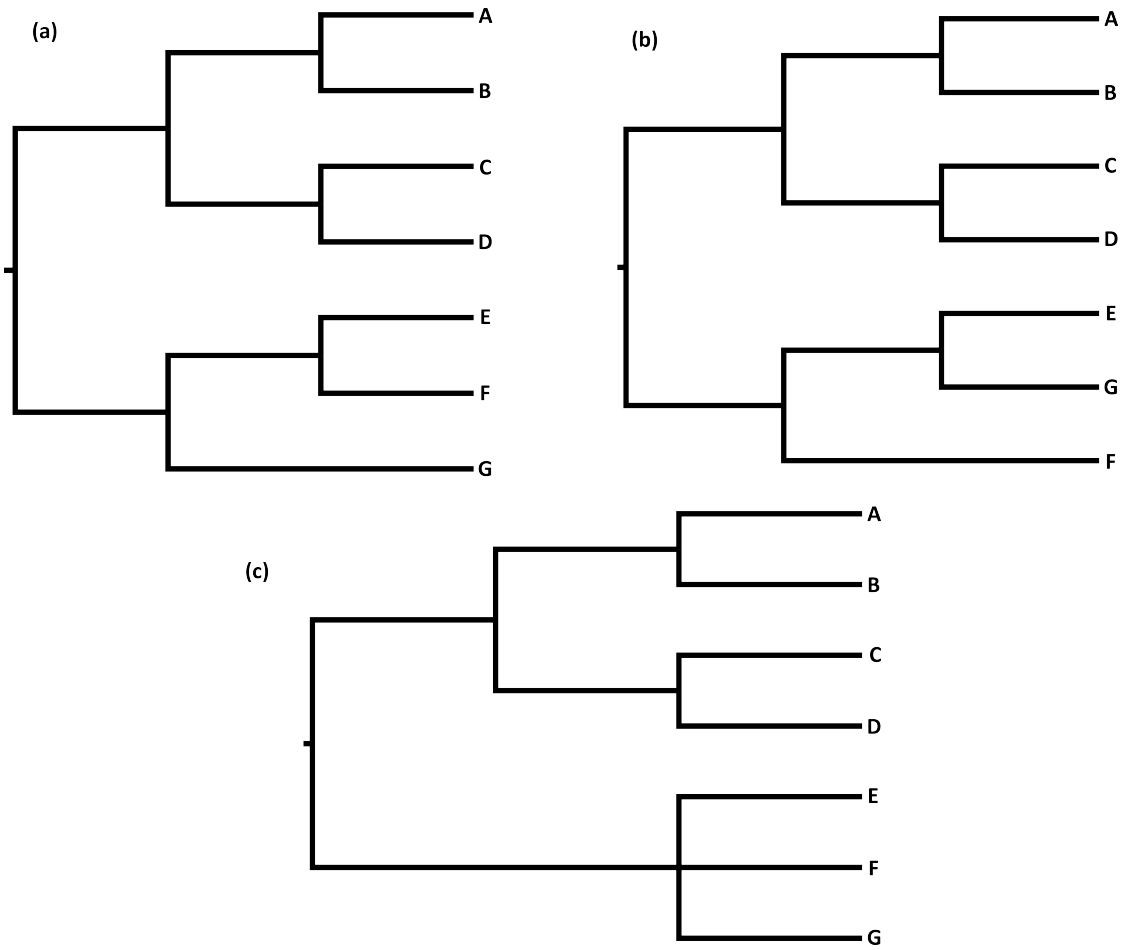


Figure 1.4: Two input trees and their strict component consensus tree**.**
(a) and (b) Two input trees. (c) The strict component consensus tree of (a) and (b).

Adams (1972) described the first consensus method but did not characterise it until 1986 (Adams, 1986). The **Adams consensus method,** as it is now known, should technically have been named the "Strict Nestings Consensus" method (SNC), due to the fact that it represents all the nestings that are common to a set of trees. Due to the fact that nestings and components represent different types of relationships on the tree, the SNC and the SCC method can often return very different results (figure 1.3).

The Adams consensus always contains all the internal branches present in the strict component consensus method, and it can be more resolved than the strict

component consensus. However, the Adams consensus tree is ambiguous (from a cladistics point of view - because nestings are not always clades) and must be interpreted differently from the strict component consensus tree (Adams, 1986). Unlike the strict component consensus trees, Adams consensus trees are more topologically sensitive to shared structure in input trees. The Adams consensus methods has been accused of producing trees that may include clades not present in any of the input trees (Sokal and Rohlf, 1981). However, this is inaccurate because nodes in an Adams consensus tree do not represent clades but nestings, hence authors claiming the Adams consensus method generates clades not present in the input set of trees (e.g. Sokal and Rohlf (1981)), are simply misinterpreting the method of Adams. Finally, it is important to recall that the Adams tree exists only for rooted trees.

1.4.1.2 Majority-rule consensus method

The Majority-rule (MJ-rule) consensus method, like the strict consensus techniques, can be used to summarize the agreement of relationship patterns in a set of trees. The most used version of this method focuses on the full splits (clades or components) in the set of trees that we want to summarize.

The MJ-rule component consensus (MJCC) tree includes all and only those clades found in the majority (typically above 50%) of the input set of trees. Other clades that induce conflict among the set of input trees are presented as unresolved polytomies. The MJCC method is often used to summarize trees in a bootstrapping framework ((Felsenstein, 1985a); (Wilkinson, 1996)), jackknifing (Farris *et al.*, 1996), with quartet puzzling (Strimmer and Von Haeseler, 1996) and when calculating

posterior probabilities in Bayesian phylogenetics (Huelsenbeck and Ronquist, 2001). In comparison to the strict component consensus tree and the Adams consensus tree the MJCC tree tends to be more resolved. However, extra resolution in the context of equally optimal trees means that the consensus tree includes relationships that are not supported by all the best interpretations of the data. For this reason, the MJ-rule consensus methods can be considered ambiguous. However, when used to summarise proportions and represent support for clades that are present in the trees, it is an excellent method. It should also be noted that the MJ-rule consensus method could be applied with a variety of thresholds (e.g. 50%, 60%, and 90%). Seen in this way, the strict consensus can be thought of as a special case of an MJ-rule tree (it is the 100% MJ-rule consensus method). Consequentially, this means that the MJ-rule tree and the strict consensus tree for two input trees are the same.

1.4.2 Supertrees

According to Wilkinson *et al.* (2004), a supertree can be defined as a phylogenetic tree that synthesizes, amalgamates or represents the evolutionary relationships displayed by a set of input trees on partially overlapping taxon sets. This definition is not different in essence from that of Semple and Steel (2000) that a supertree is any method of analysis that can amalgamate partially overlapping input trees. Consensus methods represent special cases applicable to the condition that all input trees are on fully overlapping taxa. Hence, supertree methods are a generalisation of the standard consensus methods. The supertree approach to phylogenetic

reconstruction involves the separate analyses of datasets (e.g. generation of gene trees/dataset specific trees) and their subsequent integration into a supertree (Steel, 1992; Sanderson *et al.*, 1998; Steel and Böcker, 2000; Pisani and Wilkinson, 2002; Ren *et al.*, 2009). This means that supertree methods are able to combine the phylogenetic information in a set of trees that have been inferred from all types of data (morphological data included) and using different phylogenetic methods, to reconstruct larger and more inclusive species phylogenies. The supertree approach will be discussed in more detail in chapter 2.

Chapter 2: Implementation of a Maximum Likelihood and Bayesian (MCMC) supertree method

2.1 Overview

This chapter outlays the steps towards the implementation of the Maximum Likelihood (ML) supertree method and the Bayesian (MCMC) method. As stated in section 1.2, phylogenetic trees can be constructed from either morphological data or genomic data, using an abundance of phylogenetic methods (Huelsenbeck *et al.*, 1996). The supertree approach involves building phylogenetic trees from collection of other (generally smaller and partially overlapping) trees. These can be trees that have been collected from the literature (Ruta, 2003; Ruta *et al.*, 2007; Nyakatura and Bininda-Emonds, 2012), or trees that have been derived from a series of independent data sets (e.g. a collection of alignments from a set of gene families (Creevey *et al.*, 2004; Pisani *et al.*, 2007). Both supertree methods that will be discussed in this chapter are "liberal supertrees". As mentioned in section 1.4.2 a liberal supertree method is a supertree method capable of resolving conflicts among the input trees. In the next section, I will briefly discuss some of the desired properties that a liberal supertree method should have in relation to three of the more familiar supertree methods currently available.

2.1.1 Supertree methods and their properties

In the 27 years since their introduction to classification studies (Gordon, 1986), supertrees have undergone substantial developments in terms of methods and applications (i.e. (Purvis, 1995b; Sanderson *et al.*, 1998; Jones *et al.*, 2003; Ranwez *et al.*, 2007)). The literature on new supertree methods and their variants is growing at a pace of more than 10 publications per year, and I will not achieve much by going through every supertree method available (for a more in-depth review see (Wilkinson *et al.*, 2005a; Wilkinson *et al.*, 2007)). In this section and in the rest of the thesis, I shall focus on three of the most commonly used supertree methods. These methods will be used as the standards against which to compare the new implementations for ML and Bayesian supertrees.

The first of these standard methods is the matrix representation with parsimony (MRP) supertree method. MRP was independently developed by Baum (1992) and Ragan (1992). The Standard MRP method uses additive binary coding to represent the internal nodes in input trees as pseudo-characters or elements in a matrix. The separate matrix representations (one for each tree) can then be combined to create a "supermatrix" that is then analysed using Fitch (reversible) parsimony. MRP is the supertree method of choice among researchers, and has been used to reconstruct some of the biggest, most challenging and well resolved phylogenies in the literature (Purvis, 1995b; Pisani *et al.*, 2007; Holton and Pisani, 2010; Flynn *et al.*, 2005; Daubin *et al.*, 2001). This apparent preference of MRP has courted much controversy in the literature, largely due to the fact that MRP fails to explain what the most parsimonious interpretation of the pseudo-character change mean. Whereas some investigators such as Sanderson *et al.* (1998) and Bininda-

Emonds and Bryant (1998) have claimed that the popularity of MRP is due to its potential to infer well-resolved and inclusive trees efficiently, I maintain that it is due to its ease of implementation and the familiarity of researchers with parsimony, coupled with the availability of excellent (i.e. fast) parsimony software such as PAUP (Swofford, 2003) and TNT (Goloboff *et al.*, 2008).

The second supertree method considered is the most similar supertree method (MSS) (Creevey and McInerney, 2005). The MSS supertree method compares the number of nodes separating each pair of taxa on each input tree against the number of nodes separating the same pair of taxa in the proposed supertree after it has been pruned to have the same taxon set as the considered input trees. This means that MSS uses a measure that is more reflective of the topological differences (irrespective of branch lengths) between the input tree and the pruned super tree. The MSS supertree is the tree minimizing the topological differences among the input trees

The third considered method is the Robinson-Foulds supertree method (RF) (Bansal *et al.*, 2010). Similar to the MSS supertree method, the method also prunes the proposed supertree to have the same taxon set as each of the input trees in the source data set, but the selected supertree is the one that minimizes the sum of the Robinson-Foulds distances (calculated as the sum of splits in tree A that are not in tree B and vice versa) between the proposed supertree and the input trees. It is important to note here that as I have just defined it, the RF supertree method is equivalent to Cotton and Wilkinson's Majority Rule (-) Supertree method (Cotton and Wilkinson, 2007), which returns trees with the property of being median trees to the input tree set. Indeed, this is true even though Bansal *et al.* (2010) did not

pinpoint the link between their method and the Majority Rule (-) supertree method. However, it needs to be stressed that the RF method is not necessarily finding the same solution of the Majority Rule (-) method, because it implements a heuristic strategy that is not guaranteed to find the optimal RF tree(s), which would be the Majority Rule (-) trees. In this thesis the RF supertree method has been used as an approximation of the Majority Rule (-) method, because when I started my investigation no implementation of the Majority Rule (-) method was available. Recently, such a software has been released (Kupczok, 2011b), but as most of my analyses had already been completed using the RF method, I decided to include the results obtained using RF in my thesis.

With all of these supertree methods and more readily available (including the majority rule (-) supertree method now implemented by Kupczok (2011b)), do we really need to develop more methods and software? To answer this question we first need to have a look at some of the desired properties that a liberal supertree method is expected to have. Wilkinson *et al.* (2004) proposed that the choice of a supertree method should be supported by a comparison of its accuracy in relation to other available methods. Since the ability of different liberal supertree methods to infer accurate phylogenies critically depends on their properties (i.e. how they resolve conflicts) and on the properties of data (i.e. how conflicts in the data relates to the properties of the methods) alternative supertree methods cannot be readily judged based on the results of simulations that can be easily swayed. As a consequence, Wilkinson *et al.* (2004) went on to discuss a number of properties that they think should directly bare on the accuracy of any liberal supertree method. Three key properties are: sizelessness, shapelessness and independence. A

supertree method that violates the sizeless property will tend to favour relationships in bigger trees over relationships in smaller ones when dealing with a conflict; for example, MRP has been shown to be biased towards relationships in bigger trees (Purvis, 1995a). Further to that, a liberal supertree method should not be biased towards relationships in asymmetric input trees over relationships in symmetric input trees and vice versa (i.e. it should be shapeless). Wilkinson *et al.* (2005a), investigated the shapeless property of 14 supertree methods, including MRP. Their dataset is reanalysed in chapter 3. Finally, independence (see also (Bryant, 1997)) is the property whereby extra topological information from leaves that have been pruned from the input dataset should have no bearing on the topology inferred for the remaining leaves. The properties listed above and the other (perhaps less important) properties listed by Wilkinson *et al.* (2004), are not well understood for many of the currently available supertree methods, and only recently have some studies addressed them (Wilkinson *et al.*, 2005a; Wilkinson *et al.*, 2005b). Lack of knowledge of the specific properties of liberal supertree methods, coupled with the *ad hoc* nature of most liberal supertree methods, means that any attempt to interpret the relationships that they infer can be inherently misleading and this is a major reason for the development and implementation of the ML and Bayesian supertree methods. This thesis maintains that supertree methods should have the properties highlighted in Wilkinson *et al.* (2004) . This is currently not the case with MRP which, unlike the Adams consensus method (see section 1.4.1.1) and as a consequence of its inability to meet the criteria pinpointed above, can generate truly unsupported groups (Wilkinson *et al.*, 2005b).

2.1.2 Estimating support for supertree clades

To date, the task of estimating clade support in a supertree has largely been a tedious and unsuccessful one. Creevey *et al.* (2004), suggested bootstrapping input trees (see also (Moore *et al.*, 2006)), as the bootstrap would be a natural and obvious way to measure support in supertrees. However, for the bootstrap to be applicable, the input trees must have a high level of species overlap. If a taxon is represented in, say, one out of one hundred trees, it is likely that the input tree bootstrapping procedure will produce a non-plenary (i.e. missing a taxon) pseudo-replicate (bootstrapped) dataset. In such a case, the bootstrap procedure has to be interrupted for two equally important reasons. (1) To summarise non-plenary supertrees (that are partially overlapping), one should use the majority rule supertree method. However, the majority rule supertree method (Cotton and Wilkinson, 2007), being based on tree-to-tree distances, can represent the topological relationships in the set of summarised trees but not the proportion of times that each clade appear. Hence, it is currently impossible to display support values for a set of non-plenary bootstrap trees – see Wilkinson *et al.* (2005b), for a treatment of the problem of the support that non-plenary trees can provide to the clades in a supertree. Alternatively (2) one could select only the plenary pseudo-replicates (i.e. subsample from the bootstrap generated data sets). In so doing, one would generate only plenary trees that could be summarised using the standard majority rule consensus method. Unfortunately, as the subsampling will not be random (as the selected data sets will be identified based on a specific property – i.e. they are plenary), this second approach will violate the key assumption of any bootstrap analysis (that the resampled data set are independent). As a

consequence, input tree bootstrapping has been a viable choice only for genomic applications of supertree analyses (e.g. (Creevey *et al.*, 2004; Pisani *et al.*, 2007; Holton and Pisani, 2010)), where the number of input trees is generally very high (in the thousands) and species overlap is correspondingly high.

To circumvent problems of bootstrap inapplicability, Bininda-Emonds (2003) introduced the "Quality" support (Qs) index. This index should link the frequency with which a clade in a supertree agree or disagree with the clades in the input trees. However, Wilkinson *et al.* (2005b) showed Bininda-Emonds' Qs index to be flawed and introduced an alternative called the "V" index, which unlike the Qs index is a valid approximation of supertree clade support. Price *et al.* (2005) introduced a modification of Bininda-Emonds' Qs index, but it is unclear whether Price *et al.* (2005) correction is valid (see (Baker *et al.*, 2009)), leaving Wilkinson *et al.*'s "V" as the only viable (of confirmed validity) alternative to supertree bootstrapping. However, "V" is a supertree-specific measure, it is difficult to interpret with reference to standard support measures like the bootstrap, and this reduces its utility. The general inability to measure support for clades in a supertree was the major drive underlying my interest in developing a Bayesian supertree method, as this would enable the user to estimate posterior probabilities for the clades irrespective of taxonomic overlap.

2.2 Maximum LIkelihood (ML) supertree method

It is surprising that it has taken this long for supertrees to be brought into the maximum likelihood framework. This is particularly so because since 1922, ML has grown to be one of the most employed statistical inference methods, and in the last decades it has become the most commonly used method for phylogenetic inference. The principle of ML works by estimating the value (for a given model's parameter) that makes the observed data most probable. It was recently proposed by Steel and Rodrigo (2008) that a ML supertree can be estimated, given a data set of partially overlapping trees, if the Robinson-Foulds metric is used to calculate the distances between the proposed supertrees and the input trees, and an exponential distribution is used to model the topological discordance between the input trees. That is, it is assumed that the partially overlapping input trees differ from each other because of errors, and a supertree is sought that summarises the input trees while at the same time explaining their differences as the consequence of errors in tree reconstruction. Hence, the quantity to be minimized in order to maximise likelihood is therefore the errors in the source trees.

The results of Steel and Rodrigo (2008) are summarised by Equation 5, and are based on an idealisation of the input trees being considered a sample of reconstructed subtrees extracted from an unknown true supertree. However, the reconstructed subtrees differ from the supertree pruned subtrees due to a number of reasons such as sampling errors, incomplete lineage sorting, sequencing errors, model violations e.t.c.

Equation 5: Input Tree Likelihood theorem

$$\mathbb{P}\boldsymbol{\mathcal{T}}, \boldsymbol{\Upsilon}[\boldsymbol{\mathcal{T}}'] = \boldsymbol{\alpha}\, \boldsymbol{exp}[-\boldsymbol{\beta}\boldsymbol{d}(\boldsymbol{\mathcal{T}}', \boldsymbol{\mathcal{T}}|\boldsymbol{\Upsilon})].$$

As a consequence, the ML supertree is the tree that maximises the likelihood across all the input trees by minimising the number of induced errors in the input trees.

Equation 6: Supertree Likelihood theorem

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{\beta i}\, \boldsymbol{d}(\boldsymbol{\mathcal{T}i}, \boldsymbol{\mathcal{T}}|\boldsymbol{Xi}).$$

If we take $\mathcal{T}$ to represent an hypothesized supertree and $\mathcal{T}'$ to represent an input source tree, Equation 5 (based on the assumption that $\mathcal{T}'$ has been pruned out of $\mathcal{T}$) says that the probability that $\mathcal{T}'$ was obtained from $\mathcal{T}$ can be estimated by pruning $\mathcal{T}$ so as to have the same taxon set as $\mathcal{T}'$, which is denoted Υ. Hence the probability of observing $\mathcal{T}'$ after pruning $\mathcal{T}$ of all other taxa, taking into consideration topological errors, is equal to α (a normalising constant that ensures that the sum of the likelihoods of all the supertrees is equal to 1) multiplied by the exponent of negative β (a parameter that is free to vary in relation to both the

quantity and the quality of the data) multiplied by the Robinson-Foulds distance between $\mathcal{T}'$ and (the pruned) $\mathcal{T}$ (denoted, $\mathcal{T}|\Upsilon$). Equation 6 simply states that the supertree minimizing the sum of Equation 5 values across the input trees is the ML supertree. Calculating α depends on the shape and size of $\mathcal{T}'$. Although this calculation is possible in polynomial time, it is extremely computationally expensive because it has to be calculated for every proposed supertree. Bryant and Steel (2009) suggested that we can ignore calculating α if we use a sufficiently low or high β value as in such cases the ranking of the supertrees will not be affected. Hence, in my implementation of Steel and Rodrigo's (2008) ML supertree method, the value of α is kept constant and the β values are kept out of the ranges that are proposed to be problematic by Bryant and Steel (2009). This means that our ML supertree method is, in truth, a heuristic estimator. The ML implementation is available to the public as part of my L.U.St package discussed in detail in section 2.2.3.

2.2.1 Tree representation

In order to implement the ML supertree method and embed it in a stand-alone program, the first hurdle to be jumped was to represent and store phylogenetic trees in memory. I decided to use the Python programming language for this project due to its growing popularity among researchers and programmers (Python is now used to power YouTube, Reddit, banking systems, Google, DropBox etc.). Python's surge in popularity is due to its robustness (solid, powerful, easy to debug and maintain), flexibility, availability of supporting software and the fact that it is free.

Python is an object oriented programming language. An object in programming terms is an instance of a class (Hall and Stacey, 2009), and a class can

be viewed as a template that we can use to create our own data types. Data types here can refer to real world constructs such as a person, book, or a car. Classes enable us to give objects attributes that can be used as distinguishing factors, and develop methods (associated to the objects) that we can use to perform desired operations on the objects and their attributes.

The investigation of available python libraries (such as DendroPy, Django, Scapy, Biopython) showed that no freely available code existed that could be re-used in my software to store trees in memory. Accordingly, to represent a tree properly in memory I wrote my own node and tree classes. The node class is responsible for storing all the relevant details pertaining to a particular node. These details include the descendants of said node, its ancestors, branch length information, nodes support information and other parameters. The root node is also stored as an instance of the node class. The root is defined using an added parameter to simply indicate that a given node is also the root. I also included in the node class certain operations (methods) that can be performed on a node such as those that allow tree traversal from any particular node and re-rooting of the tree to any particular node.

The tree class creates and stores tree objects. It takes as input Newick formatted trees and generates a node object for each node along with its attributes. These are the nodes that make up a tree and they are stored and linked in the right order using Python's native "dictionary data structure" (i.e. hash tables). In the process of making the ML program stand alone I created many scripts that will enable a host of desired operations on a tree to be possible and these will be discussed in section 2.4. I also developed a heuristic search strategy to search the tree space for the ML supertree and this is discussed in the next section.

2.2.2 Searching the tree space for the elusive ML supertree

Finding the ML supertree for a data set of input trees requires searching the entire tree space on the union of their taxon set. The tree space for *n* taxa is the space of all possible trees on that number of taxa (for more detail see Billera *et al.* (2001)). Searching every single tree in a tree space is computationally expensive and inefficient; in fact, when n becomes sufficiently large this task quickly becomes intractable. To circumvent this problem, many algorithms have been designed (Felsenstein, 2004). In L.U.St I implemented four alternative heuristic search strategies based on the Subtree Pruning and Regrafting (SPR) algorithm (Swofford, 2003) (figure 2.1a and 2.1b).

An SPR move starts off with a pruning step, which involves breaking up a tree into two subtrees (T_1 and T_2) by cutting a random edge. The pruning step is followed by a regrafting step, which involves choosing another edge from (say) T_1 and reinserting T_2 at that position. I devised four heuristic search strategies using the above basic swapping algorithm for exploring tree space. The alternative search strategies differ in the level of thoroughness with which they navigate the tree space. Hence they have different speed and accuracy (with accuracy decreasing as speed increases).

Search option 1

This is the most exhaustive of the search strategies implemented in L.U.St and ultimately the slowest to run.

Step 1. A new starting supertree is proposed (usually a randomly generated supertree on the union of the taxon sets of all input trees). This tree is re-rooted at every possible re-rooting point and these rooted trees are stored in a list.

Step 2. One of the newly re-rooted supertrees is chosen (without replacement) at random from the list of rooted trees and its likelihood is estimated using equations 5 and 6 above. This tree is now stored as the *current tree* and its likelihood, is stored as the *current tree likelihood*.

Step 3. A list of all the supertrees that can be generated by one round of SPR from *current tree* is generated and stored.

Step 4. One of the SPR generated supertrees is extracted (without replacement) at random and its likelihood is estimated and compared to the likelihood of the *current tree*.

> ***Step 4i***: If the likelihood of the new tree is better than the likelihood of the *current tree*, then the new tree is stored as the new *current tree* and its likelihood as the new *current tree likelihood* (old values are expunged from memory). If this were the first iteration of a search, then the new *current tree* would now also be stored as the *overall best tree* found. If this were not the first iteration of the search, then the *current tree likelihood* would be compared to the likelihood of the *overall best tree*, i.e. to the *overall best likelihood* so far. If the likelihood of the new *current tree* is better than that of the current *overall best tree,* then the new *current tree* is stored as the new *overall best tree* found so far (and the previous *overall best tree* is expunged from memory). If the likelihoods of the *overall best tree* and of the new *current tree* are the same, then the new *current tree* is added to the list of *overall best trees* found so far. If the likelihood of the new *current tree* is lower than the *overall best likelihood,* then we return to step 1 with the *new current tree* as the new starting supertree.

Step 4ii: If the likelihood of the new tree is the same as the likelihood of the *current tree*, then the new tree is stored in a list of *trees of equal likelihood* that will be visited in later steps and the search goes back to step 4.

Step 4iii: If the likelihood of the *new tree* is lower than the likelihood of the *current tree,* then the *new tree* is discarded and we return to step 4.

Step 5. If we go through the entire SPR generated tree list without finding a tree of better likelihood than the *current tree*, then we go back to step 2.

Step 6. Once we have gone through the entire set of remaining re-rooted trees the program evaluates if the list of *trees of equal likelihood* is empty.

Step 6i: If this list is not empty, then a tree is randomly extracted (without replacement) and the program goes back to step 1, where the extracted tree is treated as a new starting supertree.

Step 6ii: If the list of *trees of equal likelihood* is empty, then this iteration is ended.

Depending on the number of iterations requested by the user, either another iteration is started (and a new random tree generated) or the trees stored in the list of overall best trees found so far is returned to the user as the ML supertree(s).

Search option 2

Search option 2 is the same as search option 1 except that it skips step 2. That is: a list of supertrees re-rooted at every possible re-rooting point is not generated. In this search option the list of *trees of equal likelihood* is also re-initialised (emptied – as in Search option 2) every time a new tree with a better likelihood is found.

Search option 3

This search options is the same as search option 1 except that it only considers trees of better likelihood. Hence, a list of *trees of equal likelihood* is not generated.

Search option 4

This is the most heuristic of all of the search option available. This option does not involve generating a list of supertrees re-rooted at every possible re-rooting point and only trees of better likelihood are considered. Hence, there is no list of *trees of equal likelihood.* Basically, Search Option 4 combines the speed up strategies of Search Option 2 and Search Option 3.

Starting tree option

A way to improve speed when searching tree space is avoiding starting from a random tree. L.U.St has two alternative starting tree options. The default approach is for L.U.St to start tree searches from random trees. Alternatively L.U.St allows the user to provide a starting tree (e.g. a supertree generated with a different method – maybe an MRP tree).

The search strategies implemented were tested for accuracy and efficiency using the *Drosophila* dataset of Cotton and Wilkinson (2007) (fig. 3.3a-e). For this data set, Cotton and Wilkinson (2007) used their Majority Rule (-) supertree method and showed that there are 79 equally likely median supertrees. Steel and Rodrigo (2008) stated that the ML and the Majority rule (-) should return the same trees we decided to use this dataset, for which the correct result is known, as the gold standard. The data set was analysed using each of the search options implemented for 1, 2, 10, 100, 500, 1000 iterations. Once a particular search option was able to find the 79 equally likely trees, the analysis was stopped and the number of the

iteration at which the 79 trees were recovered was registered. The result of this experiment can be seen in Table 2.1. From the table we can conclude that option 1 is the best in terms of accuracy and speed.

ML search strategy	1 iteration (t/h)	10 iterations (t/h)	100 iterations (t/h)	500 iterations (t/h)	1000 iterations (t/h)
1	78 / 2.5	79 / 3	✓	✓	✓
2	1 / 0.1	7 / 0.13	29 / 0.33	40 / 2.5	53 / 3
3	1 / 0.1	10 / 0.33	48 / 3	78 / 12	79 / 24
4	1 / 0.1	3 / 0.15	21 / 0.35	42 /1	48 / 2

Table 2.1: Efficiency of L.U.St's ML search strategies. This table illustrates the performance of the 4 alternative SPR based heuristic search strategies, implemented in L.U.St, when used to analyse the *Drosophila* dataset of Cotton and Wilkinson (2007). Each row represents the results of the corresponding search strategy as they are numbered in the software. Ml – maximum likelihood; t/h – the total number of trees out of the 79 total median trees found during the run / the length of time in hour(s) the search strategy took to finish the analysis. Ticks represent analyses that were not done due to the search strategy having found the complete set of median trees.

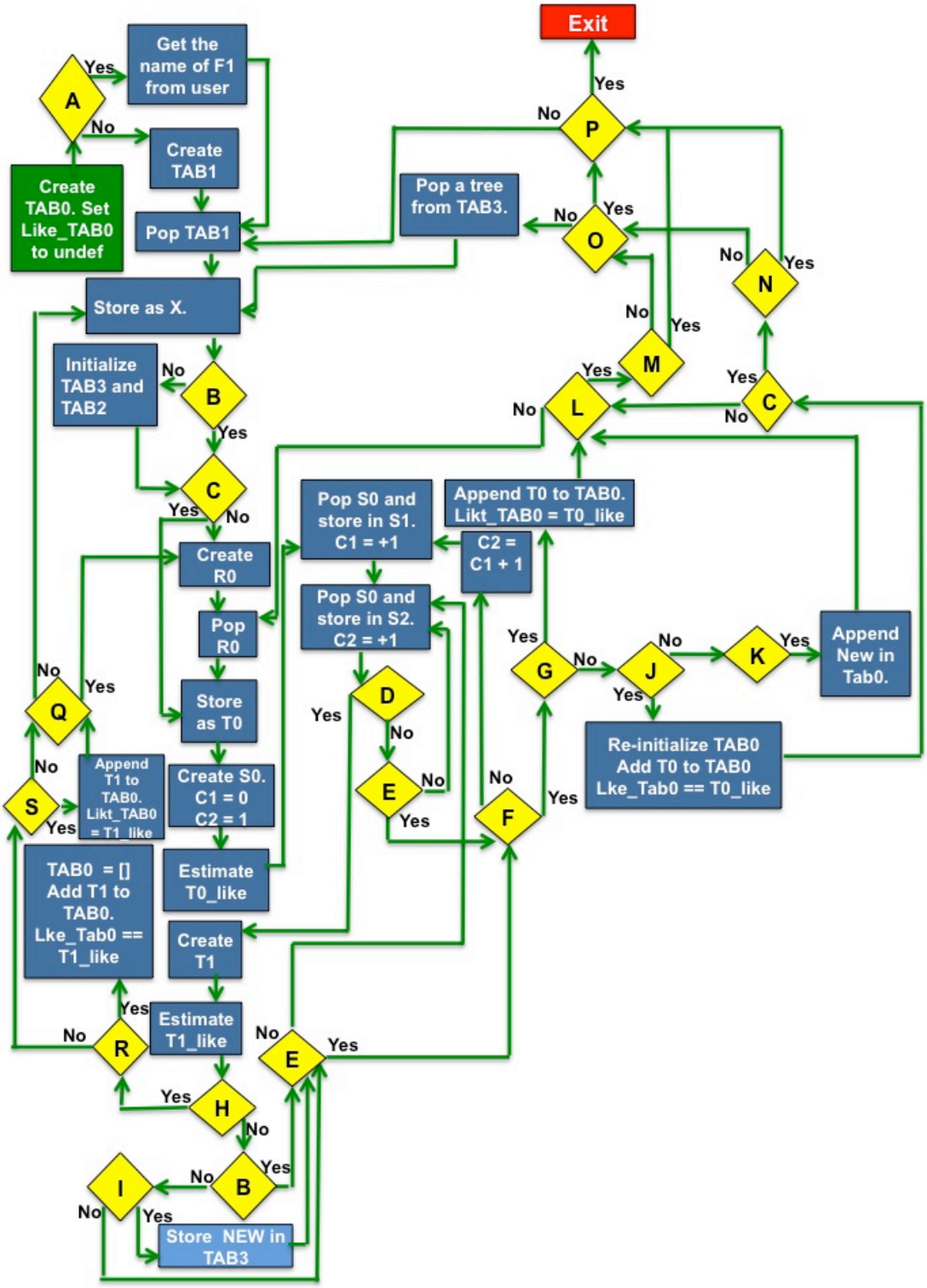


Figure 2.1a: L.U.St Maximum Likelihood supertree search strategy

Note: for a full understanding of this figure represent - see Figure 2.1b.

TAB1: A list of n randomly generated starting trees (user defines n)
F1: The file containing list of Starting supertree(s)
TAB0: Variable to store list of best trees found
Like_TAB0: Variable to store the likelihood of the trees in TAB0
TAB2: An array for the trees that we have already seen
TAB3: A dictionary to hold trees that are left to swap
R0: An array of all re-rooted trees created from all possible rooting points in a tree
T0: The re-rooted tree from R0 that we are currently working on
T0_like: The log likelihood value of T0
T1: The new tree created by one step of SPR on T0
T1_like: The log likelihood value of T1
S0: An array containing all the subtrees in a tree
S1: A variable to hold a random subtree from S0
S2: Avariable to hold a random subtree from S0
C1: A counter veriable
C2: A counter variable
Store New in Tab 0.
A: Do you have a starting tree?
B: Is this search option 3 or 4?
C: Is this search option 2 or 4
D: Is the intersection of S1 and S2 empty?
E: Is C2 equal to the cardinality of S0
F: Is C1 equal to the cardinality of S0
G: Is Like_TAB0 = undef ? ***OPTIONAL
H: Is T1_like better than T0_like?
I: Is T1_like equal to T0_like?
J: Is T0_like better than T1_like?
K: Is T0_like equal to Like_TAB0?
L: Is R0 empty?
M: Is this search option 3?
N: Is this search option 4?
O: Is TAB3 empty?
P: Is TAB1 empty?
Q: Is this search option 1?
R: Is T1_like better than Like_TAB0?
S: Is T1_like equal to Like_TAB0?

Figure 2.1b: L.U.St Maximum Likelihood supertree search strategy figure legend2.1b

2.2.3 Extending test of two trees to supertrees

The ability to estimate the likelihood of a supertree has opened up the field of supertree reconstruction to statistical hypothesis testing. Within the L.U.St package, I have included the possibility of calculating (the first time ever) tests of two trees in the supertree context. In contrast to the case of standard tests of two trees (that use site-wise likelihood values), in the case of supertrees we use input-tree specific likelihood values (that are analogous to site-wise likelihood values). Once input-tree specific likelihood values are calculated, one can use a variety of tests to compare a set of alternative supertrees for their fit to the data. L.U.St implements a winning site test (Felsenstein, 2004). In addition, it produces a CONSEL (Shimodaira and Hasegawa, 2001) compatible output file of input-tree wise likelihood values that can be fed to CONSEL to calculate the Approximately Unbiased test (Shimodaira, 2002) and other tests of two trees (i.e. Kishino Hasegawa (KH) test (Kishino and Hasegawa, 1989) and Shimodaira Hasegawa (SH) test (Shimodaira and Hasegawa, 1999)).

For L.U.St to be able to calculate tests of two trees (or output the input-tree wise likelihoods), a predefined set of supertrees (and a set of input trees that does not need to be the set of trees originally used to infer the tested alternatives) need to be provided. L.U.St will then calculate the input-tree specific likelihood score for each input tree (against every compared supertree). These values are then either directly used by L.U.St to calculate a wining site test (Felsenstein, 2004), or they are written to an output file that can be used as the input file for CONSEL. For more detail on how to run this statistical test see the provided L.U.St-manual.

2.2.4 Likelihood Utility for Supertrees (L.U.St) Package

During my PhD I wrote many scripts to perform several different tasks from extracting taxa from a dataset of trees to calculating the Approximately Unbiased (AU) test combining L.U.St and CONSEL. I believe that the availability of these scripts will be of great help to other researchers. The L.U.St package developed from this experience, it contains a variety of scripts that can be of general utility to researchers working in this area. These scripts are listed below:

Calculate_supertrees_likelihoods.py

This script allows the user to calculate the likelihood of any supertree (e.g. a supertree obtained from the literature) given a set of input trees (the input trees do not necessarily need to be the trees from which the tested tree was originally built). The input to this script can either be one supertree or a list of supertrees and the output is a file with the likelihood and RF distance values for each given supertree (see the provided L.U.St manual for details).

ExtractTaxon_file.py

This script allows the user to extract, to a user-defined file, the union of the taxon sets of a given set of input trees. For more on how to use this script, see the included L.U.St manual.

Resolve_phylogenies.py

This script enables the user to resolve the polytomous clades in a set of trees using a resolved supertree (i.e. an MRP inferred supertree). This script uses some of the capabilities of the Dendropy Python package (Sukumaran and Holder, 2010).

Polytomies are especially common when input trees have been sampled from the literature. The presented ML supertree implementation does not handle polytomies in the gene trees, so two options are provided: 1) the polytomies can be broken at random (see below), or 2) they can be resolved according to a different supertree (e.g. MRP supertree). The latter is not ideal but might be useful in some conditions. For details on how to use this script see the L.U.St manual included at the end of this thesis.

resolve_polytomies.py

Similar to the script described above, this script also gives the user the ability to resolve polytomies in phylogenies. This script, however, offers the capability to randomly resolve the polytomies in a tree. For details on how to use this script see the L.U.St manual.

deroot.py

This script uses the capability of the Dendropy Python package (Sukumaran and Holder, 2010) to allow the user to de-root rooted phylogenies. For details on how to use this script see the L.U.St manual.

Winning_site_test.py

This script allows the user to calculate the winning site test for choosing between two alternative supertree topologies given a set of input trees. For more details see section 2.2.3 above.

Statistical_test.sh and Statistical_test.py

These are two scripts that combine the capabilities of shell scripting with that of python and the CONSEL package of (Shimodaira and Hasegawa, 2001) to perform tests of trees implemented in CONSEL. For more details see section 2.2.3 above.

Note: that L.U.St and its manual are available for download from the bitBucket page - https://afro-juju@bitbucket.org/afro-juju/l.u.st.git. The manual can also be seen as Appendix A.

2.3 A Bayesian supertree method

The ability to estimate the likelihood of a supertree has another advantage. Bayesian statistical inference of phylogenies, as describe above in section 1.2.4.2.3, allows prior knowledge to be combined with the information in the data for phylogenetic reconstruction. Bayesian MCMC has already been used extensively in phylogenetics due the availability of exceptional software such as MrBayes (Ronquist *et al.*, 2012b), BEAST (Drummond and Rambaut, 2007) and P4 (Foster, 2004). The ability to estimate the likelihood of a supertree permits the introduction of supertrees to the Bayesian (MCMC) framework. This has two advantages. First, Bayesian MCMC analysis is generally faster than ML analysis. Second, Bayesian analysis allows estimation of posterior probability for clades, finally allowing for a universal measure of support for supertrees. The implementation of Steel and Rodrigo's (2008) ML supertree method was coded by Dr. Peter Foster into the already available MCMC software in P4 (Foster, 2004). The Bayesian (MCMC)

supertree method has been tested for several desired properties, including the ability to deal with large datasets (see chapters 3, 4 and 5). Differently from the L.U.St package and the ML supertree method, which I have developed in full, the Bayesian supertree method has been implemented by Dr Peter Foster (as part of a collaboration) in the package P4, and here I will only be testing his software (to complete the collaboration).

Chapter 3: Testing on Case Studies

3.1 Introduction

MRP and most other liberal supertree methods are known to suffer from either a size-related bias (Purvis, 1995a), a shape-related bias (Wilkinson *et al.*, 2005a), or both (Thorley *et al.*, 1998). A supertree method that is affected by a size bias, when faced with conflicting relationships in the input trees, will favour those in the largest of the conflicting clade(s) (Bininda-Emonds and Bryant, 1998). A supertree thats suffering from a shape bias will, in case of conflict, favour relationships in either the asymmetric or the symmetric trees (Wilkinson *et al.*, 2005a). In the case of MRP analyses, performed using the standard Baum and Ragan (Baum and Ragan, 1993) coding strategy, it is well known that the results are biased towards relationships in asymmetrical trees. However, it has been suggested that the effect might be irrelevant if large collections of informative input trees are used for the analysis (see (Kupczok, 2011a)). In the case of Purvis' coding MRP (Purvis, 1995a), which, unlike MRP, uses "?" to code for all taxa not in the clades or its sister taxon, relationships in symmetrical trees are favoured. Finally, as shown by Thorley *et al.* (1998), these two biases can add up in real examples (at the least when the inference is based on few input trees) to produce composite biases concomitantly driven by both effects. I maintain that the size and shape of an input tree should be irrelevant to its evidential significance in the supertree framework and consider the existence of these biases highly undesirable (see also (Creevey *et al.*, 2004)). Although their real

effect might be negligible (Kupczok, 2011a), the fact remains that they introduce possible doubts about the nature of actualized supertrees.

The L.U.St package includes my implementation of the ML supertree method. This is a liberal supertree method. This means that the ML supertree method will attempt to resolve any conflict among the input trees based on the available evidence. This also means that I must ensure that L.U.St's ML resolution of conflict is based solely on phylogenetic signals in the data and not on other factors (e.g. biases). Because of its applicability (see above), the MRP supertree method, despite suffering from both shape and size related biases (Bryant and Steel, 2009; Purvis, 1995a; Lapointe and Levasseur, 2004), has been widely accepted by the scientific community as the 'go to' method for the construction of supertrees from sets of less inclusive input trees. In this chapter, I shall test L.U.St's ML implementation and investigate its sensitivity to shape and size related biases. Indeed, given that the ML method should return the same result of the Majority Rule supertree (-) method (Cotton and Wilkinson, 2007), this method is not expected to be biased, yet I wanted to be sure that this was the case with respect to my specific implementation. In addition, I shall test the potential effect of these biases on the Bayesian implementation in Dr. Peter Foster's P4.

Steel and Rodrigo (2008) pointed out that when distances between trees are calculated using the symmetric difference (Robinson-Foulds distances), the ML supertrees found (given a set of input trees) correspond to the equivalent set of majority-rule consensus supertrees sensu (Cotton and Wilkinson, 2007; Barthélemy and McMorris, 1986). This is potentially a very interesting characteristic of the ML method because majority rule supertrees have the interesting statistical properties

of being median trees for the input set of trees, hence, we are measuring the central tendency in the data. This allows a non ad-hoc characterisation of the ML supertree method. In this chapter, apart from testing for biases in the ML and Bayesian (MCMC) method, I will compare alternative supertree methods, with the ML and the Bayesian one to evaluate how well they approximate the result of the Majority Rule (-) supertree method. To this end, exactly as I did when testing the performance of alternative search strategies in the previous chapter, I will be using the *Drosophila* dataset of Cotton and Wilkinson (2007). The result of the application of the Majority Rule (-) on this data set is presented in Cotton and Wilkinson (2007). The *Drosophila* dataset is composed of five phylogenetic trees overlapping on nine taxa. Using a small data set for these analyses is key because it allows me to ensure that the various methods of analyses that are being compared do not fail to return the "correct trees" simply because the problem is too complex for the search strategy they implement. As pointed out above five methods will be compared: MRP, MSS, RF, the L.U.St' ML implementation and the P4 Bayesian implementation.

3.2 Methods

3.2.1 Bias testing

The L.U.St's ML supertree implementation was tested for input tree shape effects (ITSE) using the same empirical example of Wilkinson *et al.* (2005a). To test L.U.S.T's ML supertree method for biases due to input tree shape, this data set was analysed by running it using the default heuristic search option (this is set to search option 1) for 10 iterations. The Bayesian (MCMC) supertree method was run for the same

dataset for 1,000 iterations with a β value of 1. I also reanalysed Wilkinson *et al.*'s (2005a) dataset with MRP, RF and the MSS supertree methods using their respective default settings in an attempt to be able to provide a fair comparison between L.U.St's ML implementation and the Bayesian supertree implementation on the one hand and their alternatives, on the other.

To test for biases due to input tree size, I used the example dataset from Purvis (1995a), over which most liberal supertree methods have been shown to fail. The ML analysis was run using the default heuristic search option for 10 iterations. While the Bayesian MCMC supertree method was run for 1000 iterations with a β value of 1. As above, the L.U.St's ML supertree implementation and the Bayesian (MCMC) supertree method were compared against MRP, RF and the MSS supertree method.

3.2.2 Analysis of the *Drosophila* data set

For the Bayesian analysis, I ran 2 parallel MCMC chains setting β to one for 1,000 iterations. In the ML and Bayesian supertree reconstruction, the β parameter is used to represent the quality of the trees in the data set. This is a way of differentially weighting the input trees. We can imagine that a tree constructed from high fidelity and long AA sequences has a higher β value compared to a tree constructed from a short, noisy and badly aligned AA sequences. For the ML supertree analysis, the heuristic search option 1 strategy was run, for 10 iterations. Further to that, the *Drosophila* data set was analysed using RF, MRP and MSS. Results obtained from each one of the supertree methods used for this analysis were compared. In addition, the ML and MRP scores of each one of the possible supertrees that could

be generated for the *Drosophila* data set were generated using PAUP4b10 (Swofford, 2003). The likelihood of each of these trees was calculated in L.U.St, and its parsimony score was estimated in PAUP4b10. Likelihood and parsimony scores for all these trees were plotted to evaluate the similarity and differences in scores (under the two methods) for each possible supertree.

3.3 Results

3.3.1 Testing for biases

When used to analyse the two trees (figure 3.1a and b) used by Wilkinson *et al.* (2005a) to test for input tree shape effects, the ML supertree method returned 10 supertrees. As expected if this method were not subject to tree-shape related biases, the strict consensus of these trees is fully unresolved (figure 3.1c). The mean of the Colless index (a tree balance index based on tree topology which uses an index of 1 to indicate a maximally balanced tree and an index value of 0 to indicate a maximally unbalanced tree) (Colless, 1982) for the 10 trees returned by the ML SM is 0.582, while the standard deviation is 0.026, proving that the shape of the strict consensus tree is indicative of the shape within each of the ML estimates and not due to the shape between the ML estimates. Similarly, as expected in the case of a lack of bias, the Bayesian (MCMC) supertree analysis was not able to converge: that is it was not able to decide between these two trees using the available evidence. Results of the re-analysis of the two input trees in figure 3.1a and b with MRP and MSS are presented in figure 3.1c and d, respectively, while the result obtained using RF is presented in figure 3.1e. As we shall see in the case of the *Drosophila* data set, the expectation that the RF supertree method should do as well or almost as well as

the Bayesian (MCMC) and the ML methods (given that it is an approximation of the Majority Rule (-) method) does not appear to be forthcoming. This has led me to believe that the heuristic strategy used in this approach is not effective, which I will be able to confirm after analysing the supertrees inferred by the RF method for the *Drosophila* dataset. Also in this case, the Bayesian and ML approaches seem to be the only ones (together with the Majority Rule (-) supertree method) capable of returning results consistent with the logical expectation for this example, namely that there is no shape bias in these methods.

With reference to the size bias initially highlighted by Purvis (1995a), when the ML supertree method is used to analyse the two trees in figure 3.2a and b, six supertrees of equal likelihood are found. The strict consensus of these trees is fully unresolved (figure 3.2c), as expected if this method did not suffer from a size bias. For this very simple example, the RF supertree method found the same result as the ML method, suggesting that the differences that are often observed (more on this point below) between the RF method and the L.U.St ML implementation (e.g. figure 3.3) most likely relates to the fact that the RF methods performs poorly in exploring the tree space. As in the case of the shape-bias example above, the Bayesian method failed to converge on a solution, as expected if also this method did not suffer from a size bias. In contrast, when the data is analysed using standard-MRP and MSS supertree methods as shown in figure 3.2d the topology of the largest tree is recovered. Here I have shown the susceptibility to both input tree size and shape biases for only three methods, but these are well known common ailments of all known *ad hoc* supertree methods. For an in-depth look at how other available

supertree methods deal with these biases see Wilkinson *et al.* (2005a) and Purvis (1995a).

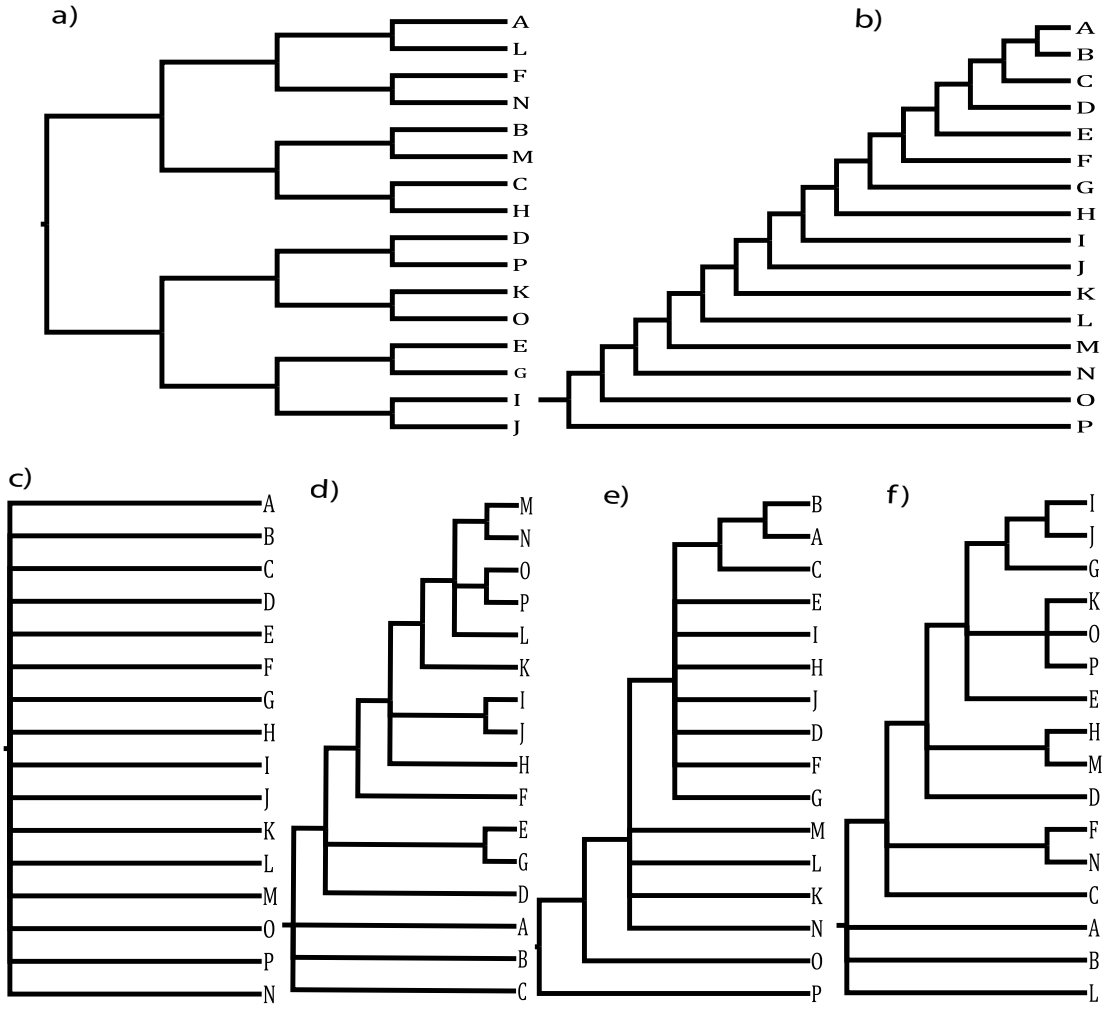


Figure 3.1: Analyses of input tree shape bias.

(a) and (b) Input trees from (Wilkinson *et al.*, 2005a) that are used as the source trees to test the shapelessness of the following supertree methods: (c) The strict consensus of the trees obtained from the ML analyses, (d) The strict consensus of the trees obtained from MRP analysis, (e) the tree obtained from the MSS analysis, and (f) the tree obtained the RF analysis.

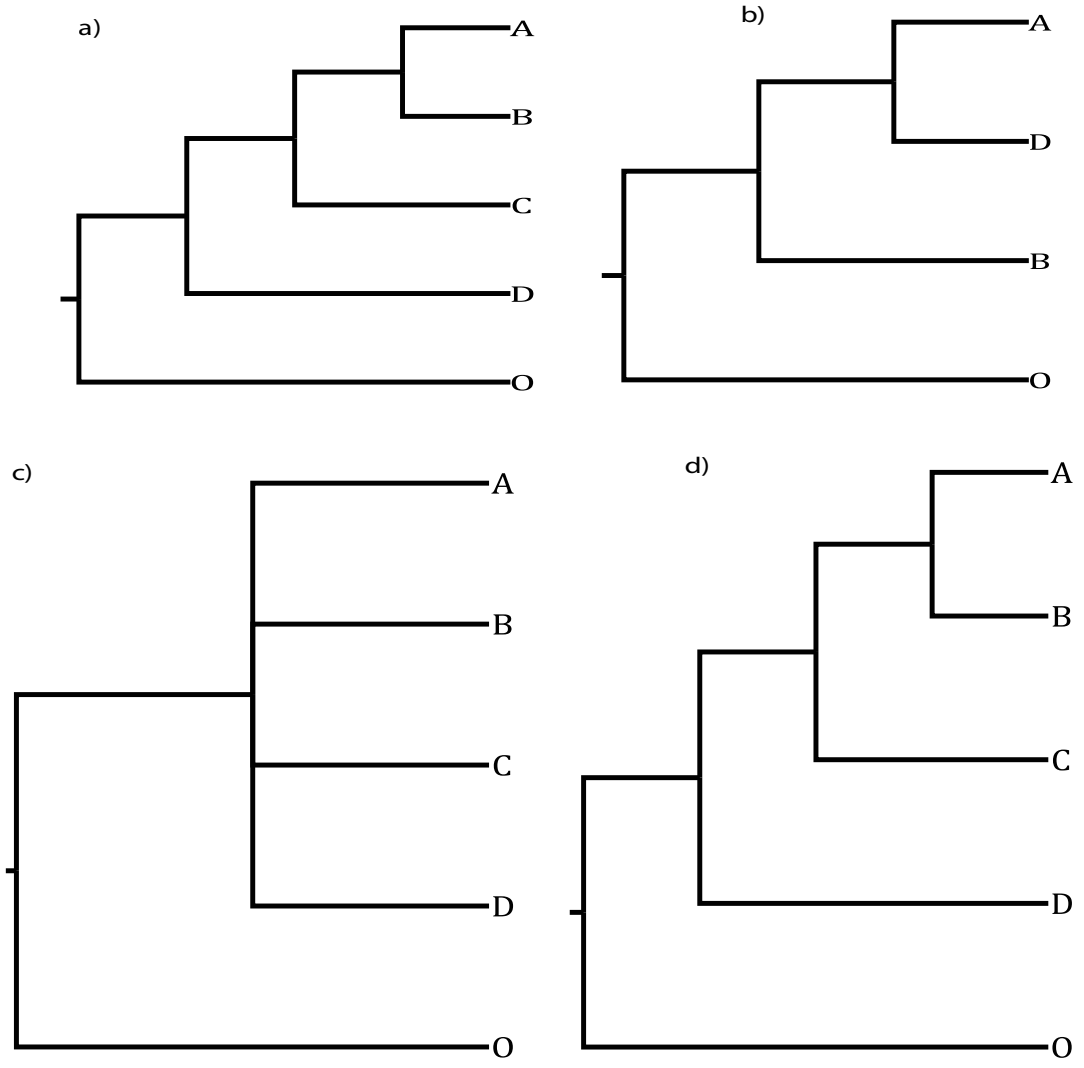


Figure 3.2: Analyses of input tree size bias.

(a) and (b) Input trees (modified from (Purvis, 1995a)) used as source trees to test the sizelessness of the following supertree methods. (c) Strict consensus of the trees obtained for the ML and RF analyses (both returned the same topology). (d) Strict consensus of the trees obtained from the MRP and MSS analyses (both returned the same topology).

3.3.2 The *Drosophila* data set

With reference to the *Drosophila* data set, the ML supertree method recovered the complete set of 79 median supertrees that were recovered by Cotton and Wilkinson using the Majority-Rule (-) method. These trees were used to construct the strict consensus tree presented in figure 3.3f. Using the Bayesian (MCMC) analysis, a tree that is topologically identical to the strict consensus tree in figure 3.3f was recovered. However, the 79 median trees identified using both ML and the Majority-Rule (-) supertrees method were obviously not recovered. This should not be viewed as a problem of the Bayesian (MCMC) supertree method, but as a result that is expected and a consequence of this method being based on an MCMC approach. The MRP analysis of the same dataset returned 77 equally parsimonious trees. As pointed out by Cotton and Wilkinson (2007), these trees represent a subset of the known (complete) set of 79 median trees identified using the ML supertree method and their Majority Rule (-) supertree method. The MSS and the RF supertree methods found 42 and 26 median supertrees respectively out of the expected total of 79.

The ML supertree and the Majority Rule (-) supertree methods were the only methods that were able to identify correctly the 79 median trees that exist for this input collection of tree topologies. Interestingly, the MRP supertree method fared quite well, when compared to the other tested approaches, as it only failed to recover 2 known median supertrees. However, MSS missed 37 and RF missed 53 of the 79 known median supertrees. The RF supertree method performed very poorly, particularly if one considers that this approach is a heuristic approximation of the Majority Rule (-) method and should be expected to approximate the result of the

latter. Instead, RF could only recover a minority (~ 32%) of the known median supertrees. It is interesting to note that none of the methods considered recovered trees that did not belong to the collection of 79 median trees; they simply failed to recover the entire set. This is an encouraging result as it suggests that, at least for the methods considered here (and for this admittedly simple example), alternative supertrees, rather then differing in their accuracy (finding the best tree), seem to differ only in their level of precision (finding all the best trees).

Given that MRP performed particularly well, I decided to estimate, plot, and compare, for each possible tree on the same leaf set of the *Drosophila* data set (135,135 supertrees in total), its likelihood score and its parsimony score. Results (figure 3.4) show that there is generally a good correspondence between the likelihood and the parsimony fit to the trees, but that this is not a universal finding and mismatches do exist.

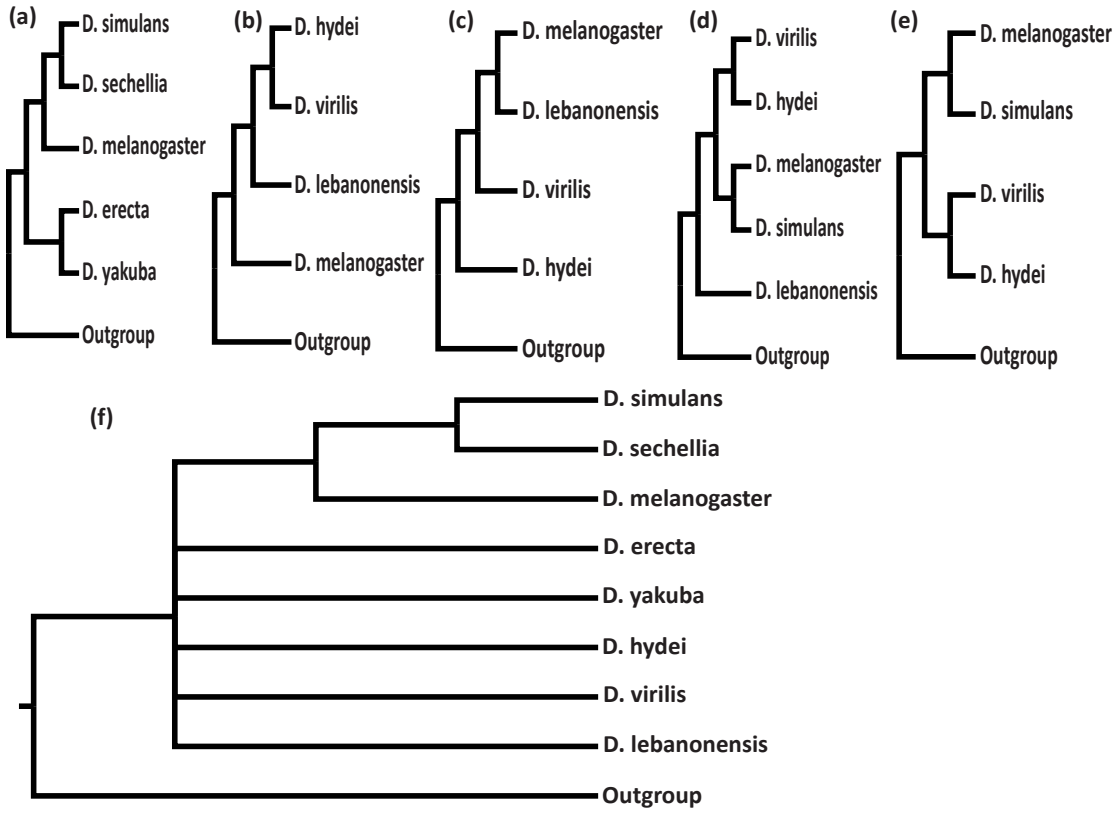


Figure 3.3: ML supertree analysis of *Drosophila* empirical dataset.

(a-e) Input trees from (Cotton and Wilkinson, 2007). (f) The strict consensus of the 79 trees retrieved by the ML supertree method. This is the same topology for the strict consensus of the 77 MRP supertrees, the strict consensus of the 42 MSS supertrees, and the strict consensus of the 26 RF supertrees.

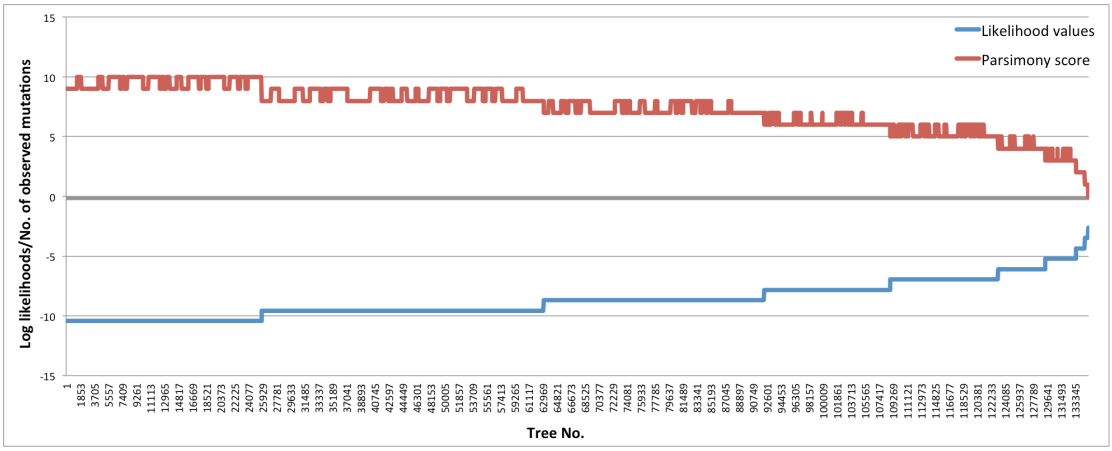


Figure 3.4: A line graph of MRP parsimony scores and likelihood scores.

This is for each of the 135,135 possible supertrees on the union of taxa of the *Drosophila* input tree from figure 1a-e. Note that the values have been scaled to allow for a better comparison. Scaling was performed by subtracting 15 from the parsimony values (blue) and while the log likelihood values (red) were left as they were.

3.4 Discussion

In the performed tests the Bayesian (MCMC) and the ML supertree approaches fared well, overall illustrating that these methods might perform better than any of the available ad hoc methods with real data sets. The maximum likelihood (ML) method returns results that are comparable to those of the Majority rule (-) consensus supertree method, and from this point of view the ML method would appear to be effectively redundant. However, its ability to transform RF distances into probabilities allows for two important and otherwise impossible advancements in supertree reconstruction: the development of Bayesian methods and the integration of the standard statistical test of two trees to the supertree context (see chapter 2 and 4). So in essence the ML supertree method appears to have taken over and

fulfilled the promise that was shown by the Majority rule (-) supertree method. Both of these advantages will be demonstrated and further explained in chapter 4, where I shall use the Bayesian (MCMC) supertree method and the L.U.St package to reanalyse a set of real world supertree data sets.

3.5 Conclusions

The results presented in this chapter show that the ML and the Bayesian (MCMC) supertree approaches are viable alternatives to MRP and to other supertree methods. With the introduction of both of these new parametric approaches, it is finally possible to have confidence in the supertrees that are being inferred.
MRP and the other ad hoc methods tested in this chapter have been proven again to suffer from either input tree shape bias or input tree size bias or both. Regardless of whether or not the effect of these biases is strong enough to affect results of analyses based on large data sets, its detection in any supertree method should warn against using such methods.

Chapter 4: Reanalysis of Real world Data sets

4.1 Introduction

The literature on supertree reconstruction is growing by the day and many promising approaches have been proposed and developed to solve the supertree problem, generating larger and more inclusive phylogenies from set smaller phylogenies on overlapping taxa. Many of these approaches are ad hoc. However, as we have seen (in the case of MRP in particular) some of these methods infer trees that could be considered good approximations of the median supertrees of a given set of trees. In addition the relationships observed in these trees are often biologically plausible, which would confirm that the trees inferred using these methods are not of an unfeasibly poor quality. However, the properties of these methods are not well understood and this implies that, in cases of conflict among the input trees, it can be difficult to evaluate whether the result of a supertree analysis is due to bias or signal in the data (Wilkinson *et al.*, 2005a). Further to that, with standard methods, calculating support for clades is difficult and developing robust statistical tests of trees are difficult or virtually impossible.

The properties of the supertree methods implemented and applied in this thesis are well formulated and understood (Steel and Rodrigo, 2008). In addition, they seem to be immune to both input tree shape and input tree size effects. In this chapter both of the supertree methods implemented will continue to be pitted against the most commonly used alternatives. In particular the aim of this chapter is to illustrate how the L.U.St package's ML and the P4 implemented Bayesian (MCMC) supertree methods can be used in real data analyses, and how they would compare

in such situations against other supertree methods (MRP, RF and MSS). The current implementation of L.U.St's ML supertree method is inadequate for extremely large analyses (as tree searches would become too slow), even if it can handle tens of taxa and hundreds of trees (when using the fastest heuristic strategies). However, such search strategies might be inaccurate. Hence, the Bayesian supertree method will be used for tree search and the L.U.St package will be used to perform tests of trees and other statistical analyses. I suggest that this is the best way to combine these tools to analyse real world data sets.

4.1.1 The Metazoan dataset

The first dataset I reanalysed was the metazoan data set of Holton and Pisani (2010). This data set included 42 taxa and 2,216 trees.

The relationships among the animals with bilateral symmetry are notoriously difficult to resolve, and a multitude of conflicting hypotheses have been proposed (Jenner and Schram, 1999). Two of these alternative hypotheses have dominated the debate on metazoan phylogeny. These are Hyman's Coelomata hypothesis (Hyman, 1940) and the Ecdysozoa hypothesis (Aguinaldo *et al.*, 1997). The former hypothesis has been the dominant view in the scientific community for a long time. It proposes that the bilateral animals fall into three groups: the Acoelomata (which include the Platyhelminthes and the Nemertea), the Pseudocoelomata (which include the Nematoda, the Nematomorpha, the Rotifera, the Gastrotricha, the Kinorhyncha, and the Priapulida), and the Coelomata (containing the remaining bilaterian phyla, e.g. Arthropoda, Mollusca, Annelida, and Vertebrata) (Philippe *et al.*, 2005b; Telford *et al.*, 2008; Holton and Pisani, 2010). The latter hypothesis

proposes a separation of the bilateral animals into two groups: Protostomia and Deuterostomia. The Ecdysozoa hypothesis further suggests that the Protostomia should be divided into two groups Lophotrochozoa and the Ecdysozoa. The Coelomata hypothesis has long been backed by evidence from both morphological and deep genomic data analyses (Hyman, 1940; Blair *et al.*, 2002; Wolf *et al.*, 2004) while most of the evidence for the Ecdysozoa hypothesis was from 18S rRNA datasets and a handful of genomic datasets, mostly expressed sequence tags (ESTs) (the product of cloned cDNA sequencing) datasets (Philippe *et al.*, 2005a; Dunn *et al.*, 2008). Holton and Pisani (2010) employed the MRP supertree method to analyse a genomic data set composed of 42 taxa overlapping on 2216 gene trees, and recovered a tree displaying the Ecdysozoa hypothesis (differently from most other deep genomic-scale analyses) (figure 4.1). Their results were used to conclude that the Hyman's hypothesis was a by-product of long branch attraction (LBA) (Holton and Pisani, 2010). Given the importance of this supertree-derived result, it is interesting to investigate whether it holds to the application of presumably better performing supertree methods such as the parametric Bayesian MCMC supertree method.

4.1.2 The carnivore data set

The metazoan dataset of Holton and Pisani (2001), being a genomic data set, contained highly overlapping trees. Hence, for the second real world dataset to test both the ML and Bayesian methods on, it was decided that a more challenging data set, the carnivore data set of Nyakatura and Bininda-Emonds (2012), should be used. This data set represents a more traditional example of an application of supertree

methods, where the data are not gene trees. Instead, the input trees have been sourced from the literature. The trees in this dataset, as one might expect, contain a considerably lower level of taxon overlap in comparison to the metazoan dataset.

Carnivores include a large number of both terrestrial and aquatic mammal species, and represent one of the largest mammalian orders. Nyakatura and Bininda-Emonds used this data set to update an original carnivore phylogeny from 1999 (Bininda-Emonds *et al.*, 1999). This update was necessary for several reasons, including taxonomic changes, the increase in available sequenced data, additional information from other types of data and the methodological improvements in the original analyses from which trees were derived (Nyakatura and Bininda-Emonds, 2012). For example, *Nandinia* (African palm civet) now forms a sister taxon to the rest of the feliform carnivores (Flynn *et al.*, 2005), and the Mephitidae have been removed from Mustelidae (Dragoo and Honeycutt, 1997).

The new dataset of Nyakatura and Bininda-Emonds (2012) is composed of 286 taxa and 558 trees. This included a “taxonomy tree”, which is a tree (derived from a taxonomic list) that these authors used to shoehorn misbehaving taxa into a supertree phylogeny. In addition to that, Nyakatura and Bininda-Emonds (2012) used a differentially weighted MRP supertree method to infer a well-resolved carnivore phylogeny that conveys the current accepted biological views of the carnivore order.

4.2 Methods

4.2.1 Supertree analysis of the Metazoa

The metazoan dataset is composed of 42 taxa overlapping on 2216 gene trees. For the analysis of this dataset two MCMC chains for 10,000 iterations were run, while sampling once every 100th iteration. I tested for convergence by comparing the log likelihoods of the trees sampled by the two chains. Different analyses were performed in which the β values were changed. Beta values tested were: 0.001, 0.01, 0.1, 0.5 and 1. Holton and Pisani (2010) previously analysed this data set using MRP and estimated support for the nodes in the tree they recovered using input tree bootstrapping. Accordingly, the Bayesian analyses performed here, can be compared against the MRP results of Holton and Pisani (2010) to clarify how similar the recovered supertrees are and how closely the Bayesian Posterior probabilities estimated using our MCMC approach compare with the bootstrap probabilities obtained for the MRP tree. In addition, in order to compare the Bayesian supertree method against other supertree methods, the data set of Holton and Pisani (2010) was reanalysed using the MSS and the RF supertree methods. MSS is implemented in the phylogenetic software CLANN (Creevey and McInerney, 2005) and was run using the default options. The RF supertree method (Bansal *et al.*, 2010) was downloaded and installed on a local server and again was run using the default parameters.

4.2.2 Supertree analysis of the Carnivores

To evaluate how the Bayesian (MCMC) supertree method fares using more traditional supertree datasets (i.e. collection of trees derived from the literature, rather then based on genomic data sets), I further tested the Bayesian (MCMC) supertree method using the carnivore dataset of Nyakatura and Bininda-Emonds (2012). The original dataset was kindly provided by Olaf Bininda-Emonds and was composed of 286 taxa and 558 input trees. This dataset included polytomous trees along with their various resolutions and the taxonomy tree. In order to analyse this dataset using the Bayesian (MCMC) supertree method, I randomly resolved the polytomous input trees, using the *Resolve_polytomies.py* script in the L.U.St package section 2.2.4. This was necessary because the Bayesian supertree can currently only deal with resolved trees. The first step in trying to resolve the polytomies was to separate polytomous trees from non-polytomous ones. This was achieved by writing a python script that utilised the python package ete2a1 (Huerta-Cepas *et al.*, 2010). The second step involved writing another python script to run the *Resolve_polytomies.py* script, which utilizes tools in the Dendropy library (Sukumaran and Holder, 2010). Each polytomous input tree was randomly resolved 10 times. I thus generated 10 datasets composed of 274, fully resolved, input trees on 271 taxa (see Appendix B for a list of the taxa). As my interest was not in obtaining a well–resolved carnivore tree but in comparing alternative supertree methods, I excluded from all Bayesian (MCMC) analyses the taxonomy tree used in the original study. Each of the 10 datasets generated was analysed using the Bayesian (MCMC) supertree method. For each analysis, two MCMC chains of 5million iterations were run, sampling once every 1000 iterations. As with the

analysis of the metazoan dataset, I ran two chains for each Bayesian analysis to check for convergence. The β value was set to one for all runs. The trees sampled after convergence from all ten runs (a total of 30,020) were then merged and a majority rule consensus tree was constructed in PAUP4b10 (Swofford, 2003). The majority rule consensus tree constructed from the 30,020 Bayesian (MCMC) sampled supertrees was compared with the MRP tree of Nyakatura and Bininda-Emonds (2012). However, the latter included the taxonomy input tree, which was not included in my analysis. In addition, Nyakatura and Bininda-Emonds (2012) also used a differential weighting scheme in their analyses, whereas equal weighting was imposed in the Bayesian (MCMC) supertree analyses performed here. This was done for simplicity, even though the Bayesian (MCMC) supertree method has the capabilities to differentially weight the input trees – see chapter 2. Accordingly, in order to carry out a fair comparison of the MRP and the Bayesian (MCMC) supertree approaches, I reanalysed the 10 modified (unpolytomised) versions of the dataset of Nyakatura and Bininda-Emonds (2012) using equally weighted MRP. The phylogenetic package CLANN was used to generate the MRP matrix and PAUP4b10 was used to analyse the matrix. Parsimony analyses in PAUP used the following parameters. (1) 100 random additions with the multree option turned off. (2) Trees that were saved from the initial set of 100 random additions were used to run the MRP analysis with the multree option turned on. This is the same strategy used, for example, by Lloyd *et al.* (2008). This returned 585,166 equally parsimonious supertrees that were summarised in a majority-rule consensus tree. Finally, a MSS and RF supertree of the Carnivora were also derived, using their respective default settings, to compare the Bayesian approach with other supertree methods.

4.2.2.1 Carnivore dataset leaf stability test

From each of the 10 data sets analysed using the Bayesian approach a set of 100 supertrees were sub-sampled (after convergence). These trees were used as input to a subsequent analysis devised to investigate the presence of rogue taxa, taxa that are unstable in their positions in the set of trees (sensu (Wilkinson, 1994)). This was done using the *LeafStability.py* script in P4 (Foster, 2004). The leaf stability test identified 26 highly unstable taxa. The fact that these taxa do not appear unstable in Nyakatura and Bininda-Emonds (2012) MRP tree is a consequence of the fact that these authors used a taxonomy tree to shoehorn unstable taxa. These 26 taxa were deleted from the set of 30,020 Bayesian supertrees and from the 585,166 most parsimonious trees (MPTs) obtained from my new, equally weighted MRP analysis (keeping trees that become identical after pruning). New majority rule trees for the Bayesian (MCMC) analysis and the equally weighted MRP analysis were derived. Finally, to allow for a full comparison across all the considered trees, these 26 taxa were also pruned from the Nyakatura and Bininda-Emonds (2012) weighted MRP supertree.

4.2.3 Statistical test of metazoan and carnivore supertrees

The Approximately Unbiased (AU) test was used to compare the alternative supertrees (Bayesian (MCMC), equally weighted MRP and differentially weighted MRP) for the carnivores, and also for the alternative supertrees obtained for the metazoans. In addition, a sample of 100 random (super)trees was generated with the same taxon set as in the carnivore dataset and a set of 1000 random (super)trees

were generated on the taxon set of the metazoan dataset (using PAUP4b10 (Swofford, 2003)). The likelihood values of these random supertrees were estimated (using the *Calculate_supertrees_likelihoods.py* script from the L.U.St package), plotted, and compared against the re-estimated likelihood values for all the set of MRP, RF, MSS and Bayesian (MCMC) supertrees that I recovered for both the metazoan and the carnivore datasets. This was done to understand better whether these methods did better than random, and how much better.

4.3 Results

4.3.1 Bayesian (MCMC) metazoan phylogeny

There were only minimal differences in the posterior probabilities of the clades in the Bayesian supertrees obtained when alternative β values (0.001, 0.01, 0.1, 0.5 and 1) were used. Therefore, from now on I shall focus on results obtained from the Bayesian (MCMC) analysis with β set to 1. Figure 4.1 shows the Majority rule consensus of the 150 trees sampled after convergence. It illustrates the set of relationships uncovered and their support (represented as nodal posterior probabilities). Supertrees inferred using the MSS and the RF supertreee methods are reported in figure 4.2a and 4.2b respectively. The RF analysis returned 15 supertrees; figure 4.2a shows the majority rule consensus of these. The Bayesian (MCMC) tree in figure 4.1 is topologically identical to the MRP tree of (Holton and Pisani, 2010). Posterior probabilities for the nodes in this tree are also entirely comparable with the bootstrap support values of the MRP tree (see (Holton and Pisani, 2010) fig.3). Importantly, the Bayesian supertree in figure 4.1 (exactly as the MRP supertree presented in Figure 3 of Holton and Pisani (2010)) recovered a set of

relationships among the considered taxa that are in full agreement with current knowledge of animal relationships (including confirmation of the Ecdysozoa hypothesis). When this phylogeny is compared with those obtained using the MSS and the RF supertree method, figure 4.2a and 4.2b respectively, it is clear that the trees obtained from these analyses do not agree with the current knowledge of animal relationships. The MSS supertree incorrectly resolves the relationships among the mammal species, while the RF supertree display a greater number of obviously incorrectly resolved nodes. Overall, taking results of the previous chapter into consideration, the RF supertree seems to be the worst performing supertree methods.

The likelihood scores for the metazoan topologies inferred by the Bayesian MCMC, MRP, RF and MSS supertree methods were compared to the likelihood scores for 1000 randomly generated metazoan supertrees. This analysis is presented in figure 4.3.

Finally, Table 4.1 illustrates the results of the test of two trees, including the AU test, which was performed to compare the topologies inferred by the MRP/Bayesian (since they returned the same topology), RF, and MSS supertrees methods. Table 4.1 shows that only the topology inferred by the Bayesian (MCMC) and by MRP cannot be rejected by the AU test. Taken together, the results of figure 4.2 and Table 4.1 show that the MRP and Bayesian (MCMC) supertree methods accommodate the metazoan data significantly better than the MSS and RF supertrees. However, they also show that the set of supertree methods considered found trees that are significantly better than randomly generated topologies. It is particularly surprising that RF found trees were significantly worse than MRP, as this

method would have been expected to return trees similar to those generated by the Majority Rule (-) method (Cotton and Wilkinson, 2007). This further confirms that the RF algorithm does not provide a particularly accurate approximation of the Majority Rule (-) method, while further confirming that MRP, despite its known problems, performs reasonably well with real-world datasets.

4.3.2 Bayesian (MCMC) carnivore phylogeny

The Bayesian (MCMC) supertree, (figure 4.4) obtained from the analysis of the carnivore dataset was quite different from the MRP tree presented in Nyakatura and Bininda-Emonds (2012), and it appears that the placement of a variety of taxa might have been erroneous in the Bayesian (MCMC) tree. However, an inspection of support levels (posterior probabilities) suggested that there could have been several rogue taxa in the dataset. These rogue taxa were not a problem in the study of Nyakatura and Bininda-Emonds (2012) because these authors used a taxonomy tree and a differential weighting scheme to shoehorn them. As pointed out in the method section of this chapter, I chose not to use a taxonomy tree or a differential weighting scheme in my investigations, as I did not want to have to evaluate factors other than the supertree method itself in investigating the performance of the Bayesian (MCMC) and ML supertree methods.

To assess the influence of rogue taxa, I performed a leaf stability analysis (see section 4.2.2.1). Twenty-six highly unstable taxa were identified by the leaf stability test (see full ranked list in Appendix B) and pruned from the sampled Bayesian supertrees, and a new majority rule consensus tree was derived (figure 4.5). The new (pruned) Bayesian supertree has generally high levels of support and is in good

agreement with the tree of Nyakatura and Bininda-Emonds at the ordinal level (see figure 4.5 and 4.6). Indeed, at this level, the only nodes where the two trees disagree can be shown to have low support in the Bayesian tree, i.e. a non-monophyletic *Viverridae* which has a posterior probability of 44%, suggesting that there is not much signal in the data to infer a monophyletic *Viverridae* clade and that the placement of these taxa cannot be considered reliable.

Figure 4.6a shows the majority rule consensus tree for the MRP analysis of the carnivore dataset, that was performed using the same dataset used for the Bayesian (MCMC) analysis and equal weighting. This result makes it possible to compare objectively the ability of the Bayesian (MCMC) supertree method to infer a biologically plausible topology for a challenging dataset, such as the carnivore dataset, with that of the MRP supertree method.

A comparison of the Bayesian (MCMC) majority rule consensus tree (figure 4.5) with the equally-weighted MRP majority rule consensus tree (figure 4.6a) and with Nyakatura and Bininda-Emonds differentially weighted MRP tree (figure 4.6b), after the removal of the 26 unstable taxa from each of them, illustrates clearly that the equally weighted Bayesian (MCMC) majority rule consensus tree represents more biologically plausible relationships and a more resolved phylogeny than the equally weighted MRP majority rule consensus tree. Indeed, when the taxonomy and differential weighting are not considered, MRP analysis of this data set returns a tree that is both biological highly implausible and extremely different from both the equally weighted Bayesian (MCMC) majority rule consensus tree and the differentially weighted MRP tree from Nyakatura and Bininda-Emonds (2012). Supertrees built with the MSS and the RF supertree methods both differed to some

extent from the supertrees inferred using the Bayesian and MRP methods; hence these results are not presented due to their poor performance with this data set.

The result of the test of two trees showed that the Bayesian MCMC majority rule consensus tree fits the data better than the trees inferred by both types of MRP analyses (see Table 4.2). Indeed, the topologies inferred by both the differentially and equally weighted MRP analyses are rejected the AU test (see Table 4.2).

Figure 4.7 shows, as in the case of the metazoan dataset, that in the case of the carnivore dataset all supertree inference methods considered here returned supertrees that are significantly better than random.

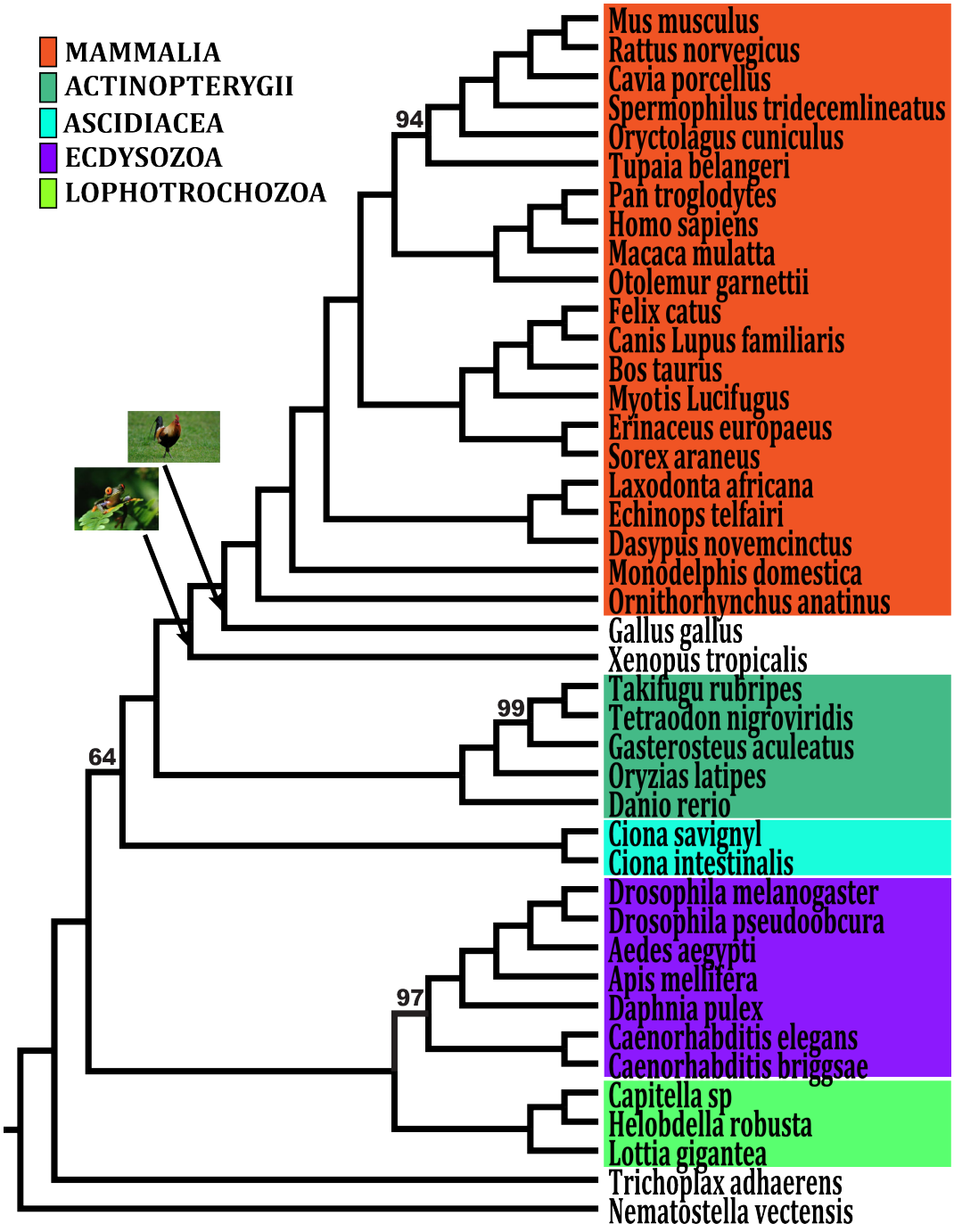


Figure 4.1: Bayesian (MCMC) phylogeny of the Metazoans.

This is the majority rule consensus tree of the 150 supertrees sampled from Bayesian (MCMC) analysis, and also represents the topology recovered by Holton and Pisani (2010) using MRP, for the analysis of the Metazoa. This data set is composed of 2216 gene trees overlapping on 42 taxa. The red coloured branch represents the branch leading to the Ecdysozoa group. Clade support is shown as posterior probability scores. Clades with no support value shown have maximum posterior probability scores.

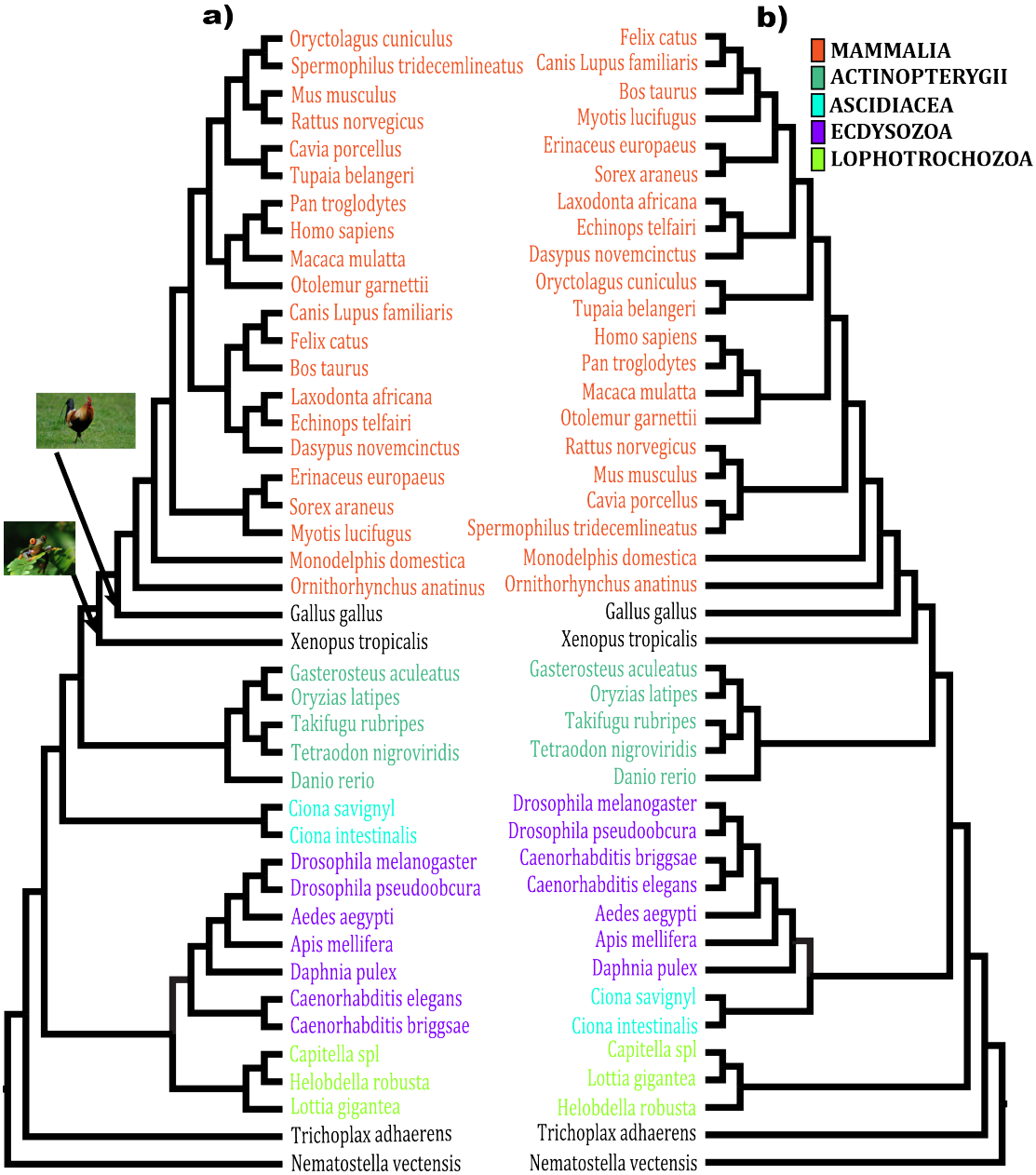


Figure 4.2: **Phylogenomic supertrees of the Metazoan.**

(a) The phylogeny recovered using MSS. (b) The majority rule consensus of the 15 phylogenies inferred using the RF supertree method. This data set is composed of 2216 gene trees overlapping on 42 taxa. The red coloured branch represents the branch leading to the Ecdysozoa group.

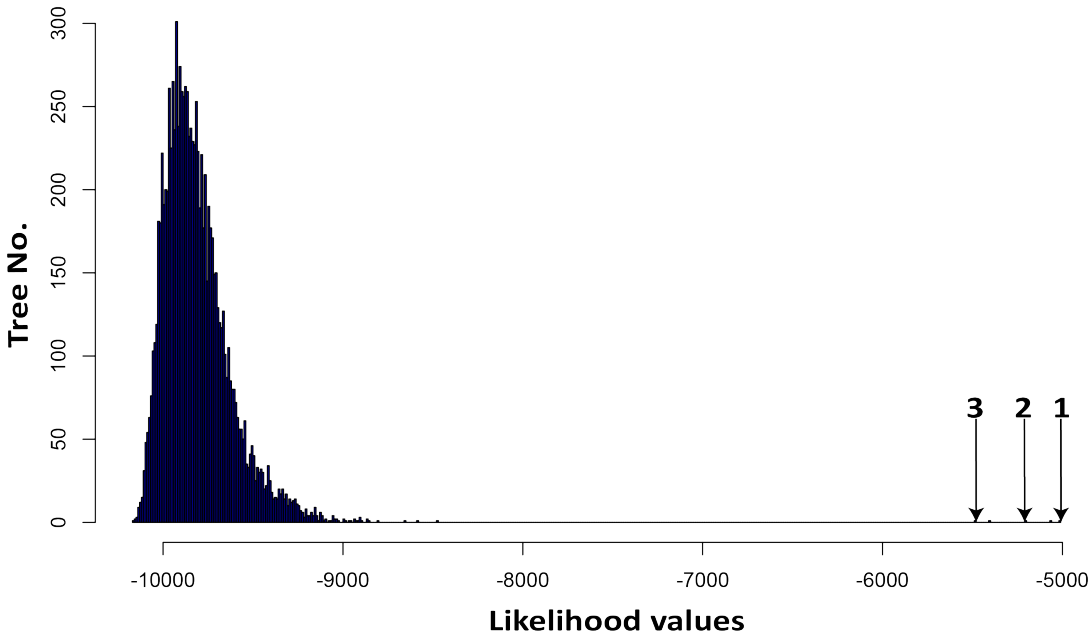


Figure 4.3: Distribution of Metazoan supertrees likelihood scores.

This graph illustrates the comparison in the distribution of the likelihood scores for 1,000 random supertrees on the same taxon set of the metazoan data set and the likelihood scores for the metazoan phylogenies inferred by the (1) Matrix representation with parsimony and Bayesian (MCMC), (2) Most similar supetree, and (3) Robinson foulds supertree method. The x-axis represent the log likelihood score while the y-axis represents the number of trees.

Supertree methods	AU test	SH test	KH test
Bayesian supertree/MRP	1	1	1
RF	2E-35	0	0
MSS	3E-12	0	0

Table 4.1: Summary of the statistical tests of the Metazoan supertrees. This table illustrates the probability values of the test of two or more trees for supertrees (implemented in L.U.St) for the phylogenies inferred for the metazoans. Legend: AU – Approximately Unbiased, SH – Shimodaira-Hasegawa, KH – Kishino-Hasegawa tests.

Supertree methods	AU test	SH test	KH test
Bayesian supertree	0.702	0.951	0.695
Equally weighted MRP	0.003	0.303	2.00E-04
Differentially weighted MRP	1.00E-49	0.005	0

Table 4.2: Summary of statistical tests of the Carnivore supertrees. This table illustrates the probability values of three of the test of two or more trees for supertrees (implemented in L.U.St) for the phylogenies inferred for the carnivores. Legend: AU – Approximately Unbiased, SH – Shimodaira-Hasegawa; KH – Kishino-Hasegawa.

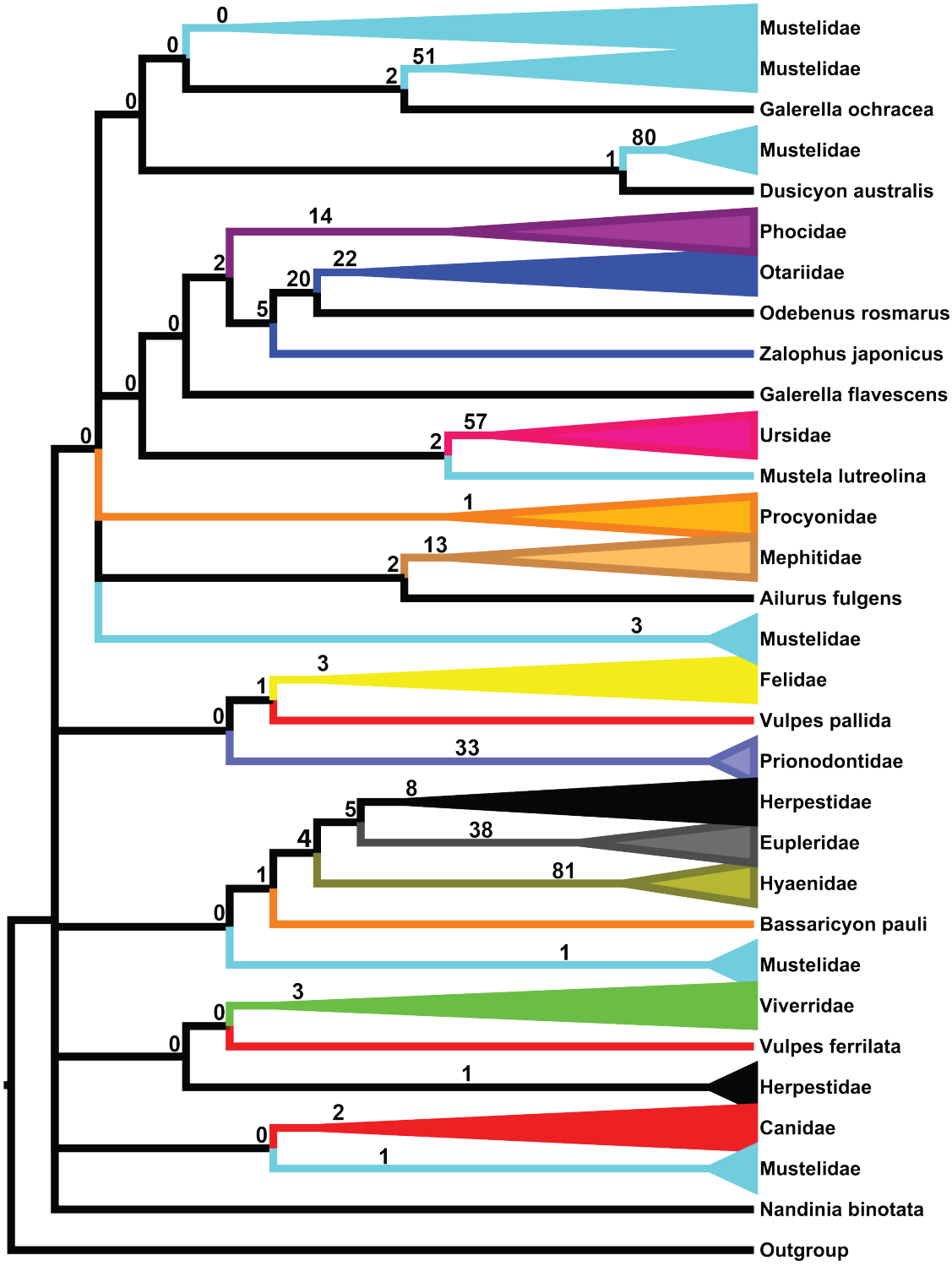


Figure 4.4: Bayesian (MCMC) phylogeny of the Carnivores.

This is the majority consensus tree of the combined 30,020 supertrees sampled by the 10 MCMC analyses after convergence and discarding of the burn-in. Clade support is shown as posterior probability scores. Clades with no support value shown have maximum posterior probability scores.

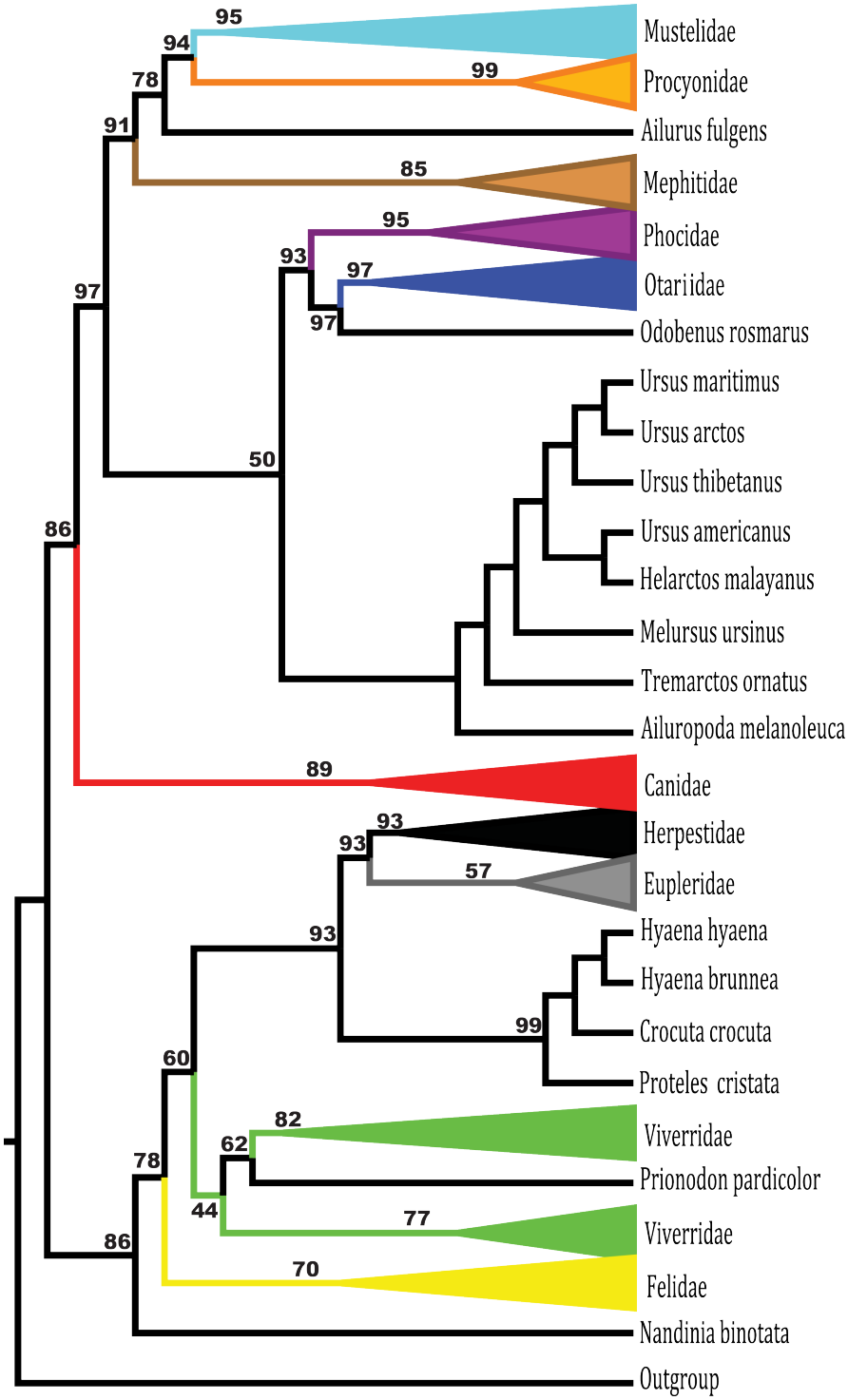


Figure 4.5: Bayesian (MCMC) Carnivore phylogeny excluding rogue taxa.

This is the majority rule consensus tree of the combined 30,020 supertrees sampled by the 10 MCMC analyses after convergence, discarding of the burn-in and pruning of the top 26 ranked unstable taxa. Clade support is shown as posterior probability scores. Clades with no support value shown have maximum posterior probability scores.

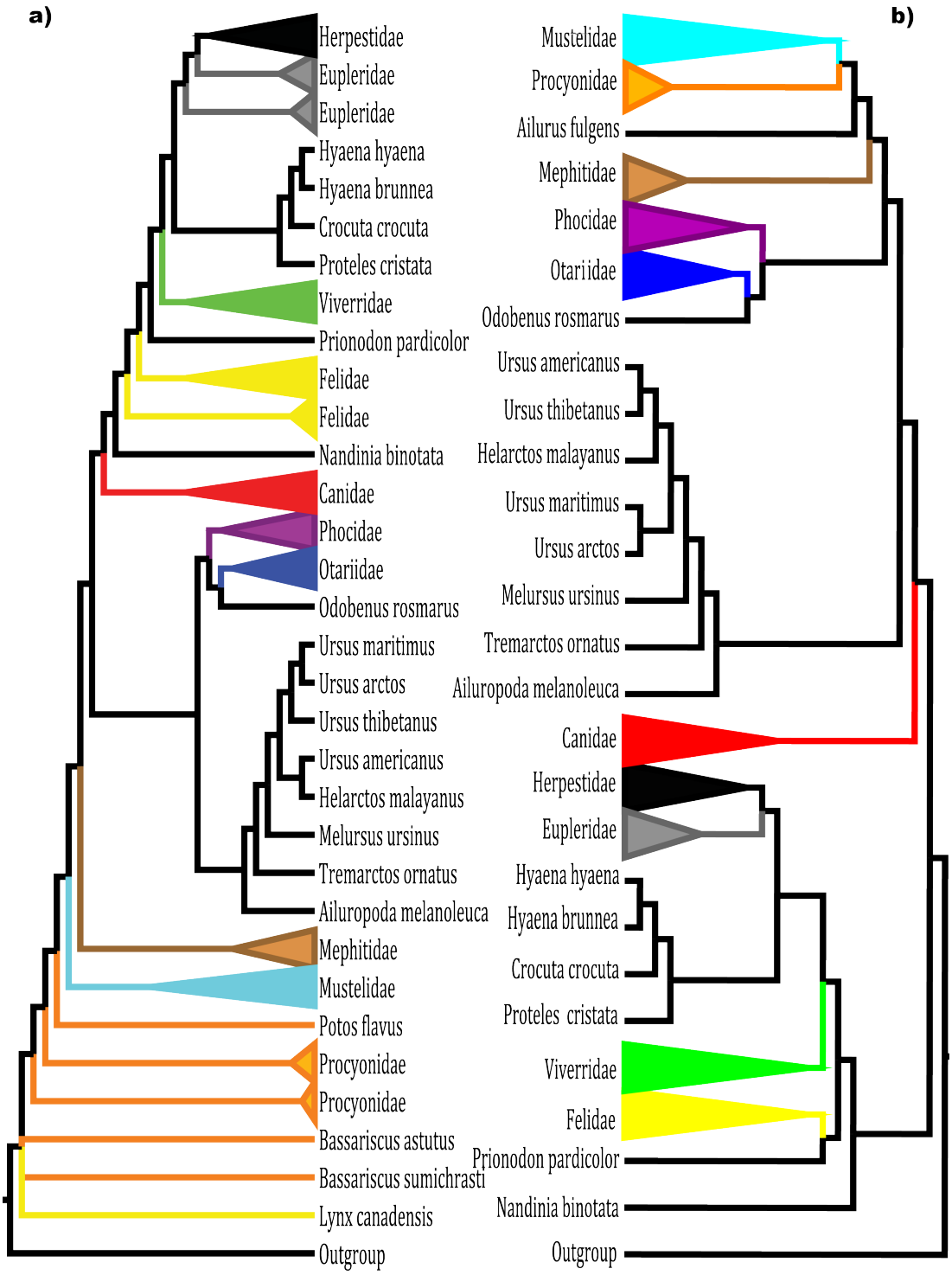


Figure 4.6: Phylogenomic supertrees of the Carnivora with rogue taxa pruned.

The top 26 ranked unstable taxa identified by the leaf stability test have been pruned from each of these phylogenies (a) Majority rule consensus tree of the 585,166 most parsimonious trees inferred by the equally weighted MRP analysis of the modified carnivore dataset, (b) Differentially weighted MRP phylogeny (presented by Nyakatura and Bininda-Emonds (2012)).

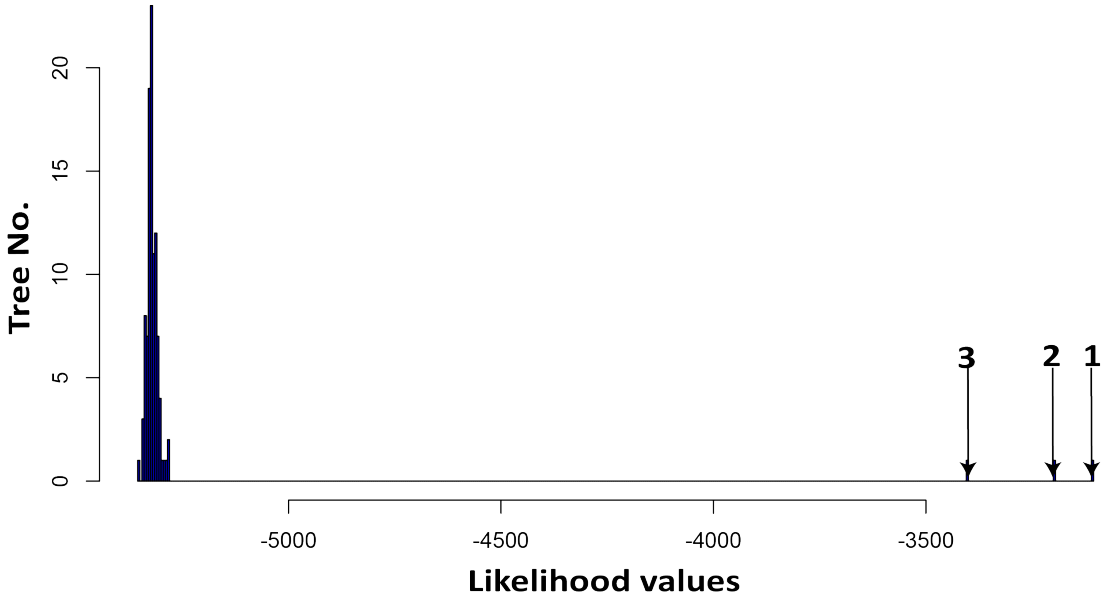


Figure 4.7: Distribution of Carnivore topologies likelihood scores.

This Graph represents the comparison of the distribution of the likelihood scores for 100 random supertrees on the same taxon set of the carnivore dataset, the dataset used for the Bayesian (MCMC) analysis, and the likelihood scores for the carnivore phylogenies inferred by (1) Bayesian (MCMC), and MRP ((2) differentially weighted) and (3) equally weighted) supertree methods. The x-axis represent the log likelihood score while the y-axis represents the number of trees

4.4 Discussion

The aim of this chapter was to see how well the Bayesian (MCMC) supertree method performs in comparison with commonly used supertree methods, in particular MRP. The results of the analysis of the metazoan dataset initially proved, even before comparing the Bayesian (MCMC) metazoan phylogeny with the MRP inferred phylogeny (Holton and Pisani, 2010), that the Bayesian (MCMC) supertree method returns phylogenies that are biologically plausible in a very efficient time. It is noticeable that, while the Bayesian (MCMC) supertree method performed as well as

the MRP supertree method in the case of the metazoan dataset (where taxon overlap is high among the input trees), this method clearly outperformed the MRP approach in the case of the Carnivore dataset, which represented a much more challenging example (due to the lower levels of overlap between the input trees). Indeed, the equally weighted Bayesian (MCMC) supertree method (once the 26 unstable taxa were removed) found a solution that is essentially consistent with that obtained using the MRP method (but only when this methods was used on a data set that included a taxonomy tree to effectively cover for the lack of phylogenetic signal, and differential weighting). Indeed, when the same data set is analysed using an "equally weighted and taxonomy tree-less" MRP approach, the result obtained is biologically implausible (see figure 4.4a). Thus, an equally weighted Bayesian (MCMC) supertree analysis performed as well as a differentially weighted MRP supertree analysis and significantly better than an equally weighted MRP supertree analysis. In addition, it is clear that the Bayesian (MCMC) supertree approach, by providing posterior probabilities for the nodes in the supertree allow for a simple interpretation of the support for the nodes in the carnivore phylogeny.

I also noticed how much more biologically plausible the phylogenies inferred by the Bayesian (MCMC) supertree method, in relation to both the metazoan and the carnivore data sets, is in comparison to the phylogenies inferred by the MSS and the RF supertree methods.

4.5 Conclusion

In conclusion, based on the result of this chapter, we can confidently say that the Bayesian (MCMC) supertreee method returns phylogenies that accomodate the data from which they have been inferred very well. The Bayesian approach performs as well as MRP for datasets with high overlap and seems to perform better for more challenging datasets. Although the ML supertree method implemented in L.U.St is currently too slow to handle large datasets, the performance of the Bayesian (MCMC) supertree method in the analysis of both the metazoan and the carnivore datasets suggests that this method can handle the most challenging of datasets and that it should be the preferred method for supertree reconstruction. On the basis of this conclusion, in the last chapter of this thesis I will exploit the potential of the new methods to address the analysis of a very large and challenging data set (a genomic data set scoring hundreds of taxa across the three domains of life). I seek to test hypotheses about the origin of cellular life, but also to evaluate how the Bayesian supertree method fares when dealing with extremely challenging data sets.

Chapter 5: Tree of Life

5.1 Introduction

Representing the evolutionary history of all extant organisms on earth as a single Tree of Life is like a mirage in the desert, the closer we think we are to it the more we realise it does not exist, at least not in the form of a standard, simple bifurcating diagram. The aim of this chapter is infer a well-resolved and inclusive prokaryotic tree of life using the new methods characterized above. From Lamarck's (Lamarck, 1809) tree diagram to Darwin's Origin famous tree figure (Darwin, 1859), from the discovery of the double helix to the discovery of the four-nucleotide bases, from the sequencing of the first gene to the sequencing of the first genome it appears that every time we think we have a new tool to understand the evolution of life on the planet (prokaryotic life in particular), we get frustrated as this long held hypothesis, even after much modifications to make it reflect current knowledge, simply does not fit the data that we observe (Bapteste *et al.*, 2009; Doolittle, 1999b; Martin, 1999). This has led to a divide within evolutionary biology community. On one side of the argument are those researchers who have proposed that we throw away the tree of life hypothesis (Gupta, 1998; Lake and Rivera, 1994; Bapteste *et al.*, 2009; McInerney *et al.*, 2011).

The argument against the tree of life hypothesis is mostly based on the emergence of the role played by lateral gene transfer (LGT) in the evolution of prokaryotes (Doolittle, 1999a; Lerat *et al.*, 2003; Bapteste and Boucher, 2008). Many authors argue that the tree of life hypothesis is inadequate as it can only decently represent eukaryotic evolution, i.e. evolution based on mechanisms that follow a

bifurcating pattern (Bapteste *et al.*, 2009), and that even within eukaryotes, its validity might be limited to animals. Hence, the problem with the tree hypothesis starts when used to represent prokaryotic evolution, and perhaps the evolution of the unicellular eukaryotes. Indeed, it has been suggested that the eukaryotic and the prokaryotic mechanisms of evolution are different (McInerney *et al.*, 2008; Puigbò, 2009). As a consequence of the above stated reasons, it has been suggested that the tree of life hypothesis is not a suitable model to represent prokaryotic evolution and hence a new and better fitting model should be employed instead. This has led to an increase in the literature advocating alternative models such as the public goods hypothesis (McInerney *et al.*, 2011) (see also (Dagan and Martin, 2006; Dagan and Martin, 2009; Halary *et al.*, 2010).

On the opposite side of the argument are the researchers who are convinced that a tree of all organisms is still a valid metaphor of life. These advocates of the tree of life hypothesis suggest that there exists a set of core genes that are immune to LGT and that these genes can be used to correctly infer a tree of life of all organism, prokaryotes included (Ciccarelli *et al.*, 2006; Puigbò, 2009; Puigbò *et al.*, 2010). However, according to McInerney and Pisani (2007), these genes do not exit. The advocates for the tree of life hypothesis such as Kurland *et al.* (2003) have responded to this by citing the lack of confidence in the accuracy of the methods used for measuring the rate of LGT i.e. gene tree incongruence (Brochier *et al.*, 2002; Lerat *et al.*, 2003; Bapteste *et al.*, 2004). Indeed, some have argued that, although the level of LGT seen in prokaryotes exceeds anything we see in eukaryotes, the tree hypothesis still offers insights into the vertical inheritance of genes in prokaryotes (we just have to find a way to separate the vertical signal of inheritance from the

horizontal signal of inheritance) or at the very least the tree of prokaryotes can be interpreted as the tree of cell division (Daubin *et al.*, 2003; Creevey *et al.*, 2004; Ciccarelli *et al.*, 2006).

The question remains: are the mechanisms underlying natural variation (such as point mutations, gene acquisition, chromosomal replication, conjugation, transformation etc.), across the prokaryote-eukaryote divide so different that different models must be employed to represent them or is there still a niche for the tree hypothesis in the representation of the genealogical relationships of prokaryotic life? An example is the case of the nitrogen fixing bacteria, *Frankia*, whose strains genomes can differ by as many as 3500 genes despite having an rRNA sequence similarity of up to 97%, this is almost 77% of some of the smaller *Frankia* strains gene repertoire (Bapteste *et al.*, 2009). Hence, we must ask ourselves whether the belief in the existence of a universal tree of life stronger than the data, the phylogenetic signal recovered from their genomes, which supports it?

Several research papers dating back to the first rRNA inferred phylogenetic tree (Woese and Fox, 1977) have attempted to shed more light on this fundamental question. Here, I present a new analysis based on the use of the Bayesian supertree method tested in the previous chapters and the use of the L.U.St package. These softwares will be used to analyse 4 new datasets. These include a data set composed of bacterial genomes only (392 species), a data set composed of archaeal genomes only (51 species), a data set composed of both bacterial and archaeal genomes (443 species), and a data set composed of both prokaryotic and eukaryotic genomes (449 species).

5.2 Methods

5.2.1 Data acquisition

All prokaryotic genomes available from the NCBI database were downloaded. The fully sequenced genomes of 6 eukaryotes were also downloaded. The sequenced genomes of the *Bigelowiella natans*, the *Arabidopsis lyrata*, the *Saccharomyces cerevisiae*, the *Trypanosoma congolense* and the *Dictyostelum purpuren* were downloaded from the joint genome institute (JGI) website. The *Cyanidioschyzon merole's* sequenced genome was downloaded from the ensembl database (see Appendix C – for a list of all taxa used).

5.2.2 Cluster of orthologous proteins

Initially two databases were assembled. The first, called the Prok dataset, was composed of bacterial and archaeal genomes only. The second, the Prok_Euk dataset, was composed of the prokaryotic genomes and the genomes of the eukaryotes mentioned above. For each database, an all-versus-all blast search (this involves blasting each sequence against every other sequence in the database) was set up (e-value = 10e-8) using BLAST 2.2.19. Homologous protein families were then identified using the Markov cluster algorithm (MCL) (van Dongen, 2000). MCL is an effective way to identify protein families based on random flow simulations.

The MCL analysis for the Prok data set returned 386,576 gene families of which 82,844 were composed of 4 or more genes. These 82,844 gene families were composed of 47,725 single gene families (potential orthologs) and 35,119 multi gene families (including both orthologs and paralogs).

Examination of the families led to the conclusion that the MCL granularity parameter (which was set to 1.4) had not been able to cut the sequences into a sufficiently large number of single gene families (with a lot of massive multi-gene families – including many paralogous groups remaining). As a consequence, there was not sufficient taxon overlap to allow for supertree reconstruction. My solution to this problem was to concatenate the multi-gene families in a new database and blast them again using the random blast approach (Creevey *et al.*, 2004). The random blast approach works by choosing a random sequence from the database and blasting it against all other sequences in the database. After that, all the sequences are removed from the database before another sequence is picked and blasted. This allows increasing the granularity and breaking multigene families into groups of orthologs. For the random blast analysis the e-value was set to 10e-16 to ensure that the large multi-gene families would be separated into constituent orthologs. The 35,119-multi gene families from the Prok data set were concatenated and analysed using random blast to give 69,070 new gene families of which 30,103 had four or more genes, and out of this we retrieved 4,734 single gene families (the remaining families were still multigene ones). These 4,734 single gene families where added to the 47,725 single gene families from the MCL analysis to obtain a total of 52,459 single gene families.

For the Prok_Euk dataset, the MCL analysis returned 432,250 gene families of which 88,038 had 4 or more genes. These were further separated into single gene (38,779 potential orthologs) and multi-gene (49,259 potential paralogs and orthologs) families. As in the case of the Prok data set, the multigene families were further split using random blast. The Random blast analysis resulted in 92,432

families, 41,353 of which had four or more genes. Of these 41,353 families, 4,732 were single gene families and were used for further analyses. The combination of the MCL derived and of the Random blast derived single gene families resulted in a total of 43,511 families that could be used for supertree reconstruction.

5.2.3 Building gene trees

To infer gene trees for the supertree analyses, the set of single gene families for the Prok and Pro_Euk datasets were aligned with the multiple sequence alignment software, PRANK (Löytynoja and Goldman, 2008). The multiple sequence alignments were then screened with Gblocks (Castresana, 2000). This software cleans up the multiple sequence alignments by removing poorly aligned positions. All multiple sequence alignments with fewer than 100 amino acids were discarded (too short to allow the generation of reliable phylogenetic trees) and the remaining multiple sequence alignments were checked for the level of phylogenetic signal they convey using the permutation tail probability (ptp) test (Archie, 1989; Faith and Cranston, 1991). Analyses were run at the generic level, this means that I was now focussed on relationships among the different genera rather than the species. The set of multiple sequence alignments that passed the ptp test were used to infer maximum likelihood trees using the RAxMl software (Stamatakis, 2006). In the RAxMl analysis, I used the GTR + Gamma model for alignments longer than 200 amino acids and an empirical LG + Gamma model for alignments shorter than 200 amino acids. When all filtering was completed, I was left with 16,463 gene trees for the Prok dataset and 17,747 gene trees for the Prok_Euk dataset. The Prok data set was then split into

two, to create two more data sets: Bac (composed solely of bacterial-specific genes – for a total 14,558 trees) and Arc (composed solely of archaeal-specific genes – for a total of 1,776 trees).

5.2.4 Supertree analysis

The trees from the Prok, Bac, Arc, and Prok_Euk data sets were used as input trees to Bayesian supertree analyses performed using the Bayesian (MCMC) supertree method discussed in chapter 2. For the Bac and the Prok data sets, the Bayesian supertree was run as follow: 2 parallel chains for 4 million iterations, sampling every 5000th iterations. The β value was set to 1. For the Arc data set, two parallel chains for 2.2 millions iteration were run while sampling every 5 thousand iterations, with the β value set to 1. For the Prok_Euk dataset, two parallel chains of 5 millions iterations were run sampling at every 5000th iterations, with the β value set to 1.

5.2.5 Testing previously proposed positions of the Eukaryotes

Four main hypotheses have been put forward to represent the tree of life. Trees representing these hypotheses were constructed using the Mesquite software (Maddison and Maddison, 2001). These hypotheses differ in their placement of the eukaryotes in the tree of life. The first hypothesis places the eukaryotes as sister group to the Creanarchaeaotes representing the hypothesis that the Crenarchaea are the closest relative of the eukaryotes (the eocyte hypothesis (Lake, 1988; Rivera and Lake, 1992; Cox *et al.*, 2008). The second hypothesis places the eukaryotes as a sister group to the Archaebacteria clade as inferred in the rRNA tree of life (widely known as the 3-domains of life hypothesis; (Woese *et al.*, 1990)). The third

hypothesis places the eukaryotes as a sister group to the Cyanobacteria (this accounts for eukaryotic genes of Cyanobacteria origin due to endosymbiotic gene transfer from the plastid to the plant nucleus (Gray, 1989; Martin *et al.*, 2002; Rivera and Lake, 2004)). The fourth hypothesis places the eukaryotes as the sister group of the Alpha-proteobacteria (this accounts for eukaryotic genes of alpha-proteobacteria origin due to endosymbiotic gene transfer from the mitochondria to the eukaryote nucleus (Martin and Müller, 1998; Andersson *et al.*, 1998). These hypotheses were analysed using the test of two trees (described in chapter 2) to see whether our data could reject some of them.

5.2.6 Identification of rogue taxa

The Concatabomination method (Siu-Ting *et al.*, Submitted) was used to identify unstable taxa. This method is a heuristic extension to the safe taxonomic reduction (STR) method of Wilkinson (1995), which uses the character information and distribution of missing data in a Baum-Ragan encoded matrix to classify taxa into taxonomic equivalents. The Concatabomination method uses a compatibility approach to test whether, if two taxa are artificially hybridised (i.e. concatabominated), the homoplasy in the matrix increases. If it does not, then the taxa are equivalent and one of them can be eliminated from the analyses. The method returns a ranked list of rogue taxa, which can be visualised as a network using Cytoscape (Shannon *et al.*, 2003).

5.3 Results

5.3.1 The Prokaryote Supertree

The two parallel chains that were set up for the Bayesian (MCMC) supertree analysis of the Prok data set converged after 2.5 million iterations. After removing the burn-in a total of 482 were left from both chains. These supertrees were read into the PAUP4b10 software package (Swofford, 2003) to construct a majority rule consensus tree (figure 5.1). The resolution of this tree is very poor, and clades with a posterior probability that is less than 0.5 are indicated by dotted lines. Of the 30 prokaryotic phyla represented in this tree by more than one genus only *Deferribacteres, Deinococcus/Thermus, Epsilon-Proteobactera, Chlorobi, Fusobacteria, Plantomycetes, Thaumarchaeota* and *Thermotogae* appear monophyletic.

The tree in figure 5.1 shows high resolution toward the tips, some with low posterior probability, however the deeper we move along the tree the poorer the resolution become. This result is not dissimilar from that found by Creevey *et al.* (2004) and Pisani *et al.* (2007). Because lack of resolution could be caused by the presence of rogue taxa, a concatabomination analysis was performed, and rogue taxa were identified and removed (see figure 5.2). The concatabomination analysis identified 16 rogue taxa. To investigate their effect on supertree topology, the 16 rogue taxa were pruned from the 482 recovered supertrees, and a new majority rule tree was calculated (to evaluate whether removing taxa affected the support for the clades in the supertree), (see figure 5.3). The removal of the 16 rogue taxa did little to resolve the tree, especially with reference to the deep branches. Although this tree showed a slight improvement in the posterior probabilities of many nodes and would have recovered an extra monophyletic phylum: the *Beta-Proteobacteria*, had

the *Beta-Proteobacteria* taxon (*Neisseria*) not been inferred within the *Gamma-Proteobacteria*. The fact is that this new tree does not represent a great improvement toward recovering a resolved prokaryotic tree of life. This observation made me question if the MCMC method was unable to recover a tree for these taxa, proving that the method is not powerful enough to deal with a dataset as challenging as this, or whether a resolved prokaryotic tree is not recovered because it does not really exist.

In an attempt to vindicate the Bayesian tree inference method, 100 supertrees were randomly generated on the same taxa as the Prok dataset using PAUP4b10 (Swofford, 2003). The mesquite software (Maddison and Maddison, 2001) was used to manipulate the Bayesian (MCMC) topology to mirror the topology presented by Ciccarelli *et al.* (2006), which we use as the standard "accepted" tree of life. Log likelihood values were calculated for the 100 random supertrees, for the Bayesian (MCMC) supertree and for the Ciccarelli topology using the L.U.St package. The software Tracer (Rambaut and Drummond, 2007) was used to show the distribution of the calculated likelihood values and the result is presented in figure 5.4. This figure shows that both the Bayesian (MCMC) and Ciccarelli topology are significantly better than random. In addition, when compared using the AU test the Bayesian topology appears to have a fit to the data that is significantly higher than that of the Ciccarelli tree. With this as evidence I can confidently rule out methodological errors or inefficiency as a cause of the lack of resolution observed in the supertrees inferred for the Prok data set. In truth, the apparently nonsensical Bayesian tree fits the data much better than a standard tree of life. Hence this data set could not have inferred a tree similar to the Ciccarelli one and we have to

conclude that the Prok data set, very simply, does not support the existence of a tree of life.

It is evident from figure 5.1 and figure 5.3 that a key topological feature of the Prok tree(s) is that the Archaebacteria appear to be substantially fragmented. For example, the Crenarchaeota and the Thaumarchaeota nest with the Alpha-Proteobacteria, while the Haloarchaea are nested within the Gamma-Proteobacteria. Finally, the methanogenic Euryarchaeaota are shown to branch within a group composed of Beta-Proteobacteria and Gamma-Proteobacteria. Importantly, all of these relationships have a low posterior probability (less than 0.5). Certainly, such values are not strong enough to suggest that an alternative to the Ciccarelli tree should be proposed. Rather, it seems that these results might be speaking against the existence of a prokaryotic tree of life.

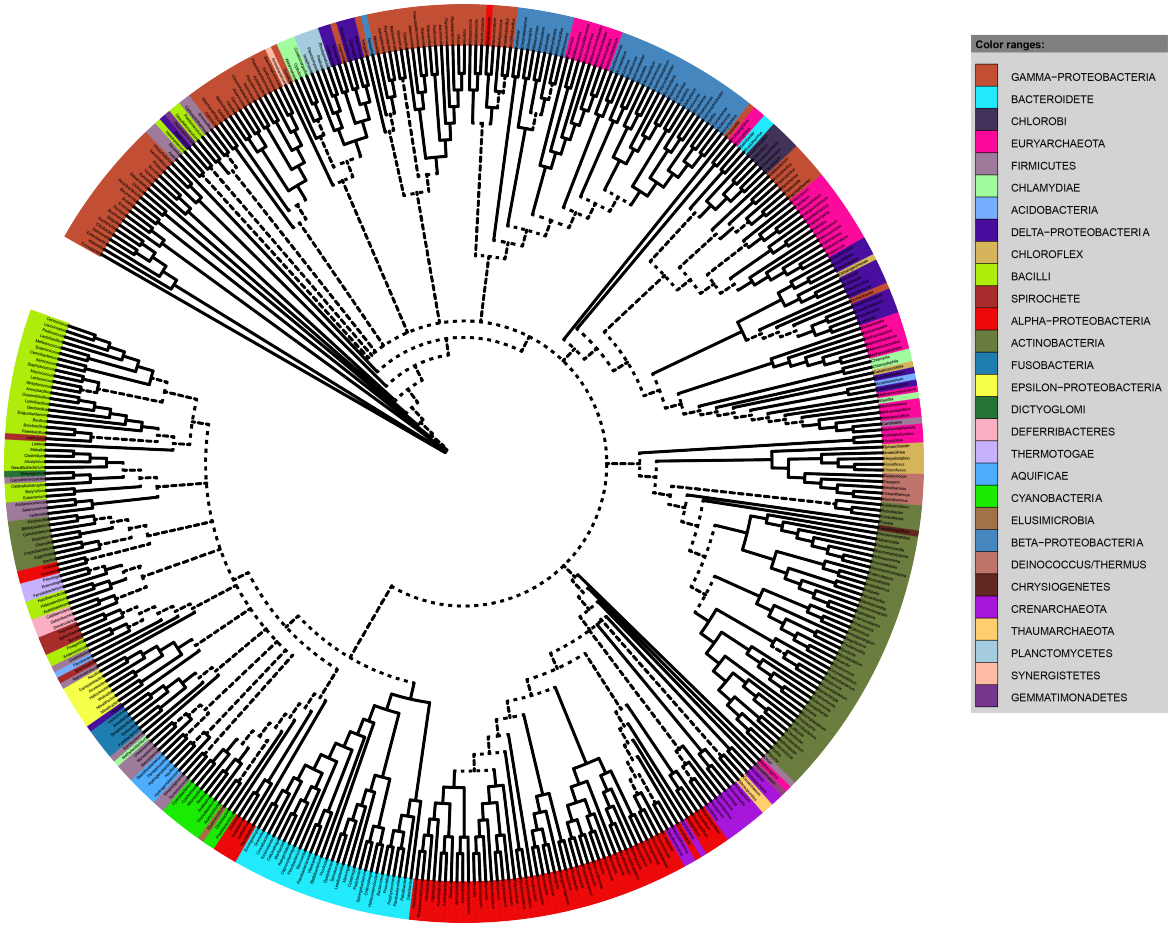


Figure 5.1: Bayesian (MCMC) phylogeny of the prokaryotes.

The majority rule consensus tree constructed from the 483 supertrees sampled from the MCMC chains (2 runs). Dotted lines represent clades with less than 0.5 posterior probabilities. **Note:** This phylogeny should be interpreted as unrooted. It is presented as a circular cladogram only to fit the page.

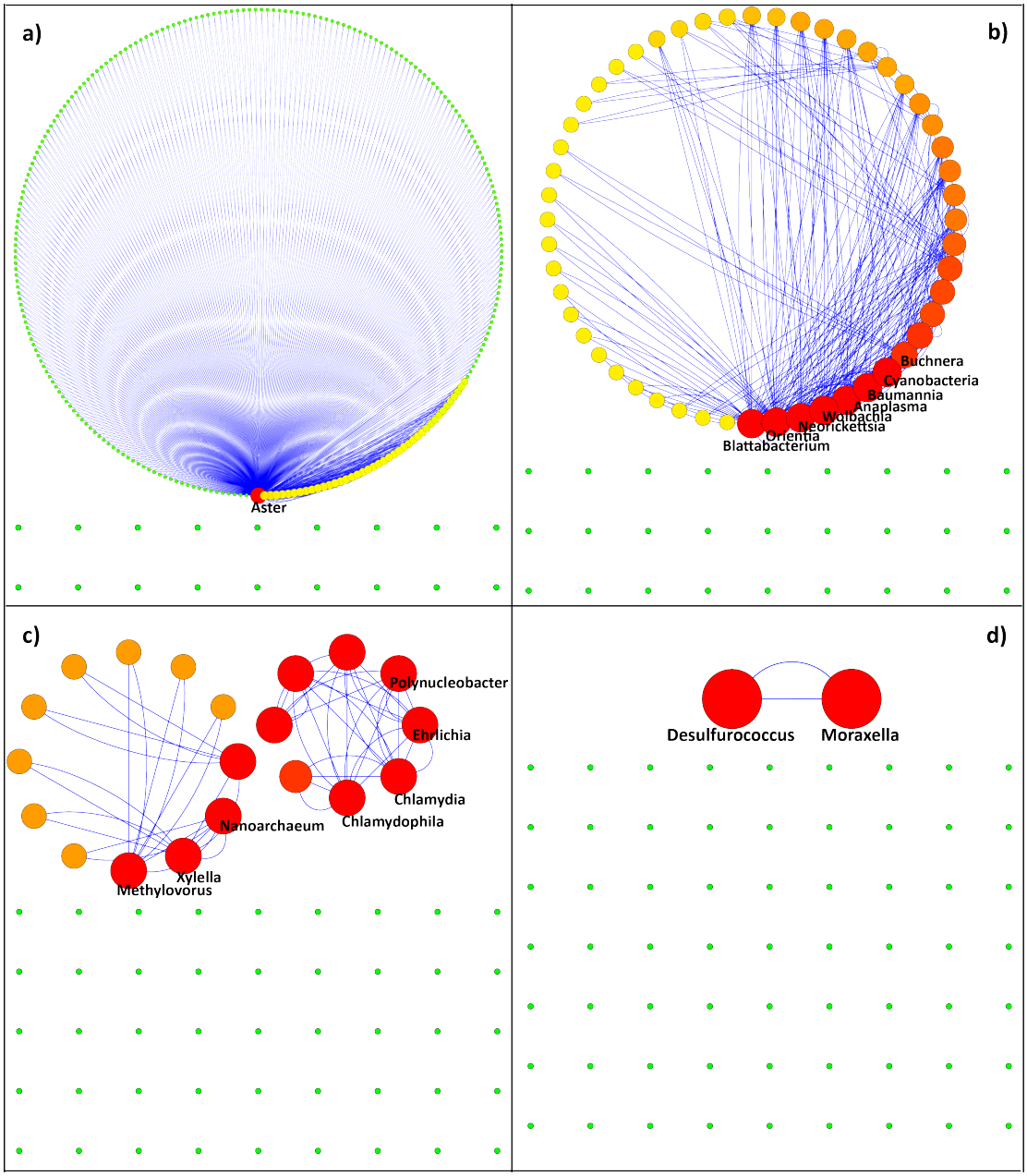


Figure 5.2: Network visualisation of taxonomic equivalents in the Prok dataset.

The green dots represent taxa that are not taxonomically equivalent to any other taxa in the dataset while the networks represent taxa that share the same information (taxonomically equivalent taxa). The highly unstable taxa are coloured in red. a) The full network-indicating *Aster* as the most unstable taxa. b) The network after *Aster* has been deleted. It also shows *Blattabacterium, Orientia, Neorickettsia, Wolbachia, Anaplasma, Baumannia, Cyanobacterium and Buchnera* as unstable taxa. c) The network following the deletion of the highly unstable taxa identified in b). At this point *Methylovorus, Chlamydophila, Chlamydia, Ehrlichia, Xylella, Nanoarchaeum and Polynucleobacter* are identified as the next group of highly unstable taxa. d) The reanalysed network following the deletion of the highly unstable taxa identified in c).

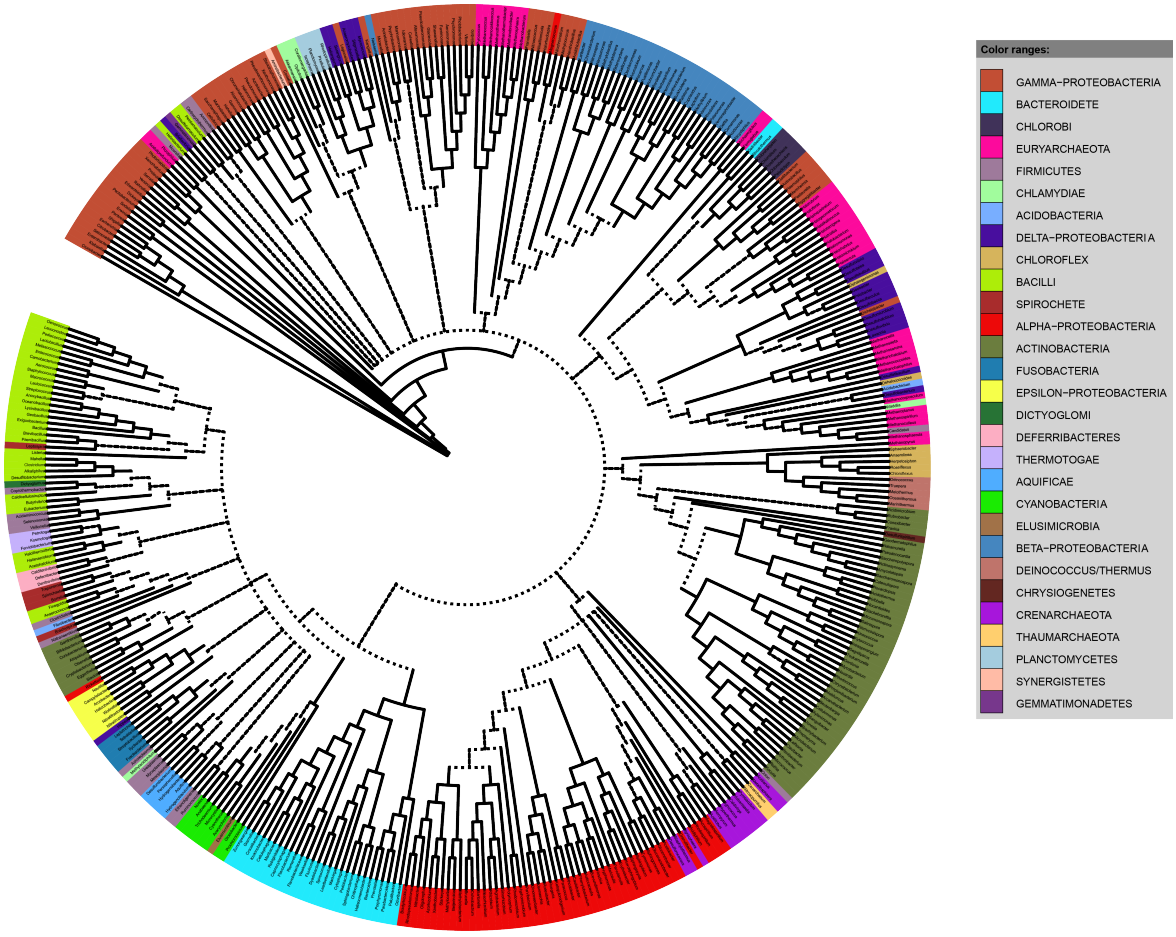


Figure 5.3: Bayesian (MCMC) phylogeny of the prokaryotes after pruning the rogue taxa. This is the majority rule tree constructed from the 483 trees sampled from the 2 runs after the 16 rogue taxa identified by the concatabomination analysis were eliminated. Dotted lines represent clades with less than 0.5 posterior probabilities. **Note:** This phylogeny should be interpreted as unrooted. It is presented as a circular cladogram only so that it could fit the page.

5.3.2 The Archaeabacterial Supertree

In light of the results from the analyses of the Prok data set, one wonders whether the Archaebacteria truly are monophyletic. A variety of recent studies have addressed the phylogenetic relationships within this domain, and an exceptional level of resolution has indeed been obtained (Brochier-Armanet *et al.*, 2008; Brochier-Armanet *et al.*, 2011). How can these results compare with those presented here (Prok analyses)? To elucidate this problem I analysed the Arc dataset to evaluate whether the signal for the relationships found by previous studies of the Archaebacteria is present also in Prok (despite the odd relationships obtained from when the Eubacteria are also included). Figure 5.4 shows the Bayesian (MCMC) supertree recovered for the Arc data set (the supertree was obtained summarising 800 trees found from the two runs after convergence). The input trees used to recover this supertree were generated from those in Prok, simply deleting all eubacterial genera. Surprisingly, the Arc supertree almost perfectly reflect current understanding of Archaea evolution (Wolf *et al.*, 2002; Ciccarelli *et al.*, 2006; Brochier-Armanet *et al.*, 2008; Brochier-Armanet *et al.*, 2011). It shows that the Haloarchaea branches from with the methanogens, and Crenarchaeota can be seen as the sister group of the Thaumarchaeota. In addition to having a topology comparable with that of other (previous) archaebacterial phylogenies, the Arc supertree is also extremely well supported (compare with the phylogenies of (Gribaldo and Brochier, 2009; Kelly *et al.*, 2011; Brochier-Armanet *et al.*, 2011), see figure 5.4). The dissimilarity between the Prok and the Arc tree is astonishing and implies that the Prok data set conveys the information generally represented in standard archaebacterial phylogenies. The question, therefore, is why isn't this

information appearing in the Prok tree? We shall address this question in the discussion.

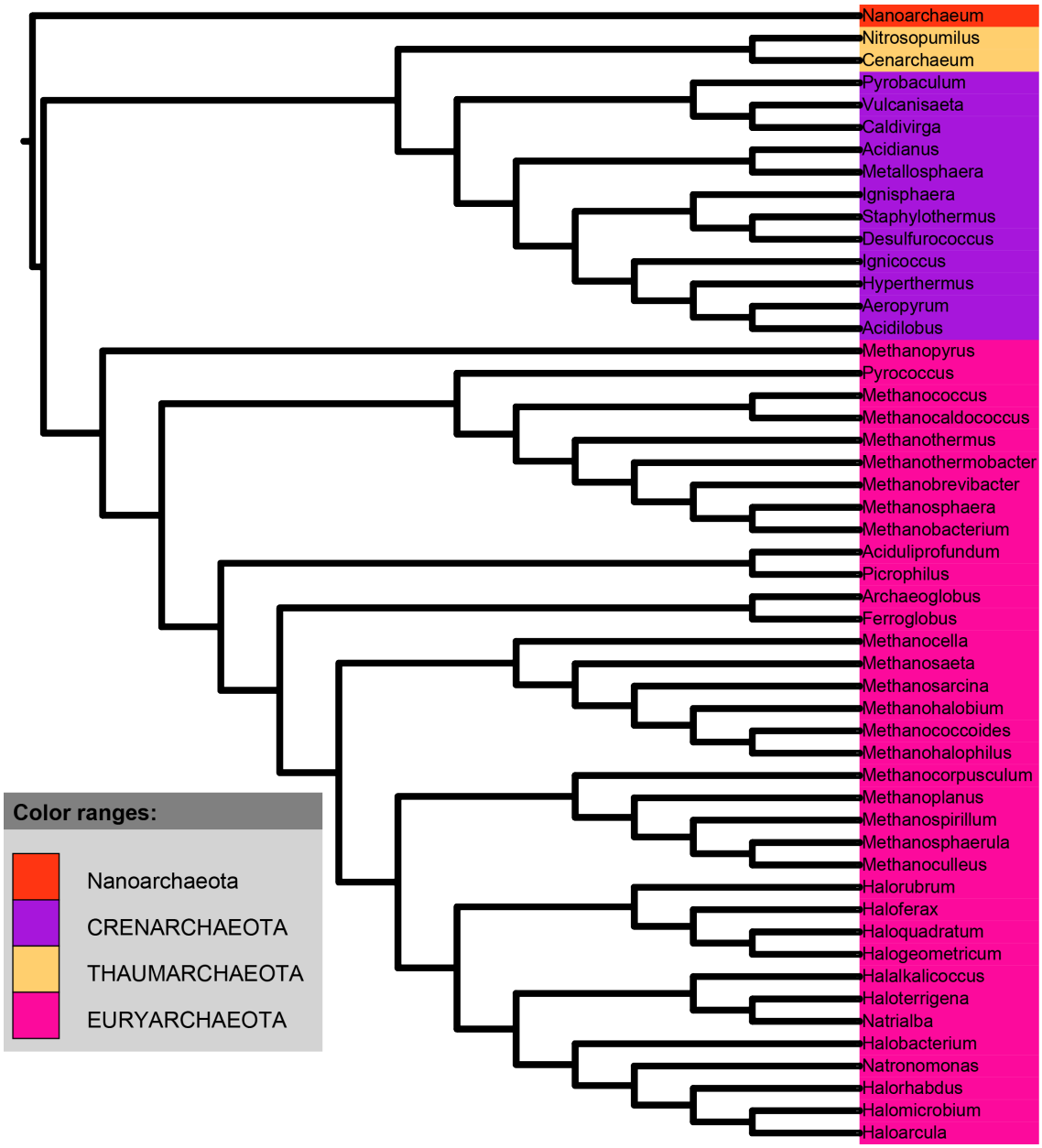


Figure 5.4: Rooted Bayesian (MCMC) phylogeny of the Archaeabacteria.

This is the majority rule tree constructed from the 800 supertrees sampled from two MCMC chains in the Bayesian supertree analysis.

5.3.3 The Eubacteria Supertree

The Bac supertree was generated by combining the 700 trees sampled (after convergence and discarding burn-in) from the two MCMC chains (figure 5.5). This tree is the obvious counterpart of the Arc tree (derived from Prok by deleting all archaebacterial lineages). This topology is an improvement on the topology recovered (for the Eubacteria clades) in the analyses of Prok (figure 5.1 and 5.3), where the Eubacteria and Archaebacteria were concomitantly included. Ultimately, also the tree of figure 5.5 failed to recover many eubacteria *phyla*. However, this tree is clearly much better than that derived from the Prok dataset.

Concatabomination analysis identified 5 excludable rogue taxa in Bac (see figure 5.6). Exclusion of these taxa did not improve the resolution of the Bac tree figure 5.7.

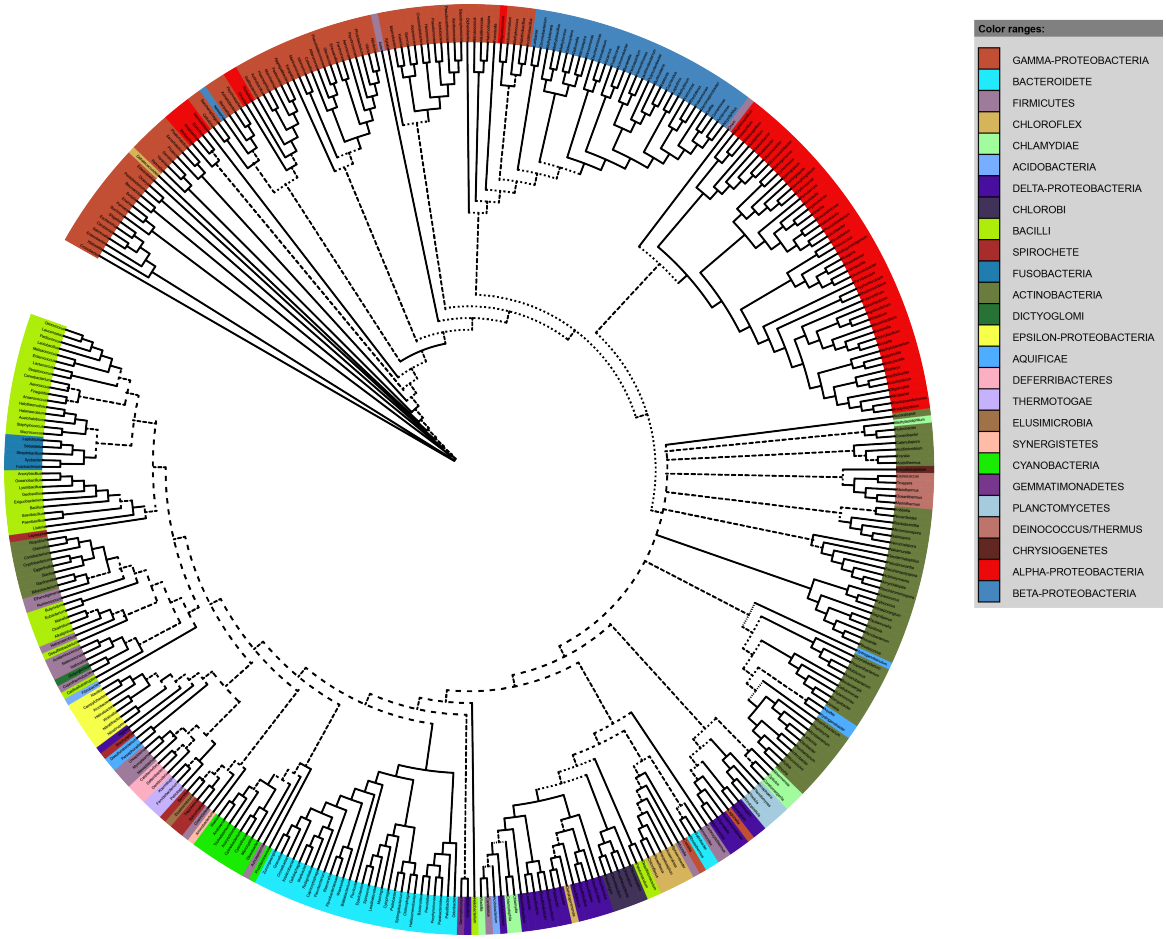


Figure 5.5: Bayesian (MCMC) phylogeny of the Eubacteria.

This is the majority rule consensus tree constructed from the 700 supertrees sampled from 2 independent Bayesian analyses after convergence. Dotted lines represent clades with less than 0.5 posterior probability. **Note:** This phylogeny should be interpreted as unrooted. It is presented as a circular cladogram only to fit the page

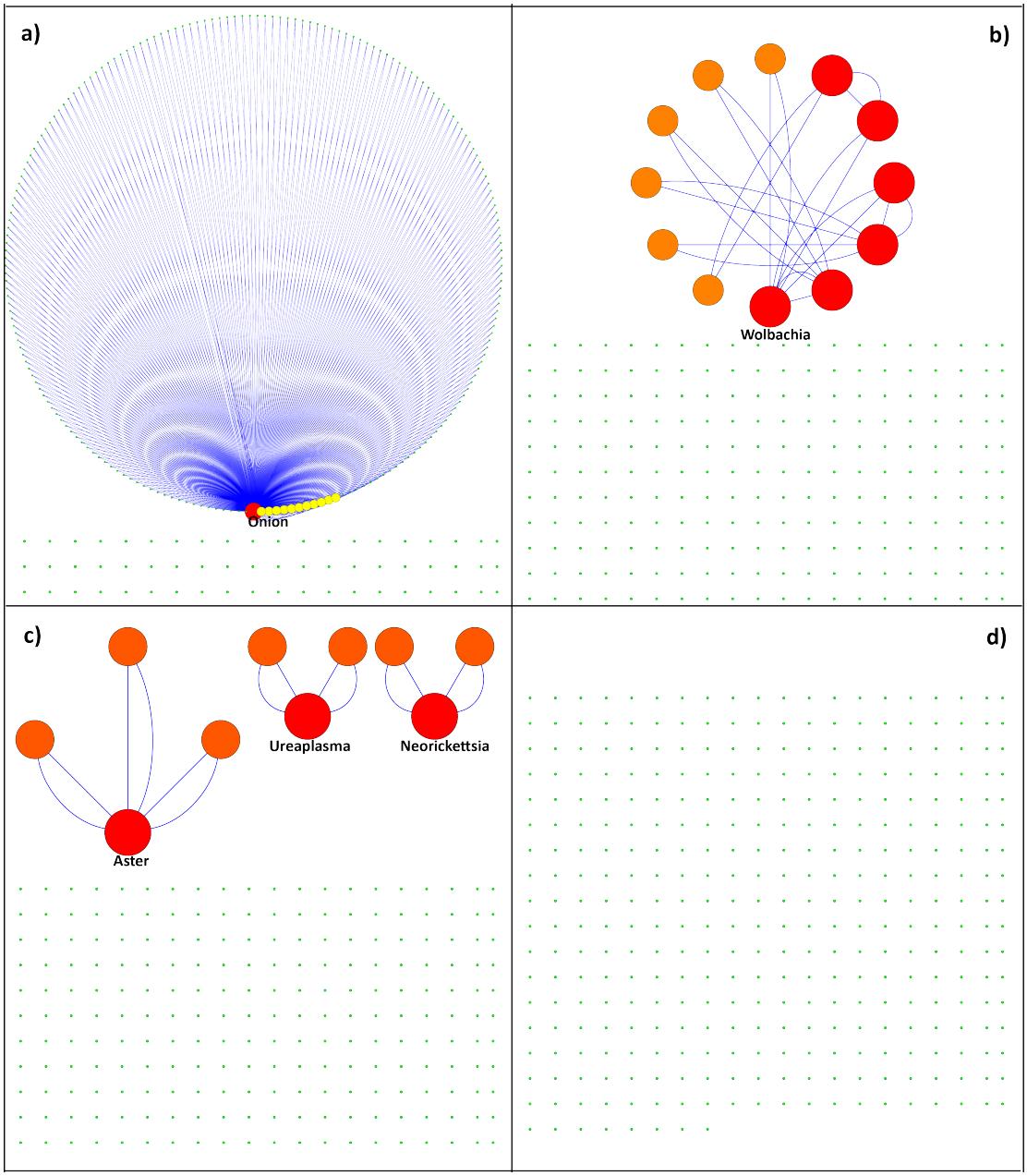


Figure 6.6: Network visualisation of taxonomic equivalents in the Bac dataset.

The green dots represent taxa that are not taxonomically equivalent to any other taxa in the dataset while the networks represent taxa are share the same information (taxonomically equivalent taxa). The highly unstable taxa are coloured in red. a) The full network indicating Onion as the most unstable taxa. b) The reanalysed network following the deletion of the Onion node. The Wolbachia node is now shown as the next highly unstable taxa in the dataset. c) The reanalysed network following the deletion of the Wolbachia node. Aster, Ureaplasma, and Neorickettsia are now shown as the next highly unstable taxa in the dataset. d) The reanalysed network following the deletion of the nodes identified as highly unstable in c). The network when all the unstable taxa in the tree have been removed.

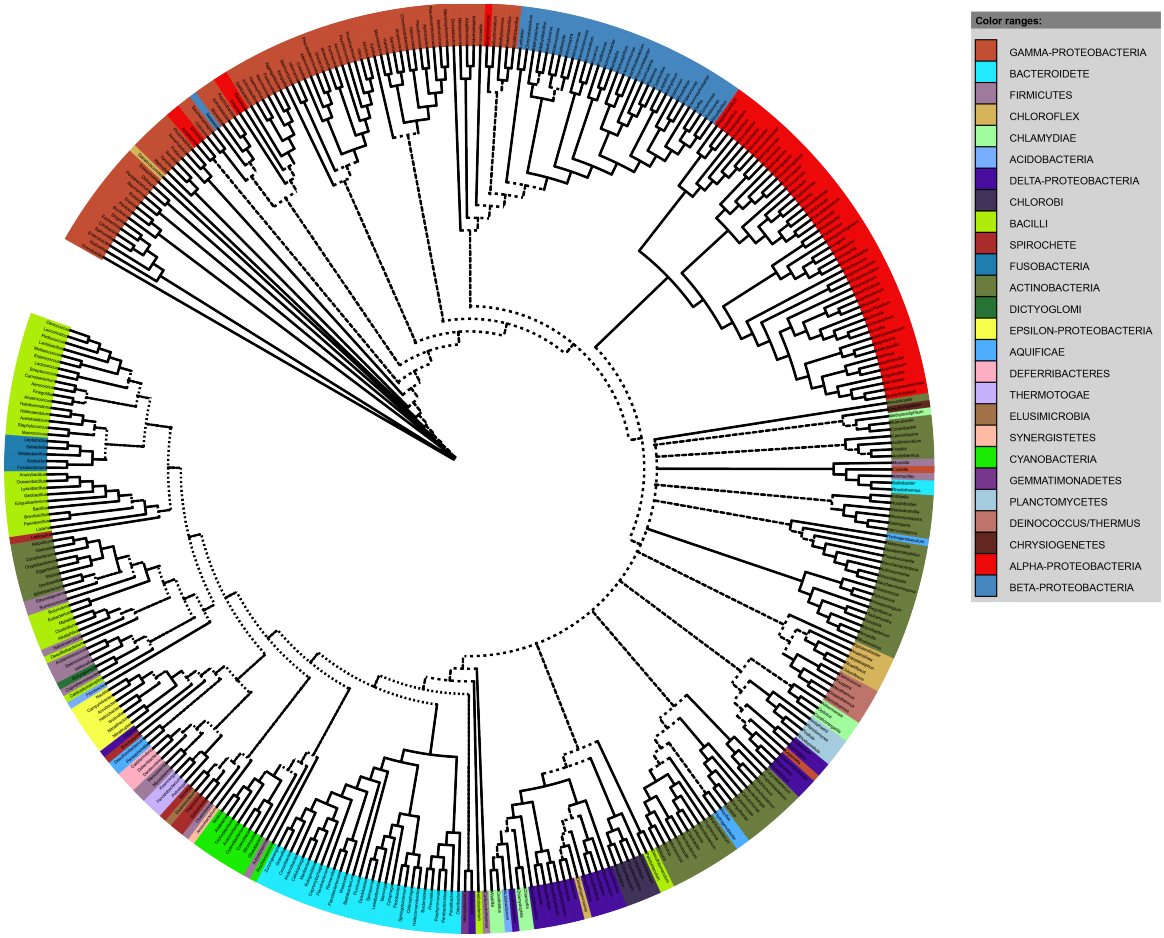


Figure 5.7: Bayesian (MCMC) phylogeny of the Eubacteria with the rogue taxa pruned. This is the majority rule consensus tree constructed from the 700 supertrees sampled from 2 independent Bayesian analyses after convergence and after the 16 rogue taxa identified by the concatabomination analysis were pruned away. Dotted lines represent clades with less than 0.5 posterior probability. Note: This phylogeny should be interpreted as unrooted. It is presented as a circular cladogram only to fit the page.

5.3.4 The position of the Eukaryotes

The Statistical test of two trees for the four "hypotheses of life" is presented in Table 5.1. The result shows that the hypothesis placing the eukaryotes in a sister group relationship with the Alpha-Proteobacteria is the only hypothesis that can be confidently rejected out of the four hypotheses. This was to be expected and is in accordance with what was showed by Pisani *et al.* (2007) who showed that the strongest signal for the outgroup of the eukaryotes is with the plant (this is unsurprising as the plastid acquisition was the latest of the symbiotic events characterising the origin of the eukaryotes and many genes of plastid origin in the plant genomes show a strong link with the Cyanobacteria).

Supertree methods	AU test	SH test	KH test
Eukaryote and Archaea as sister group	0.740	0.929	0.710
Eukaryote and Cyanobacteria as sister group	0.299	0.535	0.277
Eukaryote and Crenarchaea as sister group	0.246	0.658	0.290
Eukaryote and Alpha-proteobacteria as sister group	0.001	0.007	0.005

Table 5.1: Summary of the statistical test of the Eukaryotic relationships. This table illustrates the probability values of the test of two or more trees for supertrees (implemented in L.U.St) for the phylogenies inferred for the carnivores. The row coloured red is rejected by the AU test. Legend: AU – Approximately Unbiased, SH – Shimodaira-Hasegawa, KH – Kishino-Hasegawa.

5.4 Discussion

The idea of using trees to model evolutionary relationships, made popular by the work of Haeckel (1866), is currently facing its biggest challenge in the 21st century. This is due to new evidence showing that lateral gene transfer might be at the heart of the mechanism governing prokaryotic evolution.

The collective topologies inferred for the Prok, Bac, Arc and Prok_Euk datasets presented here provide no support for the existence of a tree of life, in particular a prokaryotic tree of life. AU tests could not distinguish between alternative hypotheses of eukaryotic relationships, while more importantly; the Prok analysis statistically rejected the standard tree of life (i.e. the Ciccarelli tree) favouring instead a topology with little biological sense (if read by assuming that the tree of life must exist). In particular this topology showed no support for the monophyly of the Archaebacteria that are distributed across the Eubacteria. This is surprising and quite shocking, if one analysed the Archaebacteria only (to the exclusion of the Eubacteria) and is able to find the traditional Archaebacteria tree. This suggests that Archaebacterial genomes are significantly enriched in eubacterial genes, and that different archaebacterial lineages have acquired genes from alternative eubacterial lineages. At the same time it seems that there is probably not much LGT going on within Archaebacteria (or that these LGT are totally randomised in direction so that the phylogenomic tree here derived is in agreement with the results of the standard rRNA tree). Overall these results are surprising and provide a very interesting insight into prokaryotic evolution. More broadly the results of the BAC data set illustrate that apart from sharing with the Archaebacteria, the Eubacteria are also much more promiscuous among themselves, to the point that a clear eubacterial phylogeny is

not recoverable from the data. Overall, we can only conclude that there is no clear evidence in genomic data for the existence of a prokaryotic tree of life.

Apart from further elucidating patterns of prokaryotic evolution, the results presented in this chapter illustrate the potential of the methods presented in this thesis.

5.5 Conclusion

While several authors have called for completely new models to be used to represent prokaryotic evolution, and some other authors have dogmatically insisted on the continuation of using the tree hypothesis as it is (Daubin *et al.*, 2003; Ciccarelli *et al.*, 2006), others have proposed using a modification of the tree hypothesis i.e. a tree showing vertical evolution with the LGT events mapped on top of it or a different interpretation of the tree of life as a tree of cell division (TOCD).

In the introduction I asked the question: is there still a place for a tree in prokaryotic evolution? The answer based on the result of this study is an overwhelming no. This is because; although there is some vertical signal visible in the prokaryotic tree this is mostly toward the tip of the tree. Deep evolutionary events are simply unresolvable based on entire genomes, not because of signal erosion, but because of the rampant role of LGT in prokaryotic evolution. Simply stated, a tree is not a good metaphor to represent the evolution of these organisms. In particular it seems that the eubacteria are particularly promiscuous among each other, while it seems that Archaebacteria are enriched in Eubacterial genes. Our results seem to suggest that while there might be patterns of transfer from

eubacteria to Archaebacteria (consistent with what was showed by Nelson-Sathi *et al.* (2012)), LGT within Archaebacteria might be quite randomly distributed (or rare). To the point that there is no strong signal (if one considers Archaebacteria only) that cancels the signal consistent with the classic Archaebacterial tree. That is, it is possible that in Archaebactera interdomain transfers were more important than intradomain ones. Given that Archaebacteria clearly engage in LGT, it seems more likely that these tend to be random and do not have a strong directional effect that could cancel out the vertical signal representing the pattern of cell division within this lineage.

These results might seem depressing (we have all learned about the tree of life and we might not necessarily have expected to see it falling apart). However, we should not be worried because even if the tree of life might ultimately be falling, evolution is real, and a new model will ultimately be described that fits the data better than a tree. From my personal point of view I can say that I am not depressed. The results obtained in this last chapter convinced me that the tools I have been developing (either in isolation or in collaboration with Peter Foster) can be extremely useful in evolutionary biology and can be used to gain new insight in the study of evolution. What better way to conclude my PhD?

Chapter 6: General Discussion & Conclusions

> "In science it often happens that scientists say, 'You know that's a really good argument; my position is mistaken,' and then they would actually change their minds and you never hear that old view from them again. They really do it. It doesn't happen as often as it should, because scientists are human and change is sometimes painful. But it happens every day. I cannot recall the last time something like that happened in politics or religion." – Carl Sagan

This thesis addresses a topic, tracing the evolutionary history of extant species from a single common ancestor, both of its constituent points of view (theoretical and applied). The search for common ancestors and relationships of relatedness has captivated researchers in the field of evolutionary biology, conservation, epidermiology etc. since the 19th century, and for some aspects this field has not changed much since then. This is evident in the fact that we are still relying on Darwin's idea that a tree of all organisms can be derived.

However, this monotony of the tree hypothesis for the representation of the evolution of life on the planet might be coming to an end. As mentioned in chapter 5 of this thesis the applicability of the tree hypothesis to the representation of evolution in prokaryotes in particular has began to be questioned (Bapteste *et al.*, 2009; McInerney *et al.*, 2011). This thesis first developed and tested tools to investigate relationships of common ancestry and then addressed this question using these new tools and genomic scale data sets. Based on the results obtained (Chapter 5) my work confirms that the hour of change is finally before us and new and better fitting models of evolution must now be developed, characterized and applied to represent evolution in the prokaryotes. However, while the tree hypothesis is no longer applicable to prokaryotic evolution (and life), there is no

doubt that trees are still of great utility (for example in the study of animal evolution). Hence the various tools developed and characterised in this thesis will continue to be of great utility in any field in which the trees represent suitable hypotheses.

From humble beginnings in the field of computer science the popularity of supertrees have soared among researchers due to their meta-analytical and combinatorial properties. Accordingly, they found application in many diverse research fields i.e. Phylogenomics, comparative biology, Taxonomy, evolutionary developments, etc. However this apparent meteoric rise in the success of supertrees does not undermine the various shortcomings of current methods.

A justified criticism of the majority for available supertree methods was that they were guilty of not treating input tree as estimates from data, treating them rather as factual statements, and hence not accounting for the uncertainties in these estimates (Cotton and Wilkinson, 2009). In the methods implemented in this thesis this problem is formally addressed by modelling error explicitly. To completely eliminate the problem pinpointed by Cotton and Wilkinson, however, one should probably combine the methods in this thesis with the use of bootstrap trees (for different data sets) as inputs instead than optimal trees. However, this is difficult because by using bootstrap input trees the calculation of supertrees become computationally much more expensive.

Secondly many researchers have questioned the black box nature of currently available supertree methods (Pisani and Wilkinson, 2002; Wilkinson *et al.*, 2005a; Wilkinson *et al.*, 2007; Ren *et al.*, 2009). This is a critic mostly levelled at the Matrix Representation with Parsimony method, mainly due to it being the most used

supertree method. Although MRP's suitability for the reconstruction of phylogenies has been discussed from the time of its inception (Steel, 1992; Baum and Ragan, 1993; Rodrigo, 1996), as mentioned in chapter 2 it is only recently that the mechanics of this and other supertree methods have been investigated (Eulenstein *et al.*, 2004; Wilkinson *et al.*, 2004; Cotton *et al.*, 2006; Wilkinson *et al.*, 2007). Based on the literature (Purvis, 1995a; Wilkinson *et al.*, 2005a; Steel and Rodrigo, 2008) and the result of chapter 3, currently available supertree methods are found wanting for some desired properties. This thesis has addressed this problem of lack of clarity regarding the properties underlying the set of available supertree methods by providing two probabilistic supertree methods with well-formulated theories that are consistent under general statistical conditions, see - (Steel and Rodrigo, 2008; Bryant and Steel, 2009).

Another major criticism of current supertree methods was that it was difficult to estimate support for nodes in these trees. This thesis provides a solution to this problem through the use of the parametric alternative, the posterior probabilities. The Bayesian (MCMC) supertree method characterised in this thesis enables the use of posterior probabilities to provide easy to interpret support values for the relationships (clades) represented in the supertrees.

Lastly some researchers have campaigned for the use of the supermatrix approach over the supertree approach as it is expected that supermatrix approaches used more of the information in the character data than supertrees (Kluge, 1989; Gatesy *et al.*, 2004; de Queiroz and Gatesy, 2007). The supermatrix approach focuses on combining data at the ground level by concatenating gene alignments to generate super alignments that can then be analysed using different phylogenetic methods

(i.e. parsimony or likelihood methods). Researchers that favour the supermatrix approach have pointed towards the fact that the supermatrix approach deals with the character data directly as evidence of it superiority, labelling it a total evidence approach and pointing out that it is able to use the hidden support in the character data (de Queiroz and Gatesy, 2007; Gatesy *et al.*, 1999). However, the supermatrix approach often implicitly assumes that all characters have undergone the same evolutionary process (at the least for some of the parameters in the substitution models used). In addition, its ability to find support for clades in the presence of missing data is unclear, and its efficiency in terms of speed is not great (Degnan and Rosenberg, 2006; Ren *et al.*, 2009; Von Haeseler, 2012). Although new methods claim to address some of these deficiencies of the supermatrix approach (Simmons and Freudenstein, 2002; Nylander *et al.*, 2004), these methods are yet to be properly characterized. The ability to use statistical analysis in the supermatrix framework and its absence in the supertree framework has been used by the proponent of the supermatrix approach as another major reason why the supermatrix approach is superior to the supertree approach (Kupczok *et al.*, 2010). However this is now a mute point as the ability to estimate the likelihood of supertrees has paved the path for the use of statistical methods such as the KH test (Kishino and Hasegawa, 1989), the SH test (Shimodaira and Hasegawa, 1999), the AU test (Shimodaira, 2002), etc. (see section 2.2.3).

Though supertrees have their limitations (Steel and Böcker, 2000), the same can be said for supermatrix approaches, hence, I join Von Haeseler (2012) in proposing that the best approach is to use both approaches and compare the results as they offer different strengths. As a last a word I'd like to say the field of

supertrees has progressed and improved very rapidly, the methods developed in this thesis are further testament to efforts to continue this improvement and based on the result of this thesis supertrees can now be applied to an even wider range of research topics.

Chapter 7: Future prospective

The development of accurate methods to reconstruct the evolutionary relationships of organisms continues to be a topic of interest among researchers and the availability of genomic data has given birth to the field of phylogenomics. Although methods presented in this thesis represent an improvement over previous supertree methods, there are still rooms for further improvements. The Robinson Foulds (RF) metric used in this thesis to calculate the distance between trees represents a quick, easy to implement and well-understood distance metric, however other distance metrics that are finer grained exist. An interesting alternative is the quartet distance metric (Estabrook *et al.*, 1985). This is characterised by the number of topological differences in the quartet sets (set of subtrees of four leaves) of two trees. This metric could offer a number of attractive advantages over the RF distances, most importantly that it can estimate with greater precision tree to tree distances (Steel and Penny, 1993). I would like to implement the ML supertree method in the future using quartet distances to measure the difference between trees.

Another possible future endeavour would be to investigate the use of alternatives to the exponential distribution, to model incongruence in the observed data.

Finally I would like to continue to improve the efficiency of the ML program (by improving tree search strategies and recoding it in C) and apply it to other biological questions of relevance.

Chapter 8: Bibliography

ADACHI, J. & HASEGAWA, M. 1992. *MOLPHY, programs for molecular phylogenetics, I: PROTML, maximum likelihood inference of protein phylogeny*, Institute of Statistical Mathematics Tokyo.

ADACHI, J. & HASEGAWA, M. 1995. Phylogeny of whales: dependence of the inference on species sampling. *Molecular biology and evolution,* 12, 177-179.

ADAMS, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic biology,* 21, 390-397.

ADAMS, E. N. 1986. N-trees as nestings: complexity, similarity, and consensus. *Journal of classification,* 3, 299-317.

AGUINALDO, A. M. A., TURBEVILLE, J. M., LINFORD, L. S., RIVERA, M. C., GAREY, J. R., RAFF, R. A. & LAKE, J. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature,* 387, 489-493.

AKANNI, W. A., CREEVEY, C. J., WILKINSON, M., FOSTER, P. G. & PISANI, D. In Prep. Parametric supertrees: Using and Testing Maximum Likelihood in Supertree Reconstruction.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215, 403-410.

ANDERSSON, S. G., ZOMORODIPOUR, A., ANDERSSON, J. O., SICHERITZ-PONTÉN, T., ALSMARK, U. C. M., PODOWSKI, R. M., NÄSLUND, A. K., ERIKSSON, A.-S., WINKLER, H. H. & KURLAND, C. G. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature,* 396, 133-140.

ANSORGE, W. 2009. Next-generation DNA sequencing techniques. *New biotechnology,* 25, 195.

ARCHIE, J. W. 1989. A randomization test for phylogenetic information in systematic data. *Systematic biology,* 38, 239-252.

BAKER, W. J., SAVOLAINEN, V., ASMUSSEN-LANGE, C. B., CHASE, M. W., DRANSFIELD, J., FOREST, F., HARLEY, M. M., UHL, N. W. & WILKINSON, M. 2009. Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. *Systematic Biology,* 58, 240-256.

BALDAUF, S. L., ROGER, A., WENK-SIEFERT, I. & DOOLITTLE, W. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *science,* 290, 972-977.

BANDELT, H.-J. & DRESS, A. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in applied mathematics,* 7, 309-343.

BANSAL, M. S., BURLEIGH, J. G., EULENSTEIN, O. & FERNÁNDEZ-BACA, D. 2010. Robinson-foulds supertrees. *Algorithms for Molecular Biology,* 5, 18.

BAPTESTE, E. & BOUCHER, Y. 2008. Lateral gene transfer challenges principles of microbial systematics. *Trends in microbiology,* 16, 200-207.

BAPTESTE, E., BOUCHER, Y., LEIGH, J. & DOOLITTLE, W. F. 2004. Phylogenetic reconstruction and lateral gene transfer. *Trends in microbiology,* 12, 406-411.

BAPTESTE, E., O'MALLEY, M. A., BEIKO, R. G., ERESHEFSKY, M., GOGARTEN, J. P., FRANKLIN-HALL, L., LAPOINTE, F.-J., DUPRÉ, J., DAGAN, T. & BOUCHER, Y. 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct,* 4, 34.

BARTHÉLEMY, J.-P. & MCMORRIS, F. 1986. The median procedure for n-trees. *Journal of Classification,* 3, 329-334.

BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 3-10.

BAUM, B. R. & RAGAN, M. A. 1993. Reply to A. G. Rodrigo's "A Comment on Baum's Method for Combining Phylogenetic Trees". *Taxon,* 42, 637-640.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L. & BIGNELL, H. R. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456, 53-59.

BILLERA, L. J., HOLMES, S. P. & VOGTMANN, K. 2001. Geometry of the space of phylogenetic trees. *Advances in applied mathematics,* 27, 733-767.

BININDA-EMONDS, O. R. 2003. Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. *Systematic Biology,* 52, 839-848.

BININDA-EMONDS, O. R. 2004a. *Phylogenetic supertrees: combining information to reveal the tree of life*, Springer.

BININDA-EMONDS, O. R. P. & BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology,* 47, 497-508.

BININDA-EMONDS, O. R. P., GITTLEMAN, J. L. & PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews,* 74, 143-175.

BLAIR, J. E., IKEO, K., GOJOBORI, T. & HEDGES, S. B. 2002. The evolutionary position of nematodes. *BMC evolutionary biology,* 2, 7.

BROCHIER, C., BAPTESTE, E., MOREIRA, D. & PHILIPPE, H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends in genetics,* 18, 1-5.

BROCHIER-ARMANET, C., BOUSSAU, B., GRIBALDO, S. & FORTERRE, P. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology,* 6, 245-252.

BROCHIER-ARMANET, C., FORTERRE, P. & GRIBALDO, S. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. *Current opinion in microbiology,* 14, 274-281.

BRYANT, D. 1997. Building trees, hunting for trees, and comparing trees. *Unpublished PhD Thesis. Department of Mathematics, University of Canterbury, New Zealand*.

BRYANT, D. & STEEL, M. 2009. Computing the distribution of a tree metric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* 6, 420-426.

BUNEMAN, P. 1974. A characterization of rigid circuit graphs. *Discrete Math,* 9, 205-212.

CASTRESANA, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution,* 17, 540-552.

CHIAPPE, L. M. 2002. Basal bird phylogeny. *Mesozoic birds: above the heads of dinosaurs*, 448-472.

CICCARELLI, F. D., DOERKS, T., VON MERING, C., CREEVEY, C. J., SNEL, B. & BORK, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *science,* 311, 1283-1287.

COLLESS, D. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology,* 31, 100-104.

COTTON, J. A., SLATER, C. S. & WILKINSON, M. 2006. Discriminating supported and unsupported relationships in supertrees using triplets. *Systematic biology,* 55, 345-350.

COTTON, J. A. & WILKINSON, M. 2007. Majority-rule supertrees. *Systematic biology,* 56, 445-452.

COTTON, J. A. & WILKINSON, M. 2009. Supertrees join the mainstream of phylogenetics. *Trends in ecology & evolution,* 24, 1-3.

COX, C. J., FOSTER, P. G., HIRT, R. P., HARRIS, S. R. & EMBLEY, T. M. 2008. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences,* 105, 20356-20361.

CREEVEY, C. J., FITZPATRICK, D. A., PHILIP, G. K., KINSELLA, R. J., O'CONNELL, M. J., PENTONY, M. M., TRAVERS, S. A., WILKINSON, M. & MCINERNEY, J. O. 2004. Does a tree–like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 271, 2551-2558.

CREEVEY, C. J. & MCINERNEY, J. O. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics,* 21, 390-392.

DAGAN, T. & MARTIN, W. 2006. The tree of one percent. *Genome Biol,* 7, 118.

DAGAN, T. & MARTIN, W. 2009. Getting a better picture of microbial evolution en route to a network of genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364, 2187-2196.

DARWIN, C. 1859. On the origins of species by means of natural selection. *London: Murray*.

DAUBIN, V., GOUY, M. & PERRIERE, G. 2001. Bacterial molecular phylogeny using supertree approach. *GENOME INFORMATICS SERIES*, 155-164.

DAUBIN, V., MORAN, N. A. & OCHMAN, H. 2003. Phylogenetics and the cohesion of bacterial genomes. *science,* 301, 829-832.

DE QUEIROZ, A. & GATESY, J. 2007. The supermatrix approach to systematics. *Trends in ecology & evolution,* 22, 34-41.

DEGNAN, J. H. & ROSENBERG, N. A. 2006. Discordance of species trees with their most likely gene trees. *PLoS genetics,* 2, e68.

DOOLITTLE, W. F. 1999a. Lateral genomics. *Trends in cell biology,* 9, M5-M8.

DOOLITTLE, W. F. 1999b. Phylogenetic classification and the universal tree. *science,* 284, 2124-2128.

DRAGOO, J. W. & HONEYCUTT, R. L. 1997. Systematics of mustelid-like carnivores. *Journal of Mammalogy*, 426-443.

DRESS, A. & STEEL, M. 1992. Convex tree realizations of partitions. *Applied Mathematics Letters,* 5, 3-6.

DRUMMOND, A. J. & RAMBAUT, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology,* 7, 214.

DUNN, C. W., HEJNOL, A., MATUS, D. Q., PANG, K., BROWNE, W. E., SMITH, S. A., SEAVER, E., ROUSE, G. W., OBST, M. & EDGECOMBE, G. D. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature,* 452, 745-749.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32, 1792-1797.

EDGAR, R. C. & BATZOGLOU, S. 2006. Multiple sequence alignment. *Current opinion in structural biology,* 16, 368-373.

EDWARDS, A. W. 1984. *Likelihood*, CUP Archive.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P. & BETTMAN, B. 2009. Real-time DNA sequencing from single polymerase molecules. *science,* 323, 133-138.

ELIAS, I. 2006. Settling the intractability of multiple alignment. *Journal of Computational Biology,* 13, 1323-1339.

ESTABROOK, G. F., MCMORRIS, F. & MEACHAM, C. A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic biology,* 34, 193-200.

EULENSTEIN, O., CHEN, D., BURLEIGH, J. G., FERNÁNDEZ-BACA, D. & SANDERSON, M. J. 2004. Performance of flip supertree construction with a heuristic algorithm. *Systematic biology,* 53, 299-308.

FAITH, D. P. & CRANSTON, P. S. 1991. Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. *Cladistics,* 7, 1-28.

FARIAS, I. P., ORTÍ, G. & MEYER, A. 2000. Total evidence: molecules, morphology, and the phylogenetics of cichlid fishes. *Journal of Experimental Zoology,* 288, 76-92.

FARRIS, J. S. 1971. The hypothesis of nonspecificity and taxonomic congruence. *Annual Review of Ecology and Systematics,* 2, 277-302.

FARRIS, J. S., ALBERT, V. A., KÄLLERSJÖ, M., LIPSCOMB, D. & KLUGE, A. G. 1996. Parsimony jackknifing outperforms neighbor - joining. *Cladistics,* 12, 99-124.

FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution*, 16-24.

FELSENSTEIN, J. 1985a. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 783-791.

FELSENSTEIN, J. 1985b. Confidence limits on phylogenies with a molecular clock. *Systematic biology,* 34, 152-161.

FELSENSTEIN, J. 2004. Inferring phytogenies. *Sunderland, Massachusetts: Sinauer Associates,* 4.

FELSENSTEIN, J. & KISHINO, H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic biology,* 42, 193-200.

FENG, D.-F. & DOOLITTLE, R. F. 1987. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution,* 25, 351-360.

FIERS, W., CONTRERAS, R., DUERINCK, F., HAEGEMAN, G., ISERENTANT, D., MERREGAERT, J., MIN JOU, W., MOLEMANS, F., RAEYMAEKERS, A. & VAN DEN BERGHE, A. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature,* 260, 500-507.

FISHER, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,* 222, 309-368.

FITCH, W. M. 2000. Homology: a personal view on some of the problems. *Trends in genetics,* 16, 227-231.

FITZPATRICK, D. A., LOGUE, M. E., STAJICH, J. E. & BUTLER, G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC evolutionary biology,* 6, 99.

FLYNN, J. J., FINARELLI, J. A., ZEHR, S., HSU, J. & NEDBAL, M. A. 2005. Molecular phylogeny of the Carnivora (Mammalia): assessing the impact of increased sampling on resolving enigmatic relationships. *Systematic biology,* 54, 317-337.

FORST, C. V. & SCHULTEN, K. 2001. Phylogenetic analysis of metabolic pathways. *Journal of molecular evolution,* 52, 471-489.

FOSTER, P. G. 2001. The Idiot's Guide to the Zen of. Likelihood in a Nutshell in Seven Days for Dummies.

FOSTER, P. G. 2004. Modeling compositional heterogeneity. *Systematic biology,* 53, 485-495.

GATESY, J., BAKER, R. H. & HAYASHI, C. 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Systematic biology,* 53, 342-355.

GATESY, J., O'GRADY, P. & BAKER, R. H. 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics,* 15, 271-313.

GAUTHIER, J., KLUGE, A. G. & ROWE, T. 1988. Amniote phylogeny and the importance of fossils. *Cladistics,* 4, 105-209.

GEISLER, J., MCGOWEN, M., YANG, G. & GATESY, J. 2011. A supermatrix analysis of genomic, morphological, and paleontological data from crown Cetacea. *BMC evolutionary biology,* 11, 112.

GOLDMAN, N., ANDERSON, J. P. & RODRIGO, A. G. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic biology,* 49, 652-670.

GOLOBOFF, P. A., FARRIS, J. S. & NIXON, K. C. 2008. TNT, a free program for phylogenetic analysis. *Cladistics,* 24, 774-786.

GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of classification,* 3, 335-348.

GRAY, M. W. 1989. The evolutionary origins of organelles. *Trends in genetics,* 5, 294-299.

GRIBALDO, S. & BROCHIER, C. 2009. Phylogeny of prokaryotes: does it exist and why should we care? *Research in microbiology,* 160, 513-521.

GUPTA, R. S. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiology and Molecular Biology Reviews,* 62, 1435-1491.

GUSFIELD, D. 1991. Efficient algorithms for inferring evolutionary trees. *Networks,* 21, 19-28.

HAECKEL, E. H. P. A. 1866. *Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*, G. Reimer.

HALARY, S., LEIGH, J. W., CHEAIB, B., LOPEZ, P. & BAPTESTE, E. 2010. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences,* 107, 127-132.

HALL, T. & STACEY, J. 2009. *Python 3 for absolute beginners*, Apress.

HARARY, F. & PALMER, E. M. 1973. Graphical enumeration. DTIC Document.

HARPER, J. T., WAANDERS, E. & KEELING, P. J. 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *International Journal of Systematic and Evolutionary Microbiology,* 55, 487-496.

HARVEY, P. H. & PURVIS, A. 1991. Comparative methods for explaining adaptations. *Nature,* 351, 619-624.

HASEGAWA, M. & FUJIWARA, M. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molecular phylogenetics and evolution,* 2, 1-5.

HASHIMOTO, T., NAKAMURA, Y., NAKAMURA, F., SHIRAKURA, T., ADACHI, J., GOTO, N., OKAMOTO, K.-I. & HASEGAWA, M. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia. *Molecular biology and evolution,* 11, 65-71.

HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* 57, 97-109.

HIGGINS, D. G. & SHARP, P. M. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene,* 73, 237-244.

HOLTON, T. A. & PISANI, D. 2010. Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm. *Genome biology and evolution,* 2, 310.

HUELSENBECK, J. P., BULL, J. & CUNNINGHAM, C. W. 1996. Combining data in phylogenetic analysis. *Trends in ecology & evolution,* 11, 152-158.

HUELSENBECK, J. P. & RONQUIST, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17, 754-755.

HUERTA-CEPAS, J., DOPAZO, J. & GABALDÓN, T. 2010. ETE: a python Environment for Tree Exploration. *BMC bioinformatics,* 11, 24.

HYMAN, L. H. 1940. The invertebrates: Protozoa through Ctenophora.

JENNER, R. A. & SCHRAM, F. R. 1999. The grand game of metazoan phylogeny: rules and strategies. *Biological Reviews,* 74, 121-142.

JONES, D., TAYLOR, W. & THORNTON, J. 1994a. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* 33, 3038-3049.

JONES, D., TAYLOR, W. & THORNTON, J. 1994b. A mutation data matrix for transmembrane proteins. *FEBS letters,* 339, 269-275.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS,* 8, 275-282.

JONES, K. E., PRICE, S. A., GRENYER, R., CARDILLO, M., HABIB, M., PURVIS, A. & GITTLEMAN, J. L. 2003. Supertrees are a necessary not-so-evil: a comment on Gatesy et al. *Systematic biology,* 52, 724-729.

JUKES, T. H. & CANTOR, C. R. 1969. {Evolution of protein molecules}.

KELLY, S., WICKSTEAD, B. & GULL, K. 2011. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proceedings of the Royal Society B: Biological Sciences,* 278, 1009-1018.

KISHINO, H. & HASEGAWA, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of molecular evolution,* 29, 170-179.

KISHINO, H., MIYATA, T. & HASEGAWA, M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of molecular evolution,* 31, 151-160.

KITCHING, I., FOREY, P. & HUMPHRIES, C. 1998. Cladistics: the theory and practice of parsimony analysis. *Systematics Association publication (*

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic biology,* 38, 7-25.

KLUGE, A. G. 2004. On total evidence: for the record. *Cladistics,* 20, 205-207.

KOSHI, J. & GOLDSTEIN, R. 1998. Mathematical models of natural amino acid site mutations. *Proteins,* 32, 289-295.

KOSHI, J. M. & GOLDSTEIN, R. A. 1995. Context-dependent optimal substitution matrices. *Protein Engineering,* 8, 641-645.

KOSHI, J. M. & GOLDSTEIN, R. A. 1997. Mutation matrices and physical-chemical properties: correlations and implications.

KUBATKO, L. S., CARSTENS, B. C. & KNOWLES, L. L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics,* 25, 971-973.

KUPCZOK, A. 2011a. Consequences of different null models on the tree shape bias of supertree methods. *Systematic Biology,* 60, 218-225.

KUPCZOK, A. 2011b. Split-based computation of majority-rule supertrees. *BMC evolutionary biology,* 11, 205.

KUPCZOK, A., SCHMIDT, H. A. & VON HAESELER, A. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms for Molecular Biology,* 5, 37.

KURLAND, C., CANBACK, B. & BERG, O. G. 2003. Horizontal gene transfer: a critical view. *Proceedings of the National Academy of Sciences,* 100, 9658-9662.

LAKE, J. A. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences.

LAKE, J. A. & RIVERA, M. C. 1994. Was the nucleus the first endosymbiont? *Proceedings of the National Academy of Sciences of the United States of America,* 91, 2880.

LAMARCK, J., BAPTISTE 1809. Philosophical zoology: An exposition with regard to the natural history of animals. *Chicago, Chicago, EEUU.[Links]*.

LAPOINTE, F.-J. & LEVASSEUR, C. 2004. Everything you always wanted to know about the average consensus, and more. *Phylogenetic supertrees: Combining information to reveal the Tree of Life*, 87-106.

LARTILLOT, N., LEPAGE, T. & BLANQUART, S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics,* 25, 2286-2288.

LERAT, E., DAUBIN, V. & MORAN, N. A. 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. *PLoS biology,* 1, e19.

LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology,* 50, 913-925.

LIU, L. & PEARL, D. K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic biology,* 56, 504-514.

LLOYD, G. T., DAVIS, K. E., PISANI, D., TARVER, J. E., RUTA, M., SAKAMOTO, M., HONE, D. W., JENNINGS, R. & BENTON, M. J. 2008. Dinosaurs and the Cretaceous Terrestrial Revolution. *Proc Biol Sci,* 275, 2483-90.

LOSOS, J. B. & GLOR, R. E. 2003. Phylogenetic comparative methods and the geography of speciation. *Trends in ecology & evolution,* 18, 220-227.

LÖYTYNOJA, A. & GOLDMAN, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *science,* 320, 1632-1635.

MADDISON, W. 1989. Reconstructing character evolution on polytomous cladograms. *Cladistics,* 5, 365-377.

MADDISON, W. P. & MADDISON, D. 2001. Mesquite: a modular system for evolutionary analysis.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y.-J. & CHEN, Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437, 376-380.

MARGUSH, T. & MCMORRIS, F. R. 1981. Consensus n-trees. *Bulletin of Mathematical Biology,* 43, 239-244.

MARTIN, W. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays,* 21, 99-104.

MARTIN, W. & MÜLLER, M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature,* 392, 37-41.

MARTIN, W., RUJAN, T., RICHLY, E., HANSEN, A., CORNELSEN, S., LINS, T., LEISTER, D., STOEBE, B., HASEGAWA, M. & PENNY, D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences,* 99, 12246-12251.

MAU, B., NEWTON, M. A. & LARGET, B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics,* 55, 1-12.

MAXAM, A. M. & GILBERT, W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences,* 74, 560-564.

MCINERNEY, J. O., COTTON, J. A. & PISANI, D. 2008. The prokaryotic tree of life: past, present… and future? *Trends in ecology & evolution,* 23, 276-281.

MCINERNEY, J. O. & PISANI, D. 2007. Paradigm for life. *SCIENCE-NEW YORK THEN WASHINGTON-,* 318, 1390.

MCINERNEY, J. O., PISANI, D., BAPTESTE, E. & O'CONNELL, M. J. 2011. The public goods hypothesis for the evolution of life on Earth. *Biol Direct,* 6, 41.

MCKERNAN, K. J., PECKHAM, H. E., COSTA, G. L., MCLAUGHLIN, S. F., FU, Y., TSUNG, E. F., CLOUSER, C. R., DUNCAN, C., ICHIKAWA, J. K. & LEE, C. C. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research,* 19, 1527-1541.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics,* 21, 1087.

METZKER, M. L. 2009. Sequencing technologies—the next generation. *Nature Reviews Genetics,* 11, 31-46.

MICKEVICH, M. F. 1978. Taxonomic congruence. *Systematic biology,* 27, 143-158.

MIYAMOTO, M. M. & FITCH, W. M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Systematic biology,* 44, 64-76.

MOORE, B. R., SMITH, S. A. & DONOGHUE, M. J. 2006. Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Systematic Biology,* 55, 662-676.

NEEDLEMAN, S. B. & WUNSCH, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology,* 48, 443-453.

NELSON-SATHI, S., DAGAN, T., LANDAN, G., JANSSEN, A., STEEL, M., MCINERNEY, J. O., DEPPENMEIER, U. & MARTIN, W. F. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proceedings of the National Academy of Sciences,* 109, 20537-20542.

NIXON, K. C. & CARPENTER, J. M. 1996. On consensus, collapsibility, and clade concordance. *Cladistics,* 12, 305-321.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology,* 302, 205-217.

NYAKATURA, K. & BININDA-EMONDS, O. R. 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC biology,* 10, 12.

NYLANDER, J. A., RONQUIST, F., HUELSENBECK, J. P. & NIEVES-ALDREY, J. 2004. Bayesian phylogenetic analysis of combined data. *Systematic biology,* 53, 47-67.

OMLAND, K. E., LANYON, S. M. & FRITZ, S. J. 1999. A Molecular Phylogeny of the New World Orioles (< i> Icterus</i>): The Importance of Dense Taxon Sampling. *Molecular phylogenetics and evolution,* 12, 224-239.

OWEN, R. 1843. On the structure and homologies of the cephalic tentacles in the pearly nautilus. *Annals and Magazine of Natural History,* 12, 305-311.

PAGE, R. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics,* 14, 819-820.

PAGEL, M. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics,* 10, 405-415.

PHILIPPE, H., LARTILLOT, N. & BRINKMANN, H. 2005a. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular biology and evolution,* 22, 1246-1253.

PHILIPPE, H., LARTILLOT, N. & BRINKMANN, H. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol,* 22, 1246-1253.

PISANI, D., COTTON, J. A. & MCINERNEY, J. O. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular biology and evolution,* 24, 1752-1760.

PISANI, D. & WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic biology,* 51, 151-155.

POSADA, D. & CRANDALL, K. A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics,* 14, 817-818.

PRAGER, E. M. & WILSON, A. C. 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. *Journal of molecular evolution,* 27, 326-335.

PRICE, S. A., BININDA - EMONDS, O. R. & GITTLEMAN, J. L. 2005. A complete phylogeny of the whales, dolphins and even - toed hoofed mammals (Cetartiodactyla). *Biological reviews,* 80, 445-473.

PUIGBÒ, P. 2009. Search for a 'Tree of Life'in the thicket of the phylogenetic forest Pere Puigbò, Yuri I Wolf and Eugene V Koonin. *Journal of biology,* 8, 59.

PUIGBÒ, P., WOLF, Y. I. & KOONIN, E. V. 2010. The tree and net components of prokaryote evolution. *Genome biology and evolution,* 2, 745.

PURVIS, A. 1995a. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology,* 44, 251-255.

PURVIS, A. 1995b. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 348, 405-421.

RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol,* 1, 53-58.

RAMBAUT, A. & DRUMMOND, A. 2007. Tracer v1. 4.

RANWEZ, V., BERRY, V., CRISCUOLO, A., FABRE, P.-H., GUILLEMOT, S., SCORNAVACCA, C. & DOUZERY, E. J. 2007. PhySIC: a veto supertree method with desirable properties. *Systematic biology,* 56, 798-817.

REN, F., TANAKA, H. & YANG, Z. 2009. A likelihood look at the supermatrix–supertree controversy. *Gene,* 441, 119-125.

RIHOUX, B. & RAGIN, C. C. 2008. *Configurational comparative methods: qualitative comparative analysis (QCA) and related techniques*, SAGE Publications, Incorporated.

RIVERA, M. C. & LAKE, J. A. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *science,* 257, 74-76.

RIVERA, M. C. & LAKE, J. A. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature,* 431, 152-155.

RODRIGO, A. G. 1996. On combining cladograms. *Taxon*, 267-274.

ROHLF, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution,* 55, 2143-2160.

RONQUIST, F., KLOPFSTEIN, S., VILHELMSEN, L., SCHULMEISTER, S., MURRAY, D. L. & RASNITSYN, A. P. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic biology,* 61, 973-999.

RONQUIST, F., TESLENKO, M., VAN DER MARK, P., AYRES, D. L., DARLING, A., HÖHNA, S., LARGET, B., LIU, L., SUCHARD, M. A. & HUELSENBECK, J. P. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology,* 61, 539-542.

ROTA-STABELLI, O., LARTILLOT, N., PHILIPPE, H. & PISANI, D. 2013. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. *Systematic biology,* 62, 121-133.

ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M., LEAMON, J. H., JOHNSON, K., MILGREW, M. J. & EDWARDS, M. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475, 348-352.

ROWE, L. & ARNQVIST, G. 2002. Sexually antagonistic coevolution in a mating system: combining experimental and comparative approaches to address evolutionary processes. *Evolution,* 56, 754-767.

RUTA, M. 2003. A species-level supertree for stylophoran echinoderms. *Acta Palaeontologica Polonica,* 48, 559-568.

RUTA, M., JEFFERY, J. E. & COATES, M. I. 2003. A supertree of early tetrapods. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270, 2507-2516.

RUTA, M., PISANI, D., LLOYD, G. T. & BENTON, M. J. 2007. A supertree of Temnospondyli: cladogenetic patterns in the most species-rich group of early tetrapods. *Proceedings of the Royal Society B: Biological Sciences,* 274, 3087-3095.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution,* 4, 406-425.

SANDERSON, M. J., PURVIS, A. & HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in ecology & evolution,* 13, 105-109.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences,* 74, 5463-5467.

SEMPLE, C. & STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics,* 105, 147-158.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research,* 13, 2498-2504.

SHIMODAIRA, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic biology,* 51, 492-508.

SHIMODAIRA, H. & HASEGAWA, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution,* 16, 1114-1116.

SHIMODAIRA, H. & HASEGAWA, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17, 1246-1247.

SIMMONS, M. P. & FREUDENSTEIN, J. V. 2002. Uninode coding vs gene tree parsimony for phylogenetic reconstruction using duplicate genes. *Molecular phylogenetics and evolution,* 23, 481-498.

SIU-TING, K., PISANI, D., CREEVEY, C. J. & WILKINSON, M. Submitted. Concatabominations: Identifying Unstable Taxa in Morphological Phylogenetics using a Heuristic Extension to Safe Taxonomic

SMITH, H. O., TOMB, J.-F., DOUGHERTY, B. A., FLEISCHMANN, R. D. & VENTER, J. C. 1995. Frequency and distribution of DNA uptake signal sequences in the Haemophilus influenzae Rd genome. *science,* 269, 538-540.

SMITH, S., BEAULIEU, J. & DONOGHUE, M. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC evolutionary biology,* 9, 37.

SNEL, B., BORK, P. & HUYNEN, M. A. 1999. Genome phylogeny based on gene content. *Nature genetics,* 21, 108-110.

SOKAL, R. R. & ROHLF, F. J. 1981. Taxonomic congruence in the Leptopodomorpha re-examined. *Systematic Zoology,* 30, 309-325.

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

STEEL, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of classification,* 9, 91-116.

STEEL, M. & BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic biology,* 49, 363-368.

STEEL, M. & RODRIGO, A. 2008. Maximum likelihood supertrees. *Systematic biology,* 57, 243-250.

STEEL, M. A. & PENNY, D. 1993. Distributions of tree comparison metrics—some new results. *Systematic biology,* 42, 126-141.

STRIMMER, K. & RAMBAUT, A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 269, 137-142.

STRIMMER, K. & VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular biology and evolution,* 13, 964-969.

SUKUMARAN, J. & HOLDER, M. T. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics,* 26, 1569-1571.

SUMMERS, M. R., SMYTHERS, G. W. & OROSZLAN, S. 1973. Thin-layer chromatography of sub-nanomole amounts of phenylthiohydantoin (PTH) amino acids on polyamide sheets. *Analytical biochemistry,* 53, 624-628.

SWIDERSKI, D. L., ZELDITCH, M. L. & FINK, W. L. 1998. Why morphometrics is not special: coding quantitative data for phylogenetic analysis. *Systematic biology,* 47, 508-519.

SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent. *Phylogenetic analysis of DNA sequences*, 295-333.

SWOFFORD, D. L. 2003. {PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.}.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. & HILLIS, D. 1990. Phylogeny reconstruction. *Molecular systematics,* 3, 407-514.

SWOFFORD, D. L., THORNE, J. L., FELSENSTEIN, J. & WIEGMANN, B. M. 1996. The topology-dependent permutation test for monophyly does not test for monophyly. *Systematic biology,* 45, 575-579.

TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci,* 17, 57-86.

TELFORD, M. J., BOURLAT, S. J., ECONOMOU, A., PAPILLON, D. & ROTA-STABELLI, O. 2008. The evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363, 1529-1537.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution*, 221-244.

THORLEY, J. L. 2000. *Cladistic information, leaf stability and supertree construction.* Citeseer.

THORLEY, J. L., WILKINSON, M. & CHARLESTON, M. 1998. The information content of consensus trees. *Advances in data science and classification.* Springer.

VAN DONGEN, S. M. 2000. Graph clustering by flow simulation.

VAN HOOFF, J. A. 1972. A comparative approach to the phylogeny of laughter and smiling.

VON HAESELER, A. 2012. Do we still need supertrees? *BMC biology,* 10, 13.

WATSON, J. D. & CRICK, F. H. The structure of DNA. Cold Spring Harbor symposia on quantitative biology, 1953. 123.

WERREN, J. H., ZHANG, W. & GUO, L. R. 1995. Evolution and phylogeny of Wolbachia: reproductive parasites of arthropods. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 261, 55-63.

WILKINSON, M. 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic biology,* 43, 343-368.

WILKINSON, M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. *Systematic biology,* 44, 501-514.

WILKINSON, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular biology and evolution,* 13, 437-444.

WILKINSON, M., COTTON, J. A., CREEVEY, C., EULENSTEIN, O., HARRIS, S. R., LAPOINTE, F.-J., LEVASSEUR, C., MCINERNEY, J. O., PISANI, D. & THORLEY, J. L. 2005a. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Systematic Biology,* 54, 419-431.

WILKINSON, M., COTTON, J. A., LAPOINTE, F.-J. & PISANI, D. 2007. Properties of supertree methods in the consensus setting. *Systematic biology,* 56, 330-337.

WILKINSON, M., PISANI, D., COTTON, J. A. & CORFE, I. 2005b. Measuring support and finding unsupported relationships in supertrees. *Systematic biology,* 54, 823-831.

WILKINSON, M. & THORLEY, J. L. 2001. No compromise on consensus. *Taxon,* 50, 181-184.

WILKINSON, M., THORLEY, J. L., PISANI, D. E., LAPOINTE, F.-J. & MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. *Phylogenetic supertrees: combining information to reveal the Tree of Life*, 564.

WOESE, C. R. & FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences,* 74, 5088-5090.

WOESE, C. R., KANDLER, O. & WHEELIS, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences,* 87, 4576-4579.

WOLF, Y. I., ROGOZIN, I. B., GRISHIN, N. V. & KOONIN, E. V. 2002. Genome trees and the tree of life. *Trends in genetics,* 18, 472-479.

WOLF, Y. I., ROGOZIN, I. B. & KOONIN, E. V. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome research,* 14, 29-36.

YANG, Z. & RANNALA, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular biology and evolution,* 14, 717-724.

Appendices

Appendix A

The L.U.St package

A phylogenetic software for inferring Maximum Likelihood supertrees

Manual Version 1.1 (July 2013)

Wasiu Akanni
Department of Biology, The National University of Ireland, Maynooth. Maynooth, Co. Kildare, Ireland.

Christopher J. Creevey
Animal & Grassland Research and Innovation Centre, Teagasc, Grange, Dunsany, Co. Meath, Ireland.

Davide Pisani
School of Biological Sciences and School of Earth Sciences, The University of Bristol. Woodland Road, Bristol BS8 1UG, United Kingdom

Contents

1 Introduction 2

2 Implementation 2

3 Citing L.U.St 2

4 Getting started 2

4.1 Options 2

4.2 Examples 3

5 Input data set 3

6 Starting Tree 3

7 Heuristic options 4

8 Output 4

9 Extra Goodies!!! 5

1. Introduction

Recent advances in supertree theory should now allow the implementation of Maximum Likelihood (ML) method, based on the use of an exponential distribution to model topological errors in phylogenies. Such approaches are expected to have distinct advantages over ad hoc approaches like the widely used Matrix representation with parsimony (MRP).
L.U.St includes the first parametric supertree method developed. The scripts that make up L.U.St are written in python and the ML method uses the subtree pruning and regrafting method to search the tree space.

2. Implementation

L.U.St comes in a folder containing all the various classes that are needed for the running of the software. L.U.St does not require any installation but as this is a python package, it requires an up to date and working version of python installed on your computer or server.

3. Citing L.U.St

Wasiu A. Akanni, Christopher J. Creevey and Davide Pisani: L.U.St: A tool for maximum likelihood supertree reconstruction. Oxford Bioinformatics Journal, (SUBMITTED).

4. Getting started

For a short description of the options available use the following command:

Python MLSupertree.py -h

Usage: python MLSupertree.py [options]

4.1. Options:

-h --help show this help message and exit

-i genetreeFILE, --input_treefile=genetreeFILE
The name of the file containing the fully resolved input gene trees in newick format (required)

-t STARTIN_OPTION, --start_tree_option=STARTIN_OPTION
'yes': if you have a starting tree. 'no' : if you want a random starting tree to be constructed by random stepwise addition Default = 'no' (required)

-s START_TREE_FILE, --start_treefile=START_TREE_FILE
The name of the file containing the fully resolved tarting supertree(s) in newick format (required ONLY IF YOU PICK YES ABOVE)

-n RANDOM_ITERATIONS, --number_of_iterations= RANDOM_ITERATIONS
The number of random starting trees/iterations Default=1

-o OUTFILE, --output_file=OUTFILE
The name of the file where you want the output of the analysis to be stored. Default= Mlsupertree_analysis_output

-c SPR_CHOICE, --heuristic=SPR_CHOICE
Please type 1: for a full exhaustive spr search
2: for a version that does not go thru every rooting point
3: a version that only considereds better trees
4: a version that does not include going through all the root and only considers trees with better likelihood.
Default=1

    - Gives the best result in tests carried out using empirical datasets

- Option 2: Faster than options 1. This search does not re-root each tree at every possible rooting point
- Option 3: Faster than option 1 and 2. The speed here is achieved by only considering the better trees (tree of equal likelihood although kept are not analysed further).
- Option 4: The fastest of them all. Only trees of better likelihood are analysed further and does not re-root the trees at every possible rooting point.

* Option 1 is recommended in terms of speed and accuracy.

8. Output

-o --output_file
The output of No-name is a file containing the Maximum Likelihood supertree(s), the likelihood value and the length of time the software took to reach completion.

9. FAQs

Q: can I used input trees containing polytomies?
A: No. At the moment the tree class can't handle polytomies.

Q: can I used starting supertrees containing polytomies?
A: No. At the moment the tree class can't handle polytomies.

Q: can I used input trees in nexus format?
A: No. At the moment the parsing scripts can only handle trees in newick format.

4.2. Examples:

python MLSupertree –i Example/Drosophila_inputtrees –t yes -o Drosophila_Mlstree_results –c 1

5. Input data set

-i –-input_treefile
The No-name software takes as input a file containing a list of newick formatted phylogenetic trees overlapping on some set of taxa (an example is included in the No-name directory in the file called Drosophila_empiricalTrees.txt).

6. Starting Tree

-t --start_tree_option
You can choose to provide your own starting tree(s) by giving the name of the file containing your newick formatted starting tree(s) to the No-name or choose for No-name to generate random starting trees.
Random starting tree(s) will be generated on the combined taxon set of the input trees using a stepwise addition technique.

Note: your starting tree must contain all the taxa in your input tree data set.

7. Heuristic Option

-c --heuristic
Five heuristic search strategies based on subtree pruning and regrafting (spr) have been implanted in No-name.

- Option 1: This is the most exhaustive search strategy implemented. It involves:
    - A full round of spr on every tree being analysed
    - Re-rooting every tree being analysed at every possible rooting point.
    - Keeping every tree found to be of equal likelihood to the current tree for future analyses

10. Extra Goodies!!!

Calculate_supertrees_likelihoods.py

- This script was written to allow the user the ability to calculate the likelihoods of a list of supertrees inferred from an input tree dataset.

Resolve_phylogenies.py

- This script was written to deal with polytomous gene trees in the maximum likelihood analysis. We came up with the idea that we can used a resolved supertree such as an MRP tree to resolve the polytomous clades in the gene trees before a maximum likelihood analysis. This should be more effective than simply randomly resolving the clades.

Winning_site_test.py

- This script calculates the winning site test between two supertrees that have been inferred from the same input tree dataset but probably using different supertree methods .

ExtractTaxon_file.py

- This script can used to create a list of all the taxon set of an overlapping input tree data set.

Statistical_test.sh **and** ***Statiscal_test.py***

- This two script work together with the CONSEL software to calculate the SH,KH, and AU test and rank supertrees inferred on the same input tree dataset but using different supetree methods (competing hypothesis).
- Both the Supertree file and the input tree dataset file should be in newick format.

- There is an example script called run_stat_test.sh. this script can be modified by changing the name of the input and output files to that of your own files.

Deroot.py

- This script offers the capability to deroot rooted trees. This script requires the DENDROPY package to be installed

generate_bootstrap_replicate_files.py

- This script enables the user to create replicate bootstrap datasets for the input tree dataset which can be used to generate bootstrap support values for the clades in the estimated ML supertree.

References

Steel, Mike, and Allen Rodrigo. "Maximum likelihood supertrees." Systematic biology 57.2 (2008): 243-250.

Bryant, David, and Mike Steel. "Computing the distribution of a tree metric." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 6.3 (2009): 420-426.

A lot of thanks to Dr David Pisani and Dr Christopher Creevey for all their help in writing this software.

Appendix A: The manual for the L.U.St package

Appendix B

Taxa	Family	A
Vulpes ferrilata	Canidae	✓
Dusicyon australis	Canidae	✓
Galerella flavescens	Herpestidae	✓
Galerella ochracea	Herpestidae	✓
Mustela lutreolina	Mustelidae	✓
Herpestes smithii	Herpestidae	✓
Nasuella olivacea	Procyonidae	✓
Procyon pygmaeus	Procyonidae	✓
Vulpes pallida	Canidae	✓
Bassaricyon pauli	Procyonidae	✓
Herpestes semitorquatus	Herpestidae	✓
Conepatus chinga	Mustelidae	✓
Mustela nudipes	Mustelidae	✓
Mustela africana	Mustelidae	✓
Mustela felipei	Mustelidae	✓
Mustela strigidorsa	Mustelidae	✓
Vulpes bengalensis	Canidae	✓
Ictonyx libyca	Mustelidae	✓
Zalophus japonicus	Otariidae	✓
Melogale personata	Mustelidae	✓
Lyncodon patagonicus	Mustelidae	✓
Herpestes vitticollis	Herpestidae	✓
Mellivora capensis	Mustelidae	✓
Galictis cuja	Mustelidae	✓
Bassaricyon beddardi	Procyonidae	✓
Prionodon linsang	Prionodontidae	✓
Ailuropoda melanoleuca	Ursidae	
Helarctos malayanus	Ursidae	
Ursus americanus	Ursidae	
Melursus ursinus	Ursidae	
Ursus thibetanus	Ursidae	
Tremarctos ornatus	Ursidae	
Ursus maritimus	Ursidae	
Ursus arctos	Ursidae	
Prionodon pardicolor	Prionodontidae	
Leopardus pajeros	Felidae	
Atelocynus microtis	Canidae	
Lycalopex sechurae	Canidae	
Cerdocyon thous	Canidae	
Lycalopex griseus	Canidae	
Lycalopex gymnocercus	Canidae	
Lycalopex culpaeus	Canidae	
Lycalopex vetulus	Canidae	

Lycalopex fulvipes	Canidae
Arctogalidia trivirgata	Viverridae
Speothos venaticus	Canidae
Chrysocyon brachyurus	Canidae
Macrogalidia musschenbroekii	Viverridae
Paradoxurus jerdoni	Viverridae
Paradoxurus zeylonensis	Viverridae
Nandinia binotata	Viverridae
Urocyon cinereoargenteus	Canidae
Urocyon littoralis	Canidae
Otocyon megalotis	Canidae
Vulpes chama	Canidae
Nyctereutes procyonoides	Canidae
Vulpes cana	Canidae
Vulpes zerda	Canidae
Vulpes corsac	Canidae
Vulpes rueppellii	Canidae
Vulpes vulpes	Canidae
Vulpes lagopus	Canidae
Vulpes velox	Canidae
Vulpes macrotis	Canidae
Canis adustus	Canidae
Canis mesomelas	Canidae
Canis simensis	Canidae
Lycaon pictus	Canidae
Cuon alpinus	Canidae
Canis aureus	Canidae
Canis lupus	Canidae
Canis latrans	Canidae
Arctocephalus gazella	Otariidae
Arctocephalus tropicalis	Otariidae
Otaria flavescens	Otariidae
Odobenus rosmarus	Odobenidae
Zalophus wollebaeki	Otariidae
Arctocephalus pusillus	Otariidae
Callorhinus ursinus	Otariidae
Eumetopias jubatus	Otariidae
Zalophus californianus	Otariidae
Neophoca cinerea	Otariidae
Phocarctos hookeri	Otariidae
Arctocephalus galapagoensis	Otariidae
Arctocephalus philippii	Otariidae
Arctocephalus townsendi	Otariidae
Arctocephalus forsteri	Otariidae
Arctocephalus australis	Otariidae
Erignathus barbatus	Phocidae

Cystophora cristata	Phocidae
Monachus monachus	Phocidae
Monachus tropicalis	Phocidae
Monachus schauinslandi	Phocidae
Lobodon carcinophaga	Phocidae
Ommatophoca rossii	Phocidae
Prionailurus iriomotensis	Felidae
Mirounga angustirostris	Phocidae
Mirounga leonina	Phocidae
Hydrurga leptonyx	Phocidae
Leptonychotes weddellii	Phocidae
Pagophilus groenlandicus	Phocidae
Histriophoca fasciata	Phocidae
Phoca vitulina	Phocidae
Phoca largha	Phocidae
Halichoerus grypus	Phocidae
Pusa caspica	Phocidae
Pusa sibirica	Phocidae
Pusa hispida	Phocidae
Conepatus semistriatus	Mephitidae
Leopardus wiedii	Felidae
Leopardus pardalis	Felidae
Leopardus colocolo	Felidae
Leopardus jacobitus	Felidae
Leopardus geoffroyi	Felidae
Leopardus guigna	Felidae
Leopardus tigrinus	Felidae
Neofelis nebulosa	Felidae
Panthera tigris	Felidae
Uncia uncia	Felidae
Panthera onca	Felidae
Panthera pardus	Felidae
Panthera leo	Felidae
Lynx canadensis	Felidae
Lynx lynx	Felidae
Lynx pardinus	Felidae
Lynx rufus	Felidae
Leptailurus serval	Felidae
Profelis aurata	Felidae
Caracal caracal	Felidae
Puma concolor	Felidae
Puma yagouaroundi	Felidae
Pardofelis marmorata	Felidae
Catopuma badia	Felidae
Catopuma temminckii	Felidae
Acinonyx jubatus	Felidae

Prionailurus viverrinus	Felidae
Prionailurus planiceps	Felidae
Prionailurus bengalensis	Felidae
Prionailurus rubiginosus	Felidae
Felis nigripes	Felidae
Felis chaus	Felidae
Felis manul	Felidae
Felis margarita	Felidae
Felis bieti	Felidae
Felis catus	Felidae
Felis silvestris	Felidae
Spilogale pygmaea	Mephitidae
Liberiictis kuhni	Herpestidae
Helogale hirtula	Herpestidae
Mungos gambianus	Herpestidae
Crossarchus alexandri	Herpestidae
Mungos mungo	Herpestidae
Herpestes naso	Herpestidae
Paracynictis selousi	Herpestidae
Bdeogale crassicauda	Herpestidae
Rhynchogale melleri	Herpestidae
Atilax paludinosus	Herpestidae
Ichneumia albicauda	Herpestidae
Crossarchus obscurus	Herpestidae
Herpestes brachyurus	Herpestidae
Herpestes urva	Herpestidae
Herpestes edwardsi	Herpestidae
Herpestes fuscus	Herpestidae
Herpestes javanicus	Herpestidae
Galerella pulverulenta	Herpestidae
Galerella sanguinea	Herpestidae
Herpestes ichneumon	Herpestidae
Bdeogale nigripes	Herpestidae
Suricata suricatta	Herpestidae
Helogale parvula	Herpestidae
Cynictis penicillata	Herpestidae
Cryptoprocta ferox	Eupleridae
Eupleres goudotii	Eupleridae
Fossa fossana	Eupleridae
Arctictis binturong	Viverridae
Paguma larvata	Viverridae
Paradoxurus hermaphroditus	Viverridae
Cynogale bennettii	Viverridae
Chrotogale owstoni	Viverridae
Diplogale hosei	Viverridae
Hemigalus derbyanus	Viverridae

Galidictis fasciata	Eupleridae
Galidia elegans	Eupleridae
Mungotictis decemlineata	Eupleridae
Salanoia concolor	Eupleridae
Genetta johnstoni	Viverridae
Proteles cristata	Hyaenidae
Crocuta crocuta	Hyaenidae
Hyaena hyaena	Hyaenidae
Hyaena brunnea	Hyaenidae
Genetta piscivora	Viverridae
Viverricula indica	Viverridae
Viverra tangalunga	Viverridae
Viverra zibetha	Viverridae
Civettictis civetta	Viverridae
Viverra civettina	Viverridae
Viverra megaspila	Viverridae
Genetta thierryi	Viverridae
Genetta abyssinica	Viverridae
Genetta bourloni	Viverridae
Poiana richardsonii	Viverridae
Genetta angolensis	Viverridae
Genetta tigrina	Viverridae
Genetta genetta	Viverridae
Genetta maculata	Viverridae
Genetta cristata	Viverridae
Genetta pardina	Viverridae
Genetta poensis	Viverridae
Genetta victoriae	Viverridae
Genetta servalina	Viverridae
Ailurus fulgens	Ailuridae
Conepatus humboldtii	Mephitidae
Mydaus javanensis	Mephitidae
Mydaus marchei	Mephitidae
Conepatus leuconotus	Mephitidae
Spilogale gracilis	Mephitidae
Spilogale putorius	Mephitidae
Mephitis macroura	Mephitidae
Mephitis mephitis	Mephitidae
Potos flavus	Procyonidae
Procyon cancrivorus	Procyonidae
Procyon lotor	Procyonidae
Bassariscus sumichrasti	Procyonidae
Bassariscus astutus	Procyonidae
Bassaricyon alleni	Procyonidae
Bassaricyon gabbii	Procyonidae
Nasua nasua	Procyonidae

Nasua narica	Procyonidae
Galictis vittata	Mustelidae
Poecilogale albinucha	Mustelidae
Taxidea taxus	Mustelidae
Vormela peregusna	Mustelidae
Ictonyx striatus	Mustelidae
Melogale moschata	Mustelidae
Martes pennanti	Mustelidae
Gulo gulo	Mustelidae
Eira barbara	Mustelidae
Martes foina	Mustelidae
Martes melampus	Mustelidae
Martes flavigula	Mustelidae
Martes americana	Mustelidae
Martes zibellina	Mustelidae
Martes martes	Mustelidae
Enhydra lutris	Mustelidae
Pteronura brasiliensis	Mustelidae
Mustela kathiah	Mustelidae
Mustela frenata	Mustelidae
Neovison vison	Mustelidae
Lutrogale perspicillata	Mustelidae
Hydrictis maculicollis	Mustelidae
Mustela erminea	Mustelidae
Aonyx cinerea	Mustelidae
Aonyx capensis	Mustelidae
Mustela itatsi	Mustelidae
Lutra lutra	Mustelidae
Lutra sumatrana	Mustelidae
Mustela lutreola	Mustelidae
Mustela sibirica	Mustelidae
Mustela altaica	Mustelidae
Mustela nivalis	Mustelidae
Lontra canadensis	Mustelidae
Lontra felina	Mustelidae
Lontra longicaudis	Mustelidae
Lontra provocax	Mustelidae
Mustela putorius	Mustelidae
Mustela nigripes	Mustelidae
Mustela eversmanii	Mustelidae
Arctonyx collaris	Mustelidae
Meles meles	Mustelidae
Meles anakuma	Mustelidae

Appendix B. The list of species used for the carnivore analysis in order as they are ranked by the leaf stability test. The top 26 ranked unstable taxa are ticked in column A.

Appendix C

TAXA	PHYLA	Origin	A	B	C
Acidobacterium	ACIDOBACTERIA	NCBI			
Fibrobacter	ACIDOBACTERIA	NCBI			
Rubrobacter	ACTINOBACTERIA	NCBI			
Conexibacter	ACTINOBACTERIA	NCBI			
Rothia	ACTINOBACTERIA	NCBI			
Kocuria	ACTINOBACTERIA	NCBI			
Micrococcus	ACTINOBACTERIA	NCBI			
Renibacterium	ACTINOBACTERIA	NCBI			
Arthrobacter	ACTINOBACTERIA	NCBI			
Slackia	ACTINOBACTERIA	NCBI			
Eggerthella	ACTINOBACTERIA	NCBI			
Cryptobacterium	ACTINOBACTERIA	NCBI			
Olsenella	ACTINOBACTERIA	NCBI			
Atopobium	ACTINOBACTERIA	NCBI			
Leifsonia	ACTINOBACTERIA	NCBI			
Clavibacter	ACTINOBACTERIA	NCBI			
Microbacterium	ACTINOBACTERIA	NCBI			
Amycolatopsis	ACTINOBACTERIA	NCBI			
Saccharomonospora	ACTINOBACTERIA	NCBI			
Actinosynnema	ACTINOBACTERIA	NCBI			
Saccharopolyspora	ACTINOBACTERIA	NCBI			
Kribbella	ACTINOBACTERIA	NCBI			
Nocardioides	ACTINOBACTERIA	NCBI			
Stackebrandtia	ACTINOBACTERIA	NCBI			
Salinispora	ACTINOBACTERIA	NCBI			
Verrucosispora	ACTINOBACTERIA	NCBI			
Micromonospora	ACTINOBACTERIA	NCBI			
Catenulispora	ACTINOBACTERIA	NCBI			
Frankia	ACTINOBACTERIA	NCBI			
Acidothermus	ACTINOBACTERIA	NCBI			
Nocardiopsis	ACTINOBACTERIA	NCBI			
Geodermatophilus	ACTINOBACTERIA	NCBI			
Kytococcus	ACTINOBACTERIA	NCBI			
Intrasporangium	ACTINOBACTERIA	NCBI			
Propionibacterium	ACTINOBACTERIA	NCBI			
Brachybacterium	ACTINOBACTERIA	NCBI			
Nocardia	ACTINOBACTERIA	NCBI			
Rhodococcus	ACTINOBACTERIA	NCBI			
Acidimicrobium	ACTINOBACTERIA	NCBI			
Arcanobacterium	ACTINOBACTERIA	NCBI			
Beutenbergia	ACTINOBACTERIA	NCBI			
Bifidobacterium	ACTINOBACTERIA	NCBI			
Cellulomonas	ACTINOBACTERIA	NCBI			

Coriobacterium	ACTINOBACTERIA	NCBI
Corynebacterium	ACTINOBACTERIA	NCBI
Gardnerella	ACTINOBACTERIA	NCBI
Gordonia	ACTINOBACTERIA	NCBI
Jonesia	ACTINOBACTERIA	NCBI
Kineococcus	ACTINOBACTERIA	NCBI
Mobiluncus	ACTINOBACTERIA	NCBI
Mycobacterium	ACTINOBACTERIA	NCBI
Nakamurella	ACTINOBACTERIA	NCBI
Pseudonocardia	ACTINOBACTERIA	NCBI
Sanguibacter	ACTINOBACTERIA	NCBI
Segniliparus	ACTINOBACTERIA	NCBI
Tropheryma	ACTINOBACTERIA	NCBI
Tsukamurella	ACTINOBACTERIA	NCBI
Xylanimonas	ACTINOBACTERIA	NCBI
Hyphomonas	ALPHA-PROTEOBACTERIA	NCBI
Hirschia	ALPHA-PROTEOBACTERIA	NCBI
Maricaulis	ALPHA-PROTEOBACTERIA	NCBI
Brevundimonas	ALPHA-PROTEOBACTERIA	NCBI
Asticcacaulis	ALPHA-PROTEOBACTERIA	NCBI
Phenylobacterium	ALPHA-PROTEOBACTERIA	NCBI
Caulobacter	ALPHA-PROTEOBACTERIA	NCBI
Mesorhizobium	ALPHA-PROTEOBACTERIA	NCBI
Ochrobactrum	ALPHA-PROTEOBACTERIA	NCBI
Brucella	ALPHA-PROTEOBACTERIA	NCBI
Bartonella	ALPHA-PROTEOBACTERIA	NCBI
Sinorhizobium	ALPHA-PROTEOBACTERIA	NCBI
Agrobacterium	ALPHA-PROTEOBACTERIA	NCBI
Rhizobium	ALPHA-PROTEOBACTERIA	NCBI
Polymorphum	ALPHA-PROTEOBACTERIA	NCBI
Methylobacterium	ALPHA-PROTEOBACTERIA	NCBI
Beijerinckia	ALPHA-PROTEOBACTERIA	NCBI
Methylocella	ALPHA-PROTEOBACTERIA	NCBI
Rhodopseudomonas	ALPHA-PROTEOBACTERIA	NCBI
Bradyrhizobium	ALPHA-PROTEOBACTERIA	NCBI
Nitrobacter	ALPHA-PROTEOBACTERIA	NCBI
Oligotropha	ALPHA-PROTEOBACTERIA	NCBI
Starkeya	ALPHA-PROTEOBACTERIA	NCBI
Xanthobacter	ALPHA-PROTEOBACTERIA	NCBI
Azorhizobium	ALPHA-PROTEOBACTERIA	NCBI
Hyphomicrobium	ALPHA-PROTEOBACTERIA	NCBI
Rhodomicrobium	ALPHA-PROTEOBACTERIA	NCBI
Ruegeria	ALPHA-PROTEOBACTERIA	NCBI
Roseobacter	ALPHA-PROTEOBACTERIA	NCBI
Jannaschia	ALPHA-PROTEOBACTERIA	NCBI
Dinoroseobacter	ALPHA-PROTEOBACTERIA	NCBI

Ketogulonicigenium	ALPHA-PROTEOBACTERIA	NCBI		
Rhodobacter	ALPHA-PROTEOBACTERIA	NCBI		
Paracoccus	ALPHA-PROTEOBACTERIA	NCBI		
Magnetospirillum	ALPHA-PROTEOBACTERIA	NCBI		
Azospirillum	ALPHA-PROTEOBACTERIA	NCBI		
Rhodospirillum	ALPHA-PROTEOBACTERIA	NCBI		
Sphingobium	ALPHA-PROTEOBACTERIA	NCBI		
Novosphingobium	ALPHA-PROTEOBACTERIA	NCBI		
Erythrobacter	ALPHA-PROTEOBACTERIA	NCBI		
Sphingopyxis	ALPHA-PROTEOBACTERIA	NCBI		
Sphingomonas	ALPHA-PROTEOBACTERIA	NCBI		
Zymomonas	ALPHA-PROTEOBACTERIA	NCBI		
Gluconobacter	ALPHA-PROTEOBACTERIA	NCBI		
Gluconacetobacter	ALPHA-PROTEOBACTERIA	NCBI		
Acetobacter	ALPHA-PROTEOBACTERIA	NCBI		
Acidiphilium	ALPHA-PROTEOBACTERIA	NCBI		
Anaplasma	ALPHA-PROTEOBACTERIA	NCBI	✓	
Ehrlichia	ALPHA-PROTEOBACTERIA	NCBI	✓	
Granulibacter	ALPHA-PROTEOBACTERIA	NCBI		
Magnetococcus	ALPHA-PROTEOBACTERIA	NCBI		
Neorickettsia	ALPHA-PROTEOBACTERIA	NCBI	✓	✓
Orientia	ALPHA-PROTEOBACTERIA	NCBI	✓	
Parvibaculum	ALPHA-PROTEOBACTERIA	NCBI		
Parvularcula	ALPHA-PROTEOBACTERIA	NCBI		
Rickettsia	ALPHA-PROTEOBACTERIA	NCBI		
Wolbachia	ALPHA-PROTEOBACTERIA	NCBI	✓	✓
Aquifex	AQUIFICAE	NCBI		
Desulfurobacterium	AQUIFICAE	NCBI		
Hydrogenobacter	AQUIFICAE	NCBI		
Hydrogenobaculum	AQUIFICAE	NCBI		
Persephonella	AQUIFICAE	NCBI		
Exiguobacterium	BACILLI	NCBI		
Paenibacillus	BACILLI	NCBI		
Brevibacillus	BACILLI	NCBI		
Geobacillus	BACILLI	NCBI		
Lysinibacillus	BACILLI	NCBI		
Oceanobacillus	BACILLI	NCBI		
Anoxybacillus	BACILLI	NCBI		
Leuconostoc	BACILLI	NCBI		
Oenococcus	BACILLI	NCBI		
Pediococcus	BACILLI	NCBI		
Macrococcus	BACILLI	NCBI		
Staphylococcus	BACILLI	NCBI		
Eubacterium	BACILLI	NCBI		
Butyrivibrio	BACILLI	NCBI		
Clostridium	BACILLI	NCBI		

Mahella	BACILLI	NCBI
Caldicellulosiruptor	BACILLI	NCBI
Alkaliphilus	BACILLI	NCBI
Anaerococcus	BACILLI	NCBI
Finegoldia	BACILLI	NCBI
Halanaerobium	BACILLI	NCBI
Halothermothrix	BACILLI	NCBI
Acetohalobium	BACILLI	NCBI
Desulfitobacterium	BACILLI	NCBI
Heliobacterium	BACILLI	NCBI
Desulfotomaculum	BACILLI	NCBI
Pelotomaculum	BACILLI	NCBI
Aerococcus	BACILLI	NCBI
Bacillus	BACILLI	NCBI
Carnobacterium	BACILLI	NCBI
Enterococcus	BACILLI	NCBI
Lactobacillus	BACILLI	NCBI
Lactococcus	BACILLI	NCBI
Listeria	BACILLI	NCBI
Melissococcus	BACILLI	NCBI
Streptococcus	BACILLI	NCBI
Spirosoma	BACTEROIDETE	NCBI
Dyadobacter	BACTEROIDETE	NCBI
Leadbetterella	BACTEROIDETE	NCBI
Marivirga	BACTEROIDETE	NCBI
Cytophaga	BACTEROIDETE	NCBI
Haliscomenobacter	BACTEROIDETE	NCBI
Chitinophaga	BACTEROIDETE	NCBI
Sphingobacterium	BACTEROIDETE	NCBI
Pedobacter	BACTEROIDETE	NCBI
Flavobacteriaceae	BACTEROIDETE	NCBI
Riemerella	BACTEROIDETE	NCBI
Weeksella	BACTEROIDETE	NCBI
Capnocytophaga	BACTEROIDETE	NCBI
Robiginitalea	BACTEROIDETE	NCBI
Maribacter	BACTEROIDETE	NCBI
Cellulophaga	BACTEROIDETE	NCBI
Gramella	BACTEROIDETE	NCBI
Zunongwangia	BACTEROIDETE	NCBI
Croceibacter	BACTEROIDETE	NCBI
Krokinobacter	BACTEROIDETE	NCBI
Flavobacterium	BACTEROIDETE	NCBI
Parabacteroides	BACTEROIDETE	NCBI
Porphyromonas	BACTEROIDETE	NCBI
Prevotella	BACTEROIDETE	NCBI
Bacteroides	BACTEROIDETE	NCBI

Paludibacter	BACTEROIDETE	NCBI		
Odoribacter	BACTEROIDETE	NCBI		
Blattabacterium	BACTEROIDETE	NCBI	✓	✓
Fluviicola	BACTEROIDETE	NCBI		
Rhodothermus	BACTEROIDETE	NCBI		
Salinibacter	BACTEROIDETE	NCBI		
Aromatoleum	BETA-PROTEOBACTERIA	NCBI		
Azoarcus	BETA-PROTEOBACTERIA	NCBI		
Nitrosospira	BETA-PROTEOBACTERIA	NCBI		
Nitrosomonas	BETA-PROTEOBACTERIA	NCBI		
Leptothrix	BETA-PROTEOBACTERIA	NCBI		
Methylibium	BETA-PROTEOBACTERIA	NCBI		
Verminephrobacter	BETA-PROTEOBACTERIA	NCBI		
Delftia	BETA-PROTEOBACTERIA	NCBI		
Comamonas	BETA-PROTEOBACTERIA	NCBI		
Acidovorax	BETA-PROTEOBACTERIA	NCBI		
Alicycliphilus	BETA-PROTEOBACTERIA	NCBI		
Rhodoferax	BETA-PROTEOBACTERIA	NCBI		
Polaromonas	BETA-PROTEOBACTERIA	NCBI		
Variovorax	BETA-PROTEOBACTERIA	NCBI		
Ralstonia	BETA-PROTEOBACTERIA	NCBI		
Cupriavidus	BETA-PROTEOBACTERIA	NCBI		
Burkholderia	BETA-PROTEOBACTERIA	NCBI		
Polynucleobacter	BETA-PROTEOBACTERIA	NCBI		✓
Pusillimonas	BETA-PROTEOBACTERIA	NCBI		
Achromobacter	BETA-PROTEOBACTERIA	NCBI		
Bordetella	BETA-PROTEOBACTERIA	NCBI		
Herminiimonas	BETA-PROTEOBACTERIA	NCBI		
Janthinobacterium	BETA-PROTEOBACTERIA	NCBI		
Laribacter	BETA-PROTEOBACTERIA	NCBI		
Chromobacterium	BETA-PROTEOBACTERIA	NCBI		
Sideroxydans	BETA-PROTEOBACTERIA	NCBI		
Gallionella	BETA-PROTEOBACTERIA	NCBI		
Dechloromonas	BETA-PROTEOBACTERIA	NCBI		
Herbaspirillum	BETA-PROTEOBACTERIA	NCBI		
Methylobacillus	BETA-PROTEOBACTERIA	NCBI		
Methylotenera	BETA-PROTEOBACTERIA	NCBI		
Methylovorus	BETA-PROTEOBACTERIA	NCBI		✓
Neisseria	BETA-PROTEOBACTERIA	NCBI		
Waddlia	CHLAMYDIAE	NCBI		
Chlamydia	CHLAMYDIAE	NCBI		✓
Chlamydophila	CHLAMYDIAE	NCBI		✓
Opitutus	CHLAMYDIAE	NCBI		
Coraliomargarita	CHLAMYDIAE	NCBI		
Akkermansia	CHLAMYDIAE	NCBI		
Methylacidiphilum	CHLAMYDIAE	NCBI		

Chlorobium	CHLOROBI	NCBI	
Prosthecochloris	CHLOROBI	NCBI	
Chlorobaculum	CHLOROBI	NCBI	
Pelodictyon	CHLOROBI	NCBI	
Chloroherpeton	CHLOROBI	NCBI	
Herpetosiphon	CHLOROFLEX	NCBI	
Roseiflexus	CHLOROFLEX	NCBI	
Chloroflexus	CHLOROFLEX	NCBI	
Anaerolinea	CHLOROFLEX	NCBI	
Sphaerobacter	CHLOROFLEX	NCBI	
Dehalococcoides	CHLOROFLEX	NCBI	
Dehalogenimonas	CHLOROFLEX	NCBI	
Desulfurispirillum	CHRYSIOGENETES	NCBI	
Anabaena	CYANOBACTERIA	NCBI	
Nostoc	CYANOBACTERIA	NCBI	
Trichodesmium	CYANOBACTERIA	NCBI	
Cyanothece	CYANOBACTERIA	NCBI	
Cyanobacterium	CYANOBACTERIA	NCBI	✓
Microcystis	CYANOBACTERIA	NCBI	
Acaryochloris	CYANOBACTERIA	NCBI	
Gloeobacter	CYANOBACTERIA	NCBI	
Prochlorococcus	CYANOBACTERIA	NCBI	
Deferribacter	DEFERRIBACTERES	NCBI	
Calditerrivibrio	DEFERRIBACTERES	NCBI	
Denitrovibrio	DEFERRIBACTERES	NCBI	
Deinococcus	DEINOCOCCUS/THERMUS	NCBI	
Truepera	DEINOCOCCUS/THERMUS	NCBI	
Meiothermus	DEINOCOCCUS/THERMUS	NCBI	
Oceanithermus	DEINOCOCCUS/THERMUS	NCBI	
Marinithermus	DEINOCOCCUS/THERMUS	NCBI	
Geobacter	DELTA-PROTEOBACTERIA	NCBI	
Pelobacter	DELTA-PROTEOBACTERIA	NCBI	
Desulfomicrobium	DELTA-PROTEOBACTERIA	NCBI	
Desulfohalobium	DELTA-PROTEOBACTERIA	NCBI	
Desulfovibrio	DELTA-PROTEOBACTERIA	NCBI	
Lawsonia	DELTA-PROTEOBACTERIA	NCBI	
Haliangium	DELTA-PROTEOBACTERIA	NCBI	
Sorangium	DELTA-PROTEOBACTERIA	NCBI	
Stigmatella	DELTA-PROTEOBACTERIA	NCBI	
Myxococcus	DELTA-PROTEOBACTERIA	NCBI	
Anaeromyxobacter	DELTA-PROTEOBACTERIA	NCBI	
Bdellovibrio	DELTA-PROTEOBACTERIA	NCBI	
Desulfarculus	DELTA-PROTEOBACTERIA	NCBI	
Desulfatibacillum	DELTA-PROTEOBACTERIA	NCBI	
Desulfobacca	DELTA-PROTEOBACTERIA	NCBI	
Desulfobacterium	DELTA-PROTEOBACTERIA	NCBI	

Desulfobulbus	DELTA-PROTEOBACTERIA	NCBI			
Desulfotalea	DELTA-PROTEOBACTERIA	NCBI			
Desulfurivibrio	DELTA-PROTEOBACTERIA	NCBI			
Hippea	DELTA-PROTEOBACTERIA	NCBI			
Dictyoglomus	DICTYOGLOMI	NCBI			
Elusimicrobium	ELUSIMICROBIA	NCBI			
Campylobacter	EPSILON-PROTEOBACTERIA	NCBI			
Nautilia	EPSILON-PROTEOBACTERIA	NCBI			
Arcobacter	EPSILON-PROTEOBACTERIA	NCBI			
Nitratiruptor	EPSILON-PROTEOBACTERIA	NCBI			
Nitratifractor	EPSILON-PROTEOBACTERIA	NCBI			
Wolinella	EPSILON-PROTEOBACTERIA	NCBI			
Helicobacter	EPSILON-PROTEOBACTERIA	NCBI			
Candidatus	FIRMICUTES	NCBI			
Acidaminococcus	FIRMICUTES	NCBI			
Ethanoligenens	FIRMICUTES	NCBI			
Acholeplasma	FIRMICUTES	NCBI			
Ammonifex	FIRMICUTES	NCBI			
Aster	FIRMICUTES	NCBI		✓	✓
Carboxydothermus	FIRMICUTES	NCBI			
Clostridiales	FIRMICUTES	NCBI			
Coprothermobacter	FIRMICUTES	NCBI			
Mesoplasma	FIRMICUTES	NCBI	✓		
Moorella	FIRMICUTES	NCBI			
Mycoplasma	FIRMICUTES	NCBI			
Natranaerobius	FIRMICUTES	NCBI			
Onion	FIRMICUTES	NCBI	✓		✓
Ruminococcus	FIRMICUTES	NCBI			
Selenomonas	FIRMICUTES	NCBI			
Ureaplasma	FIRMICUTES	NCBI	✓		✓
Veillonella	FIRMICUTES	NCBI			
Sebaldella	FUSOBACTERIA	NCBI			
Leptotrichia	FUSOBACTERIA	NCBI			
Fusobacterium	FUSOBACTERIA	NCBI			
Ilyobacter	FUSOBACTERIA	NCBI			
Streptobacillus	FUSOBACTERIA	NCBI			
Pantoea	GAMMA-PROTEOBACTERIA	NCBI			
Erwinia	GAMMA-PROTEOBACTERIA	NCBI			
Buchnera	GAMMA-PROTEOBACTERIA	NCBI		✓	
Citrobacter	GAMMA-PROTEOBACTERIA	NCBI			
Escherichia	GAMMA-PROTEOBACTERIA	NCBI			
Shigella	GAMMA-PROTEOBACTERIA	NCBI			
Enterobacter	GAMMA-PROTEOBACTERIA	NCBI			
Klebsiella	GAMMA-PROTEOBACTERIA	NCBI			
Sodalis	GAMMA-PROTEOBACTERIA	NCBI			
Baumannia	GAMMA-PROTEOBACTERIA	NCBI		✓	

Pectobacterium	GAMMA-PROTEOBACTERIA	NCBI
Dickeya	GAMMA-PROTEOBACTERIA	NCBI
Rahnella	GAMMA-PROTEOBACTERIA	NCBI
Yersinia	GAMMA-PROTEOBACTERIA	NCBI
Serratia	GAMMA-PROTEOBACTERIA	NCBI
Proteus	GAMMA-PROTEOBACTERIA	NCBI
Xenorhabdus	GAMMA-PROTEOBACTERIA	NCBI
Photorhabdus	GAMMA-PROTEOBACTERIA	NCBI
Alkalilimnicola	GAMMA-PROTEOBACTERIA	NCBI
Halorhodospira	GAMMA-PROTEOBACTERIA	NCBI
Pseudoalteromonas	GAMMA-PROTEOBACTERIA	NCBI
Glaciecola	GAMMA-PROTEOBACTERIA	NCBI
Alteromonas	GAMMA-PROTEOBACTERIA	NCBI
Colwellia	GAMMA-PROTEOBACTERIA	NCBI
Shewanella	GAMMA-PROTEOBACTERIA	NCBI
Ferrimonas	GAMMA-PROTEOBACTERIA	NCBI
Aeromonas	GAMMA-PROTEOBACTERIA	NCBI
Photobacterium	GAMMA-PROTEOBACTERIA	NCBI
Vibrio	GAMMA-PROTEOBACTERIA	NCBI
Aliivibrio	GAMMA-PROTEOBACTERIA	NCBI
Psychromonas	GAMMA-PROTEOBACTERIA	NCBI
Cellvibrio	GAMMA-PROTEOBACTERIA	NCBI
Saccharophagus	GAMMA-PROTEOBACTERIA	NCBI
Chromohalobacter	GAMMA-PROTEOBACTERIA	NCBI
Halomonas	GAMMA-PROTEOBACTERIA	NCBI
Pseudomonas	GAMMA-PROTEOBACTERIA	NCBI
Azotobacter	GAMMA-PROTEOBACTERIA	NCBI
Marinobacter	GAMMA-PROTEOBACTERIA	NCBI
Hahella	GAMMA-PROTEOBACTERIA	NCBI
Xanthomonas	GAMMA-PROTEOBACTERIA	NCBI
Stenotrophomonas	GAMMA-PROTEOBACTERIA	NCBI
Pseudoxanthomonas	GAMMA-PROTEOBACTERIA	NCBI
Pasteurella	GAMMA-PROTEOBACTERIA	NCBI
Aggregatibacter	GAMMA-PROTEOBACTERIA	NCBI
Mannheimia	GAMMA-PROTEOBACTERIA	NCBI
Actinobacillus	GAMMA-PROTEOBACTERIA	NCBI
Haemophilus	GAMMA-PROTEOBACTERIA	NCBI
Acidithiobacillus	GAMMA-PROTEOBACTERIA	NCBI
Acinetobacter	GAMMA-PROTEOBACTERIA	NCBI
Alcanivorax	GAMMA-PROTEOBACTERIA	NCBI
Allochromatium	GAMMA-PROTEOBACTERIA	NCBI
Coxiella	GAMMA-PROTEOBACTERIA	NCBI
Cronobacter	GAMMA-PROTEOBACTERIA	NCBI
Dichelobacter	GAMMA-PROTEOBACTERIA	NCBI
Edwardsiella	GAMMA-PROTEOBACTERIA	NCBI
Francisella	GAMMA-PROTEOBACTERIA	NCBI

Gallibacterium	GAMMA-PROTEOBACTERIA	NCBI		
Gamma	GAMMA-PROTEOBACTERIA	NCBI		
Halothiobacillus	GAMMA-PROTEOBACTERIA	NCBI		
Idiomarina	GAMMA-PROTEOBACTERIA	NCBI		
Kangiella	GAMMA-PROTEOBACTERIA	NCBI		
Legionella	GAMMA-PROTEOBACTERIA	NCBI		
Marinomonas	GAMMA-PROTEOBACTERIA	NCBI		
Methylococcus	GAMMA-PROTEOBACTERIA	NCBI		
Moraxella	GAMMA-PROTEOBACTERIA	NCBI		
Nitrosococcus	GAMMA-PROTEOBACTERIA	NCBI		
Psychrobacter	GAMMA-PROTEOBACTERIA	NCBI		
Salmonella	GAMMA-PROTEOBACTERIA	NCBI		
Xylella	GAMMA-PROTEOBACTERIA	NCBI	✓	✓
Gemmatimonas	GEMMATIMONADETES	NCBI		
Pirellula	PLANCTOMYCETES	NCBI		
Rhodopirellula	PLANCTOMYCETES	NCBI		
Planctomyces	PLANCTOMYCETES	NCBI		
Isosphaera	PLANCTOMYCETES	NCBI		
Spirochaeta	SPIROCHETE	NCBI		
Treponema	SPIROCHETE	NCBI		
Borrelia	SPIROCHETE	NCBI		
Brachyspira	SPIROCHETE	NCBI		
Leptospira	SPIROCHETE	NCBI		
Aminobacterium	SYNERGISTETES	NCBI		
Fervidobacterium	THERMOTOGAE	NCBI		
Kosmotoga	THERMOTOGAE	NCBI		
Petrotoga	THERMOTOGAE	NCBI		
Haloterrigena	EURYARCHAEOTA	NCBI		
Halalkalicoccus	EURYARCHAEOTA	NCBI		
Methanosphaerula	EURYARCHAEOTA	NCBI		
Methanoculleus	EURYARCHAEOTA	NCBI		
Methanospirillum	EURYARCHAEOTA	NCBI		
Methanoplanus	EURYARCHAEOTA	NCBI		
Methanocorpusculum	EURYARCHAEOTA	NCBI		
Methanosaeta	EURYARCHAEOTA	NCBI		
Methanosarcina	EURYARCHAEOTA	NCBI		
Methanohalobium	EURYARCHAEOTA	NCBI		
Methanococcoides	EURYARCHAEOTA	NCBI		
Methanohalophilus	EURYARCHAEOTA	NCBI		
Methanocella	EURYARCHAEOTA	NCBI		
Methanococcus	EURYARCHAEOTA	NCBI		
Methanocaldococcus	EURYARCHAEOTA	NCBI		
Methanosphaera	EURYARCHAEOTA	NCBI		
Methanobacterium	EURYARCHAEOTA	NCBI		
Natrialba	EURYARCHAEOTA	NCBI		

Aciduliprofundum	EURYARCHAEOTA	NCBI		
Archaeoglobus	EURYARCHAEOTA	NCBI		
Ferroglobus	EURYARCHAEOTA	NCBI		
Haloarcula	EURYARCHAEOTA	NCBI		
Halobacterium	EURYARCHAEOTA	NCBI		
Haloferax	EURYARCHAEOTA	NCBI		
Halogeometricum	EURYARCHAEOTA	NCBI		
Halomicrobium	EURYARCHAEOTA	NCBI		
Haloquadratum	EURYARCHAEOTA	NCBI		
Halorhabdus	EURYARCHAEOTA	NCBI		
Halorubrum	EURYARCHAEOTA	NCBI		
Methanobrevibacter	EURYARCHAEOTA	NCBI		
Methanopyrus	EURYARCHAEOTA	NCBI		
Methanothermobacter	EURYARCHAEOTA	NCBI		
Methanothermus	EURYARCHAEOTA	NCBI		
Natronomonas	EURYARCHAEOTA	NCBI		
Picrophilus	EURYARCHAEOTA	NCBI		
Pyrococcus	EURYARCHAEOTA	NCBI		
Vulcanisaeta	CRENARCHAEOTA	NCBI		
Caldivirga	CRENARCHAEOTA	NCBI		
Acidianus	CRENARCHAEOTA	NCBI		
Acidilobus	CRENARCHAEOTA	NCBI		
Aeropyrum	CRENARCHAEOTA	NCBI		
Desulfurococcus	CRENARCHAEOTA	NCBI		
Hyperthermus	CRENARCHAEOTA	NCBI		
Ignicoccus	CRENARCHAEOTA	NCBI		
Ignisphaera	CRENARCHAEOTA	NCBI		
Metallosphaera	CRENARCHAEOTA	NCBI		
Pyrobaculum	CRENARCHAEOTA	NCBI		
Staphylothermus	CRENARCHAEOTA	NCBI		
Nanoarchaeum	NANOARCHAEOTA	NCBI	✓	✓
Cenarchaeum	THAUMARCHAEOTA	NCBI		
Nitrosopumilus	THAUMARCHAEOTA	NCBI		
Plantae	EUKARYOTE	JGI		
Rhizaria	EUKARYOTE	JGI		
Rhodophyta	EUKARYOTE	Ensembl		
Amoebozoa	EUKARYOTE	JGI		
Fungi	EUKARYOTE	JGI		
Excavata	EUKARYOTE	JGI		

Appendix c. The list of species used for the Tree of Life analyses and the excludable rogue taxa identified by the concatabominations analyses for the Prok_Euk (A), Prok (B), and Bac (C) data sets.

Publications

L.U.St: A tool for approximated maximum likelihood supertree reconstruction

Corresponding Author:

Davide Pisani:

School of Biological Sciences and School of Earth Sciences,

The University of Bristol.

Woodland Road, BS8 1UG

Bristol, UK.

E-mail: davide.pisani@bristol.ac.uk

Abstract

Background: Supertrees combine disparate, partially overlapping trees to generate a synthesis that provides a high level perspective that cannot be attained from the inspection of individual phylogenies. Supertrees can be seen as meta-analytical tools that can be used to make inferences based on results of previous scientific studies. Their meta-analytical application has increased in popularity since it was realised that the power of statistical tests for the study of evolutionary trends critically depends on the use of taxon-dense phylogenies. Further to that, supertrees have found applications in phylogenomics where they are used to combine gene trees and recover species phylogenies based on genome-scale data sets.

Results: Here, we present the L.U.St package, a python tool for approximate maximum likelihood supertree inference and illustrate its application using a genomic data set for the placental mammals. L.U.St allows the calculation of the approximate likelihood of a supertree, given a set of input trees, performs heuristic searches to look for the supertree of highest likelihood, and performs statistical tests of two or more supertrees. To this end, L.U.St implements a winning sites test allowing ranking of a collection of *a-priori* selected hypotheses, given as a collection of input supertree topologies. It also outputs a file of input-tree-wise likelihood scores that can be used as input to CONSEL for calculation of standard tests of two trees (e.g. Kishino-Hasegawa, Shimidoara-Hasegawa and Approximately Unbiased tests).

Conclusion: This is the first fully parametric implementation of a supertree method, it has clearly understood properties, and provides several advantages over currently available supertree approaches. It is easy to implement and works on any platform that has python installed.

Keywords; Supertrees, Maximum Likelihood, Phylogenomics, tests of two trees.

Availability: *bitBucket page* - *https://afro-juju@bitbucket.org/afro-juju/l.u.st.git*

Contact: Davide.Pisani@bristol.ac.uk

Background

Supertree methods are generalisation of consensus methods to the case of partially overlapping input trees, and any method that can be used to amalgamate a collection of such trees is a supertree method [1]. Supertrees were formally introduced to the realm of the classification sciences by Gordon [2], who described a Strict Consensus Supertree method. However, the first supertree algorithm was introduced by Aho and colleagues [3] as an application to merge partially overlapping databases. Since these early works, there has been a lot of interest in supertree reconstruction particularly in evolutionary biology where supertrees have found an application as meta-analytical tools used to combine, and derive inferences from, published phylogenetic trees. Purvis [4] presented the first application of a supertree in this context merging primate phylogenies obtained from the literature to generate a supertree, and using it to test evolutionary hypotheses. Since then, the application of supertrees and more specifically their use for reconstructing large phylogenies in evolutionary biology has continued to be on the rise, paralleled by a substantial interest in the development of supertree methods. More recently, supertrees have also found important applications in genomics where they have been used to combine gene trees and derive species phylogenies [5-9].

A large number of supertree methods have been developed since the time of the Aho algorithm. However, most actual supertrees have been derived using the Matrix Representation with Parsimony (MRP) method of Baum [10] and Ragan [11]. This is due to the availability of excellent parsimony software and the general good

understanding of the theory underlying parsimony. Yet theoretical justifications for the application of parsimony to the supertree setting are weak, and MRP is mostly implemented due to the fact that it is easily applicable in practice and tends to return well-resolved trees [12]. More generally, most available supertree methods are ad hoc, their properties being often poorly known, and the rationale for their application unclear [13-15]. The only exceptions seem to be those based on generalisations of well-known consensus methods [16], and the maximum likelihood (ML) method of Steel and Rodrigo [17].

We present a Python implementation of the ML supertree method of Steel and Rodrigo [17]. The method has been shown to be consistent on general statistical conditions unlike other approaches like MRP [17], and it is closely related to the majority rule (-) supertree method [16], with which it has been suggested to share important properties, in particular the fact that the supertrees it generates have been suggested to be, like those derived using majority rule (-), median trees for the input set [17].

The method is "approximate" in the sense that, likelihood vales are not normalised for tree size. However, it has been pointed out that at the least in the context of Maximum Likelihood analyses, under specific set of parameters, this should not be a major problem [18].

The ML supertree method is available as part of the Likelihood Utility for Supertrees (L.U.St) package. L.U.St is licensed under the GNU General Public License. Once downloaded, L.U.St can be run on any platform on which python is installed.

Implementation

L.U.St's estimation of the ML supertree operates by taking as input a file containing a set of newick-formatted trees (i.e. the input trees). L.U.St's ML supertree method navigates the tree space using four alternative heuristic search strategies, varying in their speed and heuristic nature. These are all based on Subtree Pruning Regrafting (SPR) algorithm. The user can either provide a starting supertree for the search or L.U.St can generate a random starting supertree using a stepwise addition technique. It should here be noted that as in standard ML phylogenetic analyses, providing a non-random starting tree (in the case of supertree reconstruction this could be a MRP supertree) would speed up the analysis. The likelihood score of the proposed supertree is calculated by first estimating the likelihood of each input tree, given the current supertree. After that, all input-tree wise likelihood values are summed to get the likelihood of the proposed supertree. Input tree wise likelihood values are calculated assuming that each input tree can be considered a subsample of the proposed supertree generated by pruning taxa and reconstructed with or without some topological distortion or incongruence. To calculate an input tree-wise likelihood value the proposed supertree is pruned to have the same taxon set of the considered input tree. After that the symmetric difference on full splits (i.e. the Robinson-Fould's distance) [19], designated as d, between the pruned supertree and the input tree is calculated, in order to evaluate how dissimilar the input tree and the supertree are. The symmetric difference (d) is then used to calculate the input-tree likelihood using Steel and Rodrigo's formula:

$$\mathbb{P}_{\mathcal{T},\Upsilon}[\mathcal{T}'] = \alpha \exp[-\beta d(\mathcal{T}', \mathcal{T}|\Upsilon)]$$

Where α is a normalising constant and β is a value representing the quantity and quality of the data used to infer the input tree. An exponential distribution is used to

model phylogenetic error. This implies that the probability that a given input tree is a sample of the proposed supertree decrease exponentially as d increases. The likelihood of each proposed superteee is then calculated summing across all tree-wise likelihood scores.

The method is "approximate" in the sense that, likelihood vales are not normalised for tree size. This means that the likelihood we calculate is a "weighted" sum of the input tree likelihoods, where the weights correspond to the tree-specific normalising constant (α). Albeit calculating these normalising factors is in theory possible [18], it is computationally very time consuming. However, Bryant and Steel [18] pointed out that if one uses small β values, the normalising constants simplify to $\alpha=1$ irrespective of the input-tree sizes. For pragmatic reason (to maximise speed of execution), we currently do not allow the user to select β, which has been fixed to a low value ($\beta=1$) to allow $\alpha = 1$. It has been pointed out that at the least in the context of Maximum Likelihood analyses this should not cause problems [18]. But we acknowledge that the ranking of trees will be based on approximate, rather than correct, likelihood values.

L.U.St includes methods that allows for a variety of extra functions, including statistical tests for choosing between alternative hypotheses (tests of two trees – Winning site test, Kishino Hasegawa (KH) test [20], Shimidoara Hasegawa (SH) test [21] and the Approximately unbiased (AU) test [22]). Whilst the winning site test can be run natively in L.U.St, the calculation of KH, SH, AU and other tests requires the use of CONSEL [23].To our knowledge there is no other software package that allows the extension of standard tests of two trees to the supertree framework. However, tests of two trees can have great utility in supertree research, as they can be

used, for example, to investigate the extent to which current evidence (i.e. currently published trees) support alternative phylogenetic hypotheses (i.e. a set of proposed supertrees). Further to that, tests of two trees can be used in the phylogenomic context to evaluate the extent to which a set of gene-trees can reject a set of alternative phylogenetic hypotheses (i.e. a set of supertrees). Below an example of the use of test of two super(trees) in the phylogenomic context is provided.

L.U.St offers the user other useful functions to randomly resolve polytomies, deroot trees, reroot trees, resolve polytomies in a set of trees according to a user-provided input tree, create bootstrap replicates of input tree datasets, prune phyologenies, convert nexus formatted trees to the newick format and vice versa, and extract the taxon set of sets of trees.

Example: Using supertree to investigate deep placental phylogeny.

Several hypotheses have been proposed for the position of the root of the placental mammals (Fig.1). Those that received the greatest support in recent studies are: (i) the "Xenarthra root" [24], which places the xenarthrans (i.e. armadillos, the anteaters, the tree sloths etc.) as the sister group to all the remaining placentals, (ii) the "Afrotheria root" [25, 26], which places the Afrotheria (i.e. sea cows, manatees, aardvarks etc.) as the sister group to all the remaining placentals, (iii) the "Atlantogenata root" [27-29] suggesting that the sister group to the all the remaining placentals is is a clade comprising Afrotherian and the Xenarthrans. Further hypotheses that have historically been suggested include, for example (iv) the "hedgehog-1 root" placing the hedgehog (a Laurasiatherian) as the sister group of all the other placentals [30], (v) "hedgehog-2 root", placing the hedgehog as the sister group of all the placentals followed by the rodents [31], and (vi) the "murids root" placing the mouse and the rat

as the sister group of all the other placentals, and often finding the other rodents as a paraphyletic assemblage (e.g. [32], Fig.1A-F). Signals for the topologies in Fig. 1A-B, and to a lesser extent Fig. 1C, have been identified in many mammalian genes [26]. The fact that many different genes support different sets of relationships has resulted in a strong (still unresolved) debate about the correct placement of the root of the placental tree (contrast [24, 26, 29]). On the contrary, signal for the trees in Fig. 1D-F is scant and these topologies most likely represent tree reconstruction artefacts (e.g. model misspecification [33], signal saturation [34], and long branch attraction [34, 35]).

We decided to present an exemplar phylogenomic study of the mammalian relationships to illustrate our supertree software because, based on current knowledge, we can make predictions about what results to expect from our analyses and investigate whether the actualised outcomes from our software deviate from our expectations. More precisely, based on the results of [26] we expect that: (1) either the Afrotheria (fig. 1A) or the Atlantogenata (Fig. 1B) hypotheses will emerge in our optimal ML supertree (most genes in mammalian genomes support one of these two topologies). (2) Similarly, a bootstrap majority rule consensus tree will most likely display one of the two above-mentioned hypotheses (Fig. 1A or B). However, (3) as many genes are known to support both the topologies in Figs 1A-B (and to a lesser extent the tree in Fig. 1C), bootstrap support for the basal placental split in the optimal ML supertree (and in the bootstrap consensus tree) are expected to be low. (4) Tests of two trees are not expected to be able to differentiate significantly between the topologies in Fig 1a-b. Indeed, given the results of [26] we can confidently predict that the trees in Fig. 1A and 1B should be the first and second best fitting hypotheses, even though we cannot predict what their relative order will be (i.e. whether the tree

in Fig. 1A or in Fig. 1B will be the best fitting one). Similarly, (5) whilst we cannot predict whether the Xenarthra hypothesis of Fig. 1C will be significantly rejected by the Approximately Unbiased (or by another) test (e.g. Kishino-Hasegawa test), we can predict that this hypothesis should emerge as the third best one (see [26]). Finally, although we cannot make predictions about how the trees in Fig 1D-F will be ranked, given what is known of the distribution of the signal in mammal gene trees [26], we would expect all these hypotheses to be significantly rejected by the data and to emerge as the three hypotheses that worst fit our data.

To reconstruct our ML supertree of the placental mammals the gene-trees dataset of [9] was employed. This gene-trees data set was pruned to exclude irrelevant taxa using Clann [36]. Only 6 placentals (human, mouse, cat, hedgehog, elephant and armadillo) and one marsupial (the opossum) were retained. This meant that the dataset was reduced from 42 taxa overlapping on 2216 gene trees to 7 taxa overlapping on 389 gene trees (with the gene trees being partially overlapping and containing between 4 and 7 taxa).

Result and Discussion

L.U.St was used to estimate a placental ML supertree. The ML analysis was run for ten iterations with the heuristic search option set to 4 (i.e. using the fastest, least exhaustive, of the search strategies currently available in L.U.St). The pruned MRP supertree from [9] was used as starting tree. The resulting optimal ML supertree supports Afrotheria (Fig.2A). Twenty bootstrapped sets of trees were generated and ML supertree analyses were carried out for each to evaluate support for the inferred relationship of the placental mammals. A majority rule consensus was used to summarise the set of optimal supertrees from the bootstrap analyses and derive

support values for the nodes in the optimal ML tree reported in Fig. 2A. In addition to that we also report the Majority Rule consensus tree (Fig. 2B), which differently from the optimal ML supertree, supports Atlantogenata. As expected (see above) the data provides almost equal support to Afrotheria and Atlantogenata (with the ML supertree supporting Afrotheria even though in the bootstrap replicates Atlantogenata was more frequently recovered). As expected trees representing other alternative hypothesis Xenarthra root (Fig. 1C), murids root (Fig. 1D), and the two hypotheses with a hedgehog root (Figs 1E and F) obtained lower (~6% bootstrap support for the Xenarthra and murid roots hypotheses) or no support (the hypotheses where the hedgehog was the sister group of all the other taxa). L.U.St was then used to estimate, for each one of the 389 input gene-trees, its tree-wise likelihood under each of the six alternative supertree topologies in Fig. 1A-F. The input-tree-wise likelihood scores were then inputted into CONSEL to perform tests of two trees. The results from this analysis (Table 1) show that, as expected, the Approximately Unbiased test was not able to reject any of the three mainstream hypotheses (Afrotheria, Atlantogenata, and Xenarthra-root). Afrotheria emerged as the hypothesis that best fits the data (as expected given that it was represented in our optimal ML supertree), and as expected Xenarthra-root emerged as the third best-fitting hypothesis. Finally, also in this case in agreement with our expectations, all remaining hypotheses (Fig.1D-F) were significantly rejected by the data. Note that the more conservative Shimidoara-Hasegawa test was not able to reject the rodent basal hypothesis of Fig. 1D. However, this test is well known to be over-conservative [22], hence also this result is essentially in line with our expectations.

All results generated were in agreement with our expectations (see above) and apart from confirming that the phylogenetic relationships of the mammals are still far

from being resolved, they illustrate that L.U.St behave as expected and return results that reflect well current understanding of mammal evolution. Overall this illustrates that L.U.St will represent a useful tool in phylogenomics and supertree reconstruction more broadly.

Conclusions

L.U.St represent the first implementation of a maximum likelihood supertree method. This method calculates approximate ML values and has the advantage of finding a tree that has been suggested might be representative of the median of the set of input trees when the symmetric difference metric is used to calculate the tree-to-tree distance. An added advantage of having an approximate ML supertree implementation is that it allows performing statistical test on trees to choose between alternative hypotheses. The results obtained with our toy example reflect current knowledge of mammalian evolution and confirm that the L.U.St package behaves as expected when used to attempt resolving a phylogenetic problem that is well known to be difficult. Being a freely available package for the Python programming environment, L.U.St is both flexible and platform-independent while also being user friendly and easy to implement.

Table 1: Results of the test of two trees. Hypotheses tested are those from Fig. 1.

Hypotheses	Approximate Likelihoods	Ranks	AU test	SH test	KH test
Afrotheria root	-487.092	1	0.628	0.886	0.579
Atlantogenata root	-487.960	2	0.496	0.874	0.421
Xenarthra root	-493.172	3	0.128	0.614	0.146
Muridae root	-523.573	4	0.001	0.017	0.003
Erinaceous root 1	-568.739	5	9E-08	0	0
Erinaceous root 2	-586.111	6	1E-07	0	0

Figure Captions:

Figure 1 The six compared mammal phylogenies. (A) Afrotheria root; (B) Atlantogenata root; (C) Xenarthra root; (D) Rodentia root; (E) Hedgehog root hypothesis of [31]; (F) Hedgehog root hypothesis of [30].

Figure 2 Results of supertree analyses. (A) Maximum likelihood supertree of the placental mammals. (B) Bootstrap Majority Rule Consensus Supertree.

Availability and Requirements

Project name: L.U.St

Project home page: https://afro-juju@bitbucket.org/afro-juju/l.u.st.git

Operating system(s): Linux

Programming language: Python

Other requirements: Consel

License: GNU GPL

Competing interests

There were are no conflicting interests

Authors' contribution

WAA and CJC implemented the software while WAA and DP conducted the experiments and WAA, DP and MW wrote the manuscript.

Acknowledgements

This project was made possible by funding received from the Irish Research council and the UK Biotechnology and Biological Sciences Research Council (grant BB/K007440/1). It was also partially supported by the computing resources at the National University of Ireland, Maynooth and the University of Bristol, UK. DP was supported by a Science Foundation Ireland Grant SFI-RFP 11/RFP/EOB/3106.

Authors

Wasiu A. Akanni[1,2], Christopher J. Creevey[3], Mark Wilkinson[2] and Davide Pisani[1,4]

1) *Department of Biology, The National University of Ireland, Maynooth. Maynooth, Kildare, Ireland.*

2) *Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK*

3) Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Aberystwyth, Ceredigion, SY23 3FG, UK.

4) School of Biological Sciences and School of Earth Sciences, The University of Bristol. Woodland Road, BS8 1UG, Bristol, UK.

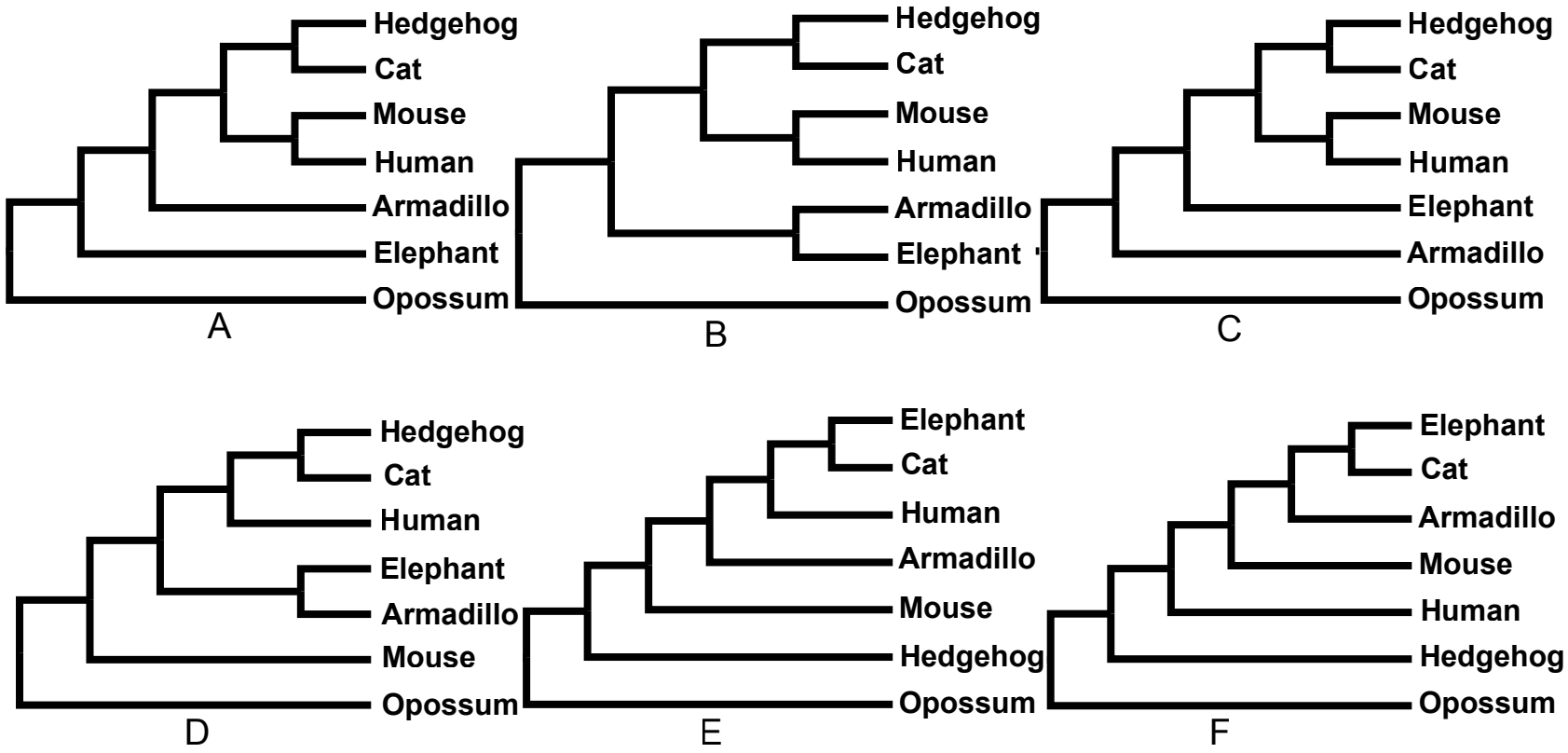
References

1. Semple C, Steel M: **A supertree method for rooted trees.** *Discrete Applied Mathematics* 2000, **105:**147-158.
2. Gordon AD: **Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves.** *Journal of classification* 1986, **3:**335-348.
3. Aho AV, Sagiv Y, Szymanski TG, Ullman JD: **Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions.** *SIAM Journal on Computing* 1981, **10:**405-421.
4. Purvis A: **A modification to Baum and Ragan's method for combining phylogenetic trees.** *Systematic Biology* 1995, **44:**251-255.
5. Daubin V, Gouy M, Perriere G: **A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history.** *Genome Research* 2002, **12:**1080-1090.
6. Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, Travers SA, Wilkinson M, McInerney JO: **Does a tree–like phylogeny only exist at the tips in the prokaryotes?** *Proceedings of the Royal Society of London Series B: Biological Sciences* 2004, **271:**2551-2558.
7. Fitzpatrick DA, Logue ME, Stajich JE, Butler G: **A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis.** *BMC evolutionary biology* 2006, **6:**99.
8. Pisani D, Cotton JA, McInerney JO: **Supertrees disentangle the chimerical origin of eukaryotic genomes.** *Molecular biology and evolution* 2007, **24:**1752-1760.
9. Holton TA, Pisani D: **Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm.** *Genome biology and evolution* 2010, **2:**310.
10. Baum BR: **Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees.** *Taxon* 1992**:**3-10.
11. Ragan MA: **Phylogenetic inference based on matrix representation of trees.** *Molecular phylogenetics and evolution* 1992, **1:**53-58.

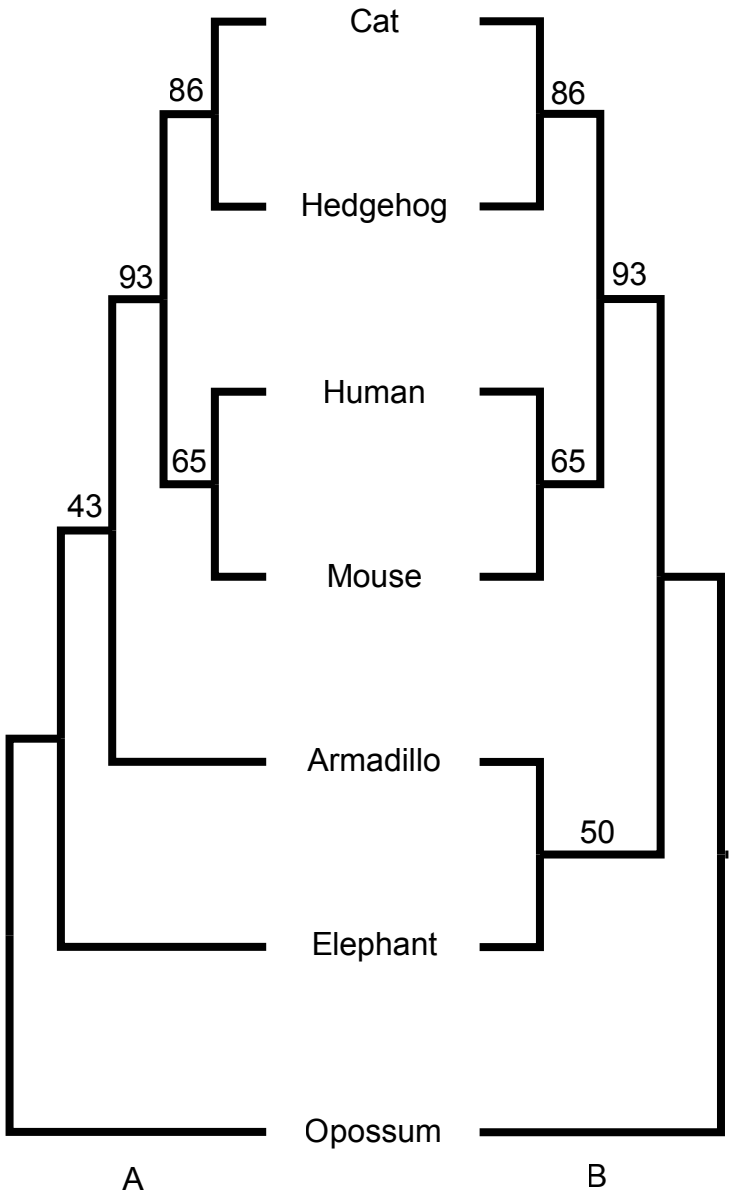
12. Wilkinson M, Cotton JA, Lapointe F-J, Pisani D: **Properties of supertree methods in the consensus setting.** *Systematic Biology* 2007, **56:**330-337.
13. Lapointe F-J, Wilkinson M, Bryant D: **Matrix representations with parsimony or with distances: two sides of the same coin?** *Systematic Biology* 2003, **52:**865-868.
14. Gatesy J, Springer MS: **A critique of matrix representation with parsimony supertrees.** In *Phylogenetic Supertrees.* Springer; 2004: 369-388
15. Wilkinson M, Thorley JL, Pisani DE, Lapointe F-J, McInerney JO: **Some desiderata for liberal supertrees.** *Phylogenetic supertrees: combining information to reveal the Tree of Life* 2004**:**564.
16. Cotton JA, Wilkinson M: **Majority-rule supertrees.** *Systematic biology* 2007, **56:**445-452.
17. Steel M, Rodrigo A: **Maximum likelihood supertrees.** *Systematic Biology* 2008, **57:**243-250.
18. Bryant D, Steel M: **Computing the distribution of a tree metric.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2009, **6:**420-426.
19. Robinson D, Foulds LR: **Comparison of phylogenetic trees.** *Mathematical Biosciences* 1981, **53:**131-147.
20. Kishino H, Hasegawa M: **Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea.** *Journal of molecular evolution* 1989, **29:**170-179.
21. Shimodaira H, Hasegawa M: **Multiple comparisons of log-likelihoods with applications to phylogenetic inference.** *Molecular biology and evolution* 1999, **16:**1114-1116.
22. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Systematic Biology* 2002, **51:**492-508.
23. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17:**1246-1247.
24. O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo Z-X, Meng J: **The placental mammal ancestor and the post–K-Pg radiation of placentals.** *Science* 2013, **339:**662-667.
25. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC: **Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis.** *Genome Research* 2012, **22:**746-754.
26. Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ: **Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals.** *Molecular biology and evolution* 2013, **30:**2134-2144.
27. Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T: **Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification.** *Science* 2011, **334:**521-524.
28. Song S, Liu L, Edwards SV, Wu S: **Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model.** *Proceedings of the National Academy of Sciences* 2012, **109:**14942-14947.

29. Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ: **Heterogeneous models place the root of the placental mammal phylogeny.** *Molecular biology and evolution* 2013, **30:**2145-2156.
30. Cao Y, Fujiwara M, Nikaido M, Okada N, Hasegawa M: **Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data.** *Gene* 2000, **259:**149-158.
31. Corneli PS, Ward RH: **Mitochondrial genes and mammalian phylogenies: increasing the reliability of branch length estimation.** *Molecular biology and evolution* 2000, **17:**224-234.
32. Misawa K, Nei M: **Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees.** *Journal of molecular evolution* 2003, **57:**S290-S296.
33. Foster PG, Cox CJ, Embley TM: **The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 2009, **364:**2197-2207.
34. Rota-Stabelli O, Lartillot N, Philippe H, Pisani D: **Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study.** *Systematic Biology* 2013, **62:**121-133.
35. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ: **A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata.** *Proceedings of the Royal Society B: Biological Sciences* 2011, **278:**298-306.
36. Creevey CJ, McInerney JO: **Clann: investigating phylogenetic information through supertree analyses.** *Bioinformatics* 2005, **21:**390-392.

Akanni et al. Fig.1



Akanni et al. Fig.2



3 An overview of arthropod genomics, mitogenomic, and the evolutionary origins of the Arthropod proteome.

Davide Pisani, Wasiu A. Akanni, Robert Carton, Lahcen I. Campbell, Eoin Mulville and Rota Stabelli Omar

3.1 Introduction

Arthropods represent the largest majority of animal biodiversity and include organisms of economic interest and key model species. It is thus unsurprising that the genome of an arthropod, the fruit fly *Drosophila melanogaster*, was among the very first ones to be sequenced (Adams et al. 2000) and that to date, about 21 *Drosophila* genomes, as well as a variety of other arthropod genomes have been sequenced. Despite this promising start, current sampling is biased toward economically relevant species, and a suitable close outgroup to the arthropods, which is necessary to polarise genomic studies, is still missing. Among the suitable outgorups to the Arthropoda, the Nematoda represent the largest majority of the extant animal biomass, and their economic importance is comparable to that of the more biodiverse arthropods. As with the Arthropoda, the importance of the nematodes is reflected in the fact that the very first animal genome to be sequenced was that of the nematode *Caenorhabditis elegans* (C. elegans sequencing consortium 1998). Despite the nematodes are phylogenetically close to the arthropods (Holton and Pisani 2010), this group is composed of highly derived species, both genetically and morphologically. Accordingly, their genomes are unlikely to be of great utility to understand arthropod genome evolution. Some genomic data (mostly in the form of transcriptomes) is now becoming available for other minor ecdysozoan phyla, and some genomes (Priapulida and Tradigradea) are on the horizon. Nonetheless, enough genomic information is now available for the Arthropoda (Table 3.1) that justifies an investigation of the evolution of their genome. Such an analysis, however, is intimately dependent the availability of a robust phylogenetic

background, and to a lower extent, robust divergence times for the nodes in the background phylogeny.

In this chapter we present an overview of arthropods mitochondrial genomics (section 3.2) and nuclear genomics (section 3.3). We then exploit the available genomic information to investigate the evolutionary origin of novel proteins (orphan gene families) in the arthropod proteome (section 3.4). We notably present the first genomic-scale data set for the Onychophora and include it in our analyses to be able to consider the closest sister group of the Arthropoda (see Campbell et al. 2011) when identifying orphan gene families. Inclusion of new data for the Onychophora is key to this study as it allows the correct identification of the orphan protein families that arose in the stem arthropod lineage.

3.2 Arthropod mitogenomes: useful, but hazardous small genomes

Each cell contains up to hundreds of mitochondria, and each mitochondrion possess many copies of their own small, typically circular, genome (usually named mitogenome or mtDNA). Therefore, mitochondrial genes largely outnumber the nuclear ones in terms of their copy number by several orders of magnitude, making mitochondrial genes easy to extract and to amplify. This is one of the conceptual advantages at the base of the success of mtDNA in phylogenetics both in Arthropoda and across Bilateria. Other reasons behind the fortune of mtDNA are: a conserved gene set, the unambiguous orthology of genes, the presence of rare genetic changes, and the availability of universal primers for many lineages. Other characteristic of

the mitogenome, however, makes it a doubled edged sword. These are: accelerated mutation rate due to uniparental inheritance, and severe biases in the composition of nucleotides that are often responsible for the dilution of the phylogenetic signal in mtDNA (Bernt et al. 2012). In this section we review some of these aspects.

3.2.1 Mitogenomic studies

Mitogenomic studies have helped throughout the 90's and 00's to elucidate some of the arthropod affinities. For example, the first robust evidences in support of the Pancrustacea came from mtDNA gene order comparisons (Boore et al. 1998) and mtDNA sequence phylogeny (Hwang et al. 2001). However, in some cases, mitogenomic studies have pointed toward likely incorrect topologies, for example suggesting a Myriapoda plus Chelicerata grouping (Hwang et al. 2001; Negrisolo et al. 2004; Pisani et al. 2004), which has been shown in recent studies to represent a long branch attraction artefact (Campbell et al. 2011; Pisani 2004; Rota-Stabelli et al. 2011; Rota-Stabelli et al. 2010; Rota-Stabelli and Telford 2008). This is a systematic error that was probably exacerbated by the use of distant outgroups and compositionally biased taxa (Rota-Stabelli and Telford 2008). Such features of the mitochondrial genomes may seriously affect phylogenetic reconstruction unless they are taken into account when inferring phylogenies (Rota-Stabelli et al. 2010).

Utility of the mitochondrial genomes is not restricted to phylogeny. The most widely used arthropod barcode is a region of approximately 650 nucleotides of the subunit 1 of the Cytochrome Oxidase complex (COX1) – a mitochondrial gene. Other

mitochondrial genes (NADH4 for example) are occasionally added to COX1 to improve resolution. This is because every mtDNA gene evolves at a different rate depending on structural constrains, while COX1 is the slower evolving of the mitochondrial genes. A possible risk with mtDNA-based barcoding is the amplification of pesudogenes numts (nuclear copies of mitochondrial genes), which may disrupt barcoding studies.

To date, more than 300 complete arthropod mitochondrial genomes and millions of partial sequences have been deposited in data banks. The taxonomic sampling is however extremely biased toward economically relevant species: 47 chelicerates (mostly acari), 53 crustaceans (mostly malacostracans), 198 insects (mosly coleopterans, dipterans and hemipterans), and only 9 myriapods. Still, most major orders and classes are now represented, thus providing an invaluable starting point for comparative analyses.

3.2.2 The structure of the arthropod mitochondrial genome

Arthropod' s mtDNA varies in size from less than 14000 bp in the spider *Ornithoctonus huwena* to more than 19000 bp in *Drosophila melanogaster*. This difference is almost entirely due to non-coding inter-genic regions, particularly the Major non-coding region commonly called "control region". Due to its low structural constrain and high tendency to accumulate A and T nucleotides, this region is also called the "AT-rich region". The AT-rich region is involved in both the replicative and transcriptional processes, and typically contains structural elements

like hairpin-loops and timidine stretches (Zhang and Hewitt 1997), elements that do not seem to be conserved throughout the arthropods.

The gene content of the arthropod mtDNA is the same as in most other bilaterian; it typically consists of 13 coding genes, 2 ribosomal RNA subunits, and 20 t-RNAs (Boore 1999). This gene set is highly conserved throughout the phylum, although few exceptions can be found. Examples include a tRNA-Ser duplication in *Thrips imaginis* (Shao and Barker 2003), a tRNA-His duplication in *Speleonectes tulumensis* (Lavrov et al. 2004) and a tRNA-Cys triplication in *Pollicipes polymerus* (Lavrov et al. 2004). Many arthropod mitochondrial coding genes lack a stop codon (TAA or TAG), and possess a single T or TA at the 3-terminal end. The correct stop codon is then assembled by the poly-adenilation of an excised, presumedly policistronic, transcript. Although most arthropod mitogenomes use the invertebrate genetic code, it has been shown that some lineages use a slightly different code (Abascal et al. 2006). Remarkably, this new genetic code is scattered throughout the arthropod tree.

Although the gene content is conserved throughout the arthropods, the gene order may vary significantly (Lavrov et al. 2004). Comparative studies have determined an arthropod ancestral gene order, which is represented (retained) by *Limulus polyphemus*, while the pancrustaceans gene order differs from that of all the other arthropods by the position of one of the two leucine t-RNAs. T-RNAs in general are mostly responsible for variation in gene order as they are hotspots of recombination. Less often, coding genes change their position or swap strand,

allowing for variation in gene specific strand asymmetry, as detailed in the next paragraphs.

3.2.3 Arthropod mitogenomes: a composition nightmare

The main source of compositional heterogeneity in mtDNA is mutational pressure, which is correlated with a deficiency in the mtDNA repair system and with a consequent inefficiency at replacing erroneous insertions of A nucleotides (Reyes et al. 1998). Compared to other metazoans, arthropod lineages are typically enriched in A and T. In the absence of strong purifying selection, this mutational pressure affects also encoded proteins, which are enriched in amino acids encoded by A+T rich codons (Foster and Hickey 1999; Foster et al. 1997; Rota-Stabelli et al. 2010). The effect of this mutational pressure depends on structural constraints acting on the genes: more conserved genes such as Cox1 accumulate less A+T mutations than poorly constrained genes such as Atp8. In addition, not all positions of a gene are affected in a similar way: while the 1st and 2nd codon positions are more constrained by the genetic code, the 3rd codon positions are more prone to accumulate A+T mutations and experience saturation of replacement events (Figure 3.1a). Interestingly, 1st codon positions show a different A+T replacement pattern than the 2nd. This advocates the employment of different model of evolution for the 1st and 2nd codon positions and the exclusion of the 3rd codon positions when performing phylogenetic reconstruction from nucleotide sequences. This would, at least

partially, obviate for possible artifactual attraction in the case that unrelated species have a similarly increased A+T content.

The A+T content is not homogenously distributed along the arthropod phylogeny: some groups such as Pycnogonida, the Acari, and some insects are more A+T rich than other lineages (Figure 3.1b). This uneven distribution of nucleotide content may have been responsible for the artifactual attraction of for example Acari and Pycnogonida (Podsiadlowski and Braband 2006). In some species such as bees and the green-bug (grey dots in figure 3.1b) the A+T content reaches extremely high values, the highest ever reported for eukaryotic coding genes.

Strand asymmetry is another type of compositional heterogeneity affecting mtDNA. This bias is related with the origin and direction of mtDNA replication (Reyes et al. 1998), and leads one strand to become enriched in G (and to a lesser extent in T) while the other strand become enriched in C (and less in A). Strand asymmetry is generally expressed in terms of GC-skew. Although all genes in a mitochondrial genome usually have a similar A+T content, homologous genes from different organisms may have extremely different, sometimes opposite, GC (and AT) skew: this depends on the strand on which the gene is located, and on its position relative to the origin of replication (Lavrov et al. 2000). Therefore there is a link between strand asymmetry and gene order.

In arthropods most mtDNA coding genes are charcaterised by a negative GC-skew (they have more C than G), while four genes that lies on the opposite strand are characterised by a positive GC-Skew. This situation is characteristic, in

particular, of species characterised by the arthropod ancestral gene order (as in Figure 3.2a). In some species, the GC-skew is opposite for all the genes, although the gene order is substantially identical to that of the ancestral arthropods (Figure 3.2b). In such cases, it is the origin of replication (the control region) that underwent a modification, for example a duplication or an inversion of strand. In other cases all genes may have been translocated on the same strand, so that all the genes posses either a positive or a negative GC-skew (Figure 3.2c).

3.2.4 The hazard of using arthropod mitochondrial genomes for phylogenetics

It has been shown that both sources of compositional heterogeneity (A+T mutational pressure, and strand asymmetry) may play strong roles in generating artefactual mitogenomic phylogenies (Hassanin et al. 2005; Rota-Stabelli et al. 2010). Compositional problems are worsened by the accelerated rate of evolution of mitogenomic sequences, which is related to the uniparental inheritance characterizing the mitochondria. An effective approach to deal with these problems is to improve models of mitochondrial sequence evolution both at the nucleotide (Hassanin et al. 2005) and protein level (Abascal et al. 2007; Rota-Stabelli et al. 2009), as well as to exclude more affected genes or codon positions. Sophisticated evolutionary models which account for among site and among branches heterogeneity (Blanquart and Lartillot 2008; Foster 2004) are useful to lessen the effects of these mitochondrial compositional biases. Another obvious approach is to enlarge or modify taxonomic sampling. More taxa may break problematic branches

and reduce the number of homoplasies responsible for long branch (or compositional) attractions. Exclusion of problematic taxa may results in the same beneficial effect. It is therefore advisable to conduct an explorative compositional analysis of the properties of the considered mitochondrial genomes prior to phylogenetic inference. This is particularly true for the arthropods, which include some highly derived lineages, parasites for example, whose particular life style is responsible for bottle-neck events and therefore extreme acceleration of substitution rates or divergent nucleotide compositions.

Compositional biases (and related phylogenetic artifacts) have been primarily studied using mitogenomics datasets (Foster et al. 1997). The advent of the phylogenomic-type (nuclear) datasets has been initially seen as a relief in terms of compositionally related biases. This may however not be the case: the community is just noticing that even large genomic datasets are not free from compositional problems that can cause serious phylogenetic artefacts (Nabholz et al. 2012; Stabelli et al. 2012). Still, the origins of such biases in nuclear genomic data are largely not known.

3.3 Arthropod comparative genomics

The study of arthropod genomics started with the sequencing of the genome of the Fruit fly *Drosophila melanogaster* (Adams et al. 2000). Currently, genomic data are available for a relatively large number of arthropods allowing the first attempts at performing comparative genomic analyses of the Arthropoda (Vieira

and Rozas 2011). However, the majority of the currently available arthropod genomes are from closely related species (mostly insects), and a coherent set of conclusions about the arthropod nuclear genomes (as presented for the mitochondrial genomes above) are still lacking.

3.3.1 Uneven taxonomic sampling

The biased taxonomic distribution of the available arthropod genomes is a persisting problem. This is because it does not allow detailed investigations of key questions in arthropod evolution, like the origin of the arthropod subphyla. Initiatives exist that aim at increasing the amount of available genomic information for the Arthropoda. Paramount among these projects are the 1KITE project – 1000 Insects Transcriptome Evolution project (http://1kite.org/), and the i5K (http://www.arthropodgenomes.org/wiki/i5K) project which plans to sequence the complete genomes of 5000 insects and related arthropod species. Unfortunately, as commendable as these projects are, they fall short of adequately capturing the breadth of the evolutionary diversity within the Arthropoda. The 1KITE project will not even attempt to generate data for non-hexapod species, whilst about 87% of the species currently nominated for sequencing as part of the i5K project are hexapods. Only 0.7% belongs to Myriapoda and only 2.8% belong to the Crustacea. This is an important issue with the current initiatives, as the heterogeneous species sampling discussed above might bias future comparative analyses in unexpected ways.

An important aspect to which current large-scale genome sequencing projects are not given sufficient attention is that of the arthropod outgroups. To increase the power of comparative analyses, adequate outgroups should also be sequenced, but large-scale sapling initiatives are not considering the outgroups of the Arthropoda. Indeed, to date, the only arthropod outgroups available with at the least one fully sequenced genome are the nematodes. Yet species belonging to this phylum are too distantly related and too divergent from the Arthropoda (see also above) to be of significant utility in arthropod comparative genomics. Other more closely related genomes (those of the Onychophora and of the Tardigrada) should be sequenced and used instead. Indeed, as part of this chapter, to obviate the lack of genomic-scale data sets for the arthropod outgroups, we shall present a genome-wide trascriptomic data set obtained using next generation sequencing.

The 1KITE and i5K projects have not produced data yet. However, a relative abundance of arthropod genomes has been accumulating in recent years, albeit with a biased taxonomic distribution. The genomes of 21 *Drosophila* species have been sequenced and are publicly available. Trascriptomic, Proteomic and Genomic data, as well as abundant functional annotations, for 12 of these species can be found in the specialised database Flybase (http://flybase.org/). Other key insects for which genomic information is available include the mosquitoes *Aedes aegypti* (Nene et al. 2007) and *Anopheles gambiae* (Holt et al. 2002), the honeybee *Apis Mellifera* (The honeybee sequencing consortium 2006), the beetle *Tribolium castaneum* (Richards et

al. 2008), the body louse *Pediculus humanus* (Kirkness et al. 2010)**,** the pea aphid *Acyrthosiphon pisum* (The pea aphid sequencing consortium 2010), and the butterfly *Bombyx mori* (The silkworm sequencing consortium 2008). A variety of other insects, e.g. ants and other butterflies have also been sequenced (The butterfly sequencing consortium 2012; Suen et al. 2011), and for many of these species taxon-specific databases exist (e.g. Butterflybase – http://butterflybase.ice.mpg.de/*).* However, differently from Flybase, which is a mature database providing, for example, a genome browser, and allowing complex searches (using Gene Ontology – GO terms and developmental stages), most of these species-specific databases are still quite immature. In any case, they represent an important resource and their utility is bound to increase with time.

While hexapod genomes are relatively abundant, the situation changes drastically when moving to other arthropod subphyla. Only one complete crustacean genome (that of the water flea *Daphnia pulex* – Colbourne et al. 2011), and one complete chelicerate genomes, that of the two spotted spider mite *Tetranychus urticae* (Grbic et al. 2011) have been released. Finally, the complete genome of one myriapod, the centipede *Strigamia maritima* (Genbank access id: GCA_000239455.1), and that of a second chelicerate *Ixodes scapularis* (Genbank access id: GCA_000208615.1) are now publicly available, although not yet released.

Apart from standard genomic studies, a variety of large-scale transcriptome-wide sequencing studies have been performed, and EST data are thus available for many taxa. Even though these studies do not provide information about

untranslated genomic regions, a large amount of useful data has been provided using these approaches. One of the earliest studies that employed EST generated next generation sequencing (454 sequencing) to gain a first snapshot of an arthropod genome was the transcriptome sequencing of the emperor scorpion *Pandinus imperator (Roeding et al. 2009)*. More recently similar approaches have started to allow interesting insights into chelicerate venomes (Rendon-Anaya et al. 2012), and allowed the development of the new science of Venomics.

3.3.2 Heterogeneity of genome sizes and shortage of miRNA

Important aspects of the key, publicly available, arthropod genomes are reported in Table 3.1. From this table it is clear that the arthropod genomes are fairly variable. Their lengths in MB vary substantially with one of the chelicerate genomes being the smallest while the other is the biggest overall. Similarly, GC content is quite variable with *Ixodes* having the highest GC content and the pea aphid the lowest. Also the number of predicted genes varies substantially between genomes, with *Daphnia* having more than 30,000 protein-coding genes and *Ixodes* only 5,867 protein coding genes. An obvious observation emerging from an analysis of Table 3.1 is that the sequenced chelicerate taxa cannot be particularly good resources for evolutionary biologists. *Ixodes* and *Tetranychus* are highly specialized species unlikely to reflect what the analysis of more standard chelicerate genomes will uncover.

Next generation sequencing approaches have also allowed our understanding of regulatory (non-coding) microRNA to increase substantially. Genome wide screening performed for taxa belonging to all the arthropod subphyla and to the arthropod outgroups (Campbell et al. 2011; Rota-Stabelli et al. 2011) allowed identification of several arthropod specific microRNA (miR275 and iab-4), mandibulata specific ones (miR-965 and miR-282) and chelicerate specific ones (miR-3931). These studies also showed that arthropods, differently from other lineages (like the mammals or the annelids) have significantly less lineage specific microRNAs, suggesting that arthropod genomes, from this point of view, evolve quite differently from those of other animal lineages.

Overall, current genomic scale information available across the Arthropoda is still too fragmentary to allow the development of a coherent view of arthropod genome evolution. However, in the last section of this chapter, we shall attempt to start obviating to this problem, by presenting an evolutionary analysis of the arthropod proteomes that exploits the transcriptomic data we generated for the Onychophora.

3.4. A genomic phylostratigraphic analysis of the arthropod proteomes.

An interesting aspect of the arthropod genome evolution that availability of current metazoan and arthropod genomes allow us to address (given also the data we generated for the Onychophora) is that of the origin of the Arthropod specific protein coding genes (i.e. genes found only within Arthropoda). Studies of this type

have been named “genomic phylostratigraphic analyses” by Domazet-Loso et al. (2007). To complete such studies (in addition to genomic information) one needs information about phylogeny and divergence times. The relationships among the arthropods and divergence times used are summarised below.

3.4.1 A robust phylogenetic framework for genomic studies

Comparative genomics must be anchored on a phylogenetic tree. Significant progresses in our understanding of the ecdysozoan relationships have been made (Campbell et al. 2011). Similarly, some agreement on the phylogenetic relationships within the Arthropoda has recently emerged (Regier et al. 2010; Rota-Stabelli et al. 2011) but see Rota-Stabelli et al. (2012). With reference to the current study we shall consider the Lobopodia to be the sister group of the Tardigrada within a monophyletic Panarthropoda. We shall further assume Nematoida (Nematoda plus Nematomorpha) to be the sister group of Panarthropoda, with the Scalidophora (here Priapulida and Kinorincha) representing the sister group of Nematoida plus Panarthropda. That is, we shall assume the phylogenetic relationships of the Ecdysozoa inferred by Campbell et al. (2011) to represent our working hypothesis. Campbell et al. (2011) only performed a Bayesian analysis of their data set and did not present bootstrap support for their results. Given that they did not find particularly strong support (low posterior probabilities) for some key ecdysozoan nodes (Nematoida + Panarthropoda and Mandibulata), and given that there still are few studies (e.g. Meusemann et al. 2010) whose results contradict those of Campbell

et al. (2011), we present here a novel statistical analysis – a non parametric bootstrapping – of the dataset used in Campbell et al, (2011), or which a detailed explanation is given in the Appendix to this chapter.

Results of the bootstrap analysis that consider all the taxa in Campbell et al. (2011) is in agreement with the Bayesian analyses of Campbell et al. (2011). This analysis shows a lack of support for many important nodes, including Nematoida (which was not recovered), Nematoda plus Arthropoda (BP = 41), Panarthropoda (BP = 66), Lobopodia (BP = 61), and Mandibulata (BP = 64), see Fig. 3.3. We performed a leaf stability analysis (results not shown – but see appendix) illustrating that Nematomorpha is the most unstable taxon in the data set. The nematomorph in Campbell et al. (2011) emerged as the sister group of the Nematoda in agreement with most previous studies. Yet, in Fig. 3.3 Nematomorpha is not the sister group of the Nematoda. Instead, it emerges as the sister of a Nematoda + Arthropoda clade. This is an artefact caused by high number of missing data in the Nematomorpha (which is the most incomplete taxon in Campbell et al. 2011), and that is unstable in bootstrapped data sets. Upon removal of the unstable Nematomorpha, the bootstrap support for all the other nodes increases significantly. Arthropoda plus Nematoda reach 100%, Panarthropoda increase to 76%, and Lobopodia to 70%. In conclusion, when accounting for unstable taxa, Arthropoda has a bootstrap support of 100% and Mandibulata of 76%. This confirms that there is a good level of support for the clades in Fig.3.3 and those in Campbell et al. (2011).

3.4.2 Expanding our understanding of the arthropod comparative genomics

Given our poor understanding of the processes through which the arthropod (nuclear) genomes evolved, we shall here present a genomic Phylostratigraphic analysis (Domazet-Loso et al. 2007) of their genome. The aim of this analysis is gaining some information on the evolutionary processes responsible for the origin and evolution of the Arthropoda. Domazet-Loso et al. (2007) performed a similar analysis, but various new genomes have been published since their study, allowing for a much greater precision in the identification of orphan genes along the Ecdysozoan and Arthropod phylogeny. To better identify proteins that are arthropod specific, we extended our analyses to include a variety of ecdysozoans and non-ecdysozoan genomes. Particularly we included representatives of the Lophotrochozoa, of the Deuterostomia and two non-bilaterian metazoans (a sponge – *Amphimedon queenslandica*, and a cnidarian – *Hydra magnipapillata*) – see Fig. 3.4. In addition, and most importantly, here we added data for an onychophoran transcriptome, which allowed pinpointing protein families that are specific to the Arthropoda (i.e. that originated after the Onychophora-Arthropoda split). Finally, more reliable molecular clock divergence times (Erwin et al, 2011) are now available and they have been used here to define rates of orphan gene acquisitions through time allowing for a better estimation of rates of new protein family acquisitions in Ecdysozoa and Arthropoda.

3.4.3 The evolution of orphan gene families in arthropoda

We used the MCL algorithm (Enright et al, 2002) to identify protein families in the set of considered genomes, and identified, for each internal node in Fig. 3.4, all the proteins universally distributed in the taxa descending from each given node. These are orphan families that evolved in the branch underlying the considered node. The average number of new families acquired across all the internodes of the considered phylogeny is 1025. When this value is normalised (dividing by the total number of proteins in the considered set of genomes (79,052 protein coding genes), the 1025 protein families that are gained as novel orphan genes correspond to ~ 1.2%.

Within Arthropoda, and more broadly Panarthropoda, only the origin of the Diptera (with 2.05% of new protein families being aquired), show a statistically significant rate of novel gene families acquisition (Fig. 3.4 and 3.5). Genomic data were not available for the Myriapoda when we assembled our data set, but it is clear, given the low level of proteins that originated in the branch separating Arthropoda and Pancrustacea (1.49%), that also the origin of Mandibulata cannot be marked by a spike in the origin of new protein families (Fig. 3.4 and 3.5).

The most surprising result emerging from this analysis is that the deepest nodes in the Ecdysozoan phylogeny: (origin of Nematoida plus Arthropoda, origin of Lobopodia, and origin of Arthropoda) are not characterised by above than average acquisitions of new genes families (Fig. 3.5). When the number of orphan families (N-orph) acquired along a branch is divided by the length (in millions of years) of

the branch along which the N-orph accumulated, the pattern in Fig. 3.5a change quite significantly: even the mild, but somewhat continuous, increment in the rate of N-orph acquisition disappears (Fig. 3.5b). All internodes within Ecdysozoa (on the path leading to Arthropoda and within Arthropoda) roughly exhibit the same rate of new protein acquisition per million of year. Constancy of the rate of protein family acquisition through time (from the Precambrian to the Jurassic – see Fig. 3.5b) suggests that this rate (identified with a red line in Fig. 3.5b) might represent the neutral background rate of new protein family origination in Ecdysozoa. The only internode where this neutral rate is modified is represented by the stem dipteran lineage. Along this lineage (Fig. 3.5b) the rate is significantly increased suggesting that orphan gene family acquisition was an important phenomenon in the evolution of this group.

A functional analysis of the orphan proteins that originated along the stem dipterian lineage (see Appendix for methodological details) provide a view of what kind of gene families are acquired along this branch (Fig. 3.6). When comparing the average trend estimated across all the considered stem lineages but the dipteran, with the trend observed in the dipteran, two conclusions can be reached. The first is that the trends observed are comparable in shape (i.e. there is a proportionality in the number of new genes acquired on average across the Arthropoda and specifically in Diptera). The second is that when the numbers of genes in each GO category is analysed, it is clear that for two GO terms (Metabolic processes and Cellular Processes) the increase observed in Diptera is significantly higher (greater than the limiting values of a 99% confidence interval calculated across all the other

internodes – Fig. 3.6). A further significantly increased category (exceeding the 95% confidence interval calculated across all the other – non-dipteran internodes) is the Localization proteins category. Finally other GO categories for which new proteins are accumulated in Diptera to levels that are above average (but not significantly so) are: Biological Regulation, Response to stimulus, multicellular organismal processes, signalling, developmental processes, and cellular component organisation.

3.4.4 conserved rate of gene gain with some surprises

It is fairly obvious from the above results that, at the least within Ecdysozoa, the origin of new protein families (orphan genes accumulation) did not play a particularly significant role in the evolution of what we recognise as high level, taxonomic groups (Phyla and assemblages of phyla). In particular, we have shown here that the origin of the arthropod body plan was not characterised by an unusual rate of new protein families acquisition. One can thus argue that other processes, like the re-wiring of developmental networks (and more generally protein-protein interaction networks), might have been much more important (see also Erwin et al. (2011). Yet, these hypotheses needs to be tested, and will be tested in the future when more data will be available.

On the other hand the origin of the Diptera is markedly signed by a substantial increase in the origin of orphan families. This is interesting because it suggests that (1) if increases in rate existed somewhere else in the ecdysozoan tree we should have been able to identify them (i.e. our results do not seem to represent

a methodological artifact), and (2) orphan gene acquisition is not always an unimportant process in animal evolution: hence the need to investigate it. With reference to the Diptera, it is clear that the strong acceleration in rate of new families acquisition observed implies that, new functionalities emerged in this part of the ecdysozoan tree, and it is clear that these protein families played a role in the origin of this group. Our current GO analyses did not allow us to obtain a detailed description of what the newly acquired dipteran functions are. However, as more data will become available, more precise results will be possible to be derived. One can only conjecture, given also the unimpressive amount of orphan families being fixed on the Holometabola-stem lineage, that the origin of key innovations affecting the emergence of novel life cycles or substantially modified morphological features is generally fuelled by re-wiring of the developmental networks and by differential expressions of genes, whilst origin of novel protein families probably has a greater impact on adaptations to novel environmental challenges.

3.5. Conclusions

Here we have tried to summarise available mitogenomic and nuclear genomic information currently available for the Arthropoda. There are a large number of mitochondrial genomes available to date but it is unclear if something that will be of any utility will be gained from the analyses of these genomes. They might have some limited utility in phylogenetics compositional bias studies, and barcoding, but

probably not much utility to understand large-scale evolutionary patterns in Arthropoda.

Arthropod genomics, on the other end is still in its infancy, very few genomes are available at this stage but within five years we will probably have thousands of genes available (particularly thanks to large scale efforts like the i5k). One wonders what will be gained from having so many genomes. Perhaps a lot, but their biased taxonomic distribution might prove to be a limit of these data sets. Data analysis will be prohibitively complex and serious bioinformatic recourses will be necessary for these data to be of any utility. In any case, the initial analysis we present in this chapter suggests that, if adequate bioinformatic resources will be available, availability of a multitude of arthropod genomes will allow gaining detailed information on the origin and evolution of this important phylum. Yet, sequencing projects should not forget that arthropod outgroups are necessary and important to increase the power of comparative analyses.

No matter what the future will reserve, it is clear that Arthropod comparative genomics is still in its infancy. We are just at the dawn of what will be a laborious and complex research task which will involve the continuous effort of many research groups, from all around the world for, probably, several research cycles.

FIGURE CAPTIONS:

Figure 3.1: Compositional heterogeneity in arthropods miotgenome. (A) A+T% content of the three codon positions plotted against that calculated on the whole mtDNA. 2nd codon position is the most constrained, while 3rd codon position changes so dramatically that reaches plateau in some species. (B) A+T % calculated on the whole mtDNA in different arthropod lineages. Nucleotide content varies between and within classes.

Figure 3.2: strand asymmetry in arthropod mitogenome. Each gene in the mtDNA is characterised by a different propencity of accumulating mutations toward G or C. This is because different genes lies on different strands and each strand has his own mutational pressure, described here by the GC-Skew statistics. (A) in most arthropods the majority of genes are on the same strand and possess a negative GC skew; the ORF of Nadh4, Nadh5 Nadh4l and Nadh1 are on the opposite strand, and as a consequences these genes accumulate more G and have a positive GC-skew. (B) Some phylogentically unrelated arthropods experienced an inversion of the replicative system, which leads to a complete inversion of GC-skew for each of the genes. (C) Some taxa underwent genomic rearrangement so that all genes are on the same strand.

Figure 3.3: The phylogeny of the Ecdysozoa. The tree represent a Bayesian-bootstrap analysis performed under CAT+G of the data set of Campbell et al. (2011). Values at the nodes represent bootstrap proportions. * = 100% support. The leftmost value represents the bootstrap proportion obtained for a data set including all the sequences in Campbell et al. (2011). The rightmost value represents the

bootstrap proportion obtained when the most unstable taxon in the data set (the nematomorph *Spinochordodes*) was excluded.

Figure 3.4. Orphan protein gains in Arthropoda. The number below each node quantifies the orphan families that evolved along the branch subtending the considered node. The number in black above each node represents the numbers of protein coding genes inferred to have existed (using squared parsimony) in the common ancestor represented by the considered node. The red value above the node represents the rate of orphan gene acquisition along the branch subtending the considered node. These values are normalised (they have been calculated as the number of orphans divided by the total number of proteins in the collection of considered proteomes). The numbers reported for each terminal taxa are: the number of orphan families that originated along the terminal branch, and the number of genes in the genome of the corresponding organism (in bold). Note that the numbers of orphans for the terminal taxa are misleading and should not be considered to represent the number of new genes that emerged in the species at the tip of the tree. Instead they represent the number of orphan in the group the species represent. For example, the number of orphans in *Hydra* represents the orphans that where acquired by the Coelenterata (to which *Hydra* belong and that *Hydra* represents) rather then by Hydra itself.

Figure 3.5. Protein gains through time. This figure represents (A) the normalised rates of orphan acquisition (red values in Fig.3.4). This panel illustrates that the

normalised rates are quite variable across all the considered nodes. Note that the values were ordered from oldest to youngest to make the Figure more readable. (B) rates of orphan acquisition per millions of years. This chart was derived dividing the values in Fig. 3.5A by the length (in million years) of the branch along which the considered orphans originated. This figure clearly illustrates how the row rates and the rates per million year are substantially different, and that normalising for the time of duration of the considered internodes is key to obtain values that are biologically meaningful. The red line represents the average rate across the considered lineages (but excluding the Diptera). This was done to estimate the average rate orphan protein acquisition (i.e. the neutral rate).

Figure 3.6. The function of the newly acquired families. This graph displays the average number of orphans (across all the internodes but the Diptera) for each GO (Gene Ontology)-category. We also reported the values representing, respectively, the limits of the 95% and 99% confidence intervals. Values observed for the dipteran stem lineage are reported. This figure shows that for two GO categories the number of orphan acquired in Diptera is higher than the value bounding the 99% confidence interval over all the other internodes, and that a variety of other GO categories are overrepresented with reference to the other, considered, internal branches.

Appendix: methods for the analyses presented in chapter 3

A. Generation of the Onychophoran transcriptome.

Total RNA was extracted from three single individual of *Peripatoides novaezealandiae* using TriZol©. A cDNA library was constructed and sequenced by the Trinity College Dublin Next Generation Sequencing Facility to an estimated coverage of <100 using 100 paired end reads on two IlluminaHiseqII lanes. Row data was inspected for its quality and the resulting paired-ends reads were assembled using Abyss (Simpson et al. 2009) with k-mer of 45. This resulted in ~27,000 assembled transcripts (with lengths variable between ~ 70 to 1750 base pairs). Approximately 17,000 of these transcripts had a significant blast hit against an annotated gene.

B. Mitogneomic compositional analyses

We downloaded a set of 90 arthropods mitochondrial genomes in order to represent as homogenously as possible the whole phylum. Coding genes were extracted and processed with DAMBE (Xia and Xie 2001) to obtain composition for each codon position.

C. Phylogenetic Analyses

We investigated whether the low posterior probabilities observed for some nodes by Campbell et al. (2011) were caused by the presence of unstable taxa. We estimated leaf stability indices (Thorley and Wilkinson 1999) using P4 (Foster 2004), and performed Bayesian bootstrap analysis (under CAT + G – the same model

used by Campbell et al. (2011), using the entire data set of Campbell et al. (2011). To perform the Bayesian bootstrap analyses, 100 bootstrapped data sets were generated starting from the alignment of Campbell et al. (2011). For each bootstrapped data set a Bayesian analysis (2 independent runs) was performed under CAT +G (using Phylobayes – Lartillot et al. 2009). Results from each Bayesian analysis were summarised to generate a Bayesian majority rule consensus tree, and the resulting 100 trees were then summarised to generate a bootstrap majority rule consensus (results in Fig. 3.3).

Identification of novel gene families

We downloaded the entire proteomes for the taxa in Fig. 3.4, and used MCL (Enright et al, 2002) to define protein families. A Perl script written by LC was used to partition these gene families with reference to their taxon coverage. This allowed identifying protein families that are exclusive and universally distributed within each one of the clades in Fig. 3.4. These protein families must have been present in the clade last common ancestor (LCA), and must have been gained along the stem lineage of the considered clade. Because different genomes have different numbers of protein coding genes, the absolute numbers of newly acquired protein coding families for each internode can be misleading. We thus normalised numbers of orphan families by dividing these numbers by the total number of protein coding genes in the set of considered genomes (sum of the values in bold at the tips of Fig.3.4). The normalised orphans counts (N-orph) can be interpreted as the fraction

of some, abstract, pan-metazoan genome that was acquired at each internode of Fig. 3.4. Finally, we calculated rates of new orphan acquisition per million of years, dividing the N-orph values by the length of the internode along which the N-orph was acquired. As above, this allows the amount of orphan families gained each million year, along each internode in Fig. 3.4, to be expressed as proportions of a reference (abstract) "pan-metazoan" genome. The estimates of divergence times of Erwin et al. (2011) were used to calculate branch durations in million of years. For each internal node in our phylogeny we also estimated (using squared parsimony – as implemented in Mesquite – http://mesquiteproject.org) the expected size of the genome of the corresponding LCA. This was done to allow evaluating what proportion of each LCA genome was gained via new orphan family acquisition, along the corresponding stem lineage. Because Squared Parsimony is unlikely to be a particularly robust estimator or ancestral size we suggest these numbers should be considered with caution, and only to represent a rough approximation of the true LCA-genomes dimensions.

Once the orphan gene families were identified for every internode of Fig.3.4, BLAST2Go (www.blast2go.com) was used to obtain functional information for each of these families. For each protein family, the BLAST2Go analysis was performed for one protein family member only, and we assumed, by homology-implication, that all the other proteins in the same orphan family had the same (or similar) function.

Acknowledgments

We would like to thank Prof. Minelli for inviting us to contribute a chapter to this book and for the patient demonstrated during the editing process. DP and RC are supported by a Science Foundation Ireland Research Frontier Programme (SFI-RFP) grant SFI-RFP 11/RFP/EOB/3106. ORS by a Marie Curie -Trento Province COFUND Fellowship. WAA by an IRCSET PhD studentship.

References

The C elegans genome consortium (1998) Genome sequence of the nematode C elegans: a platform for investigating biology: Science 282:2012-2018.

The honeynee genome consortium (2006) Insights into social insects from the genome of the honeybee Apis mellifera: Nature 443:931-949.

The Heliconium genome consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species: Nature 487:94-98.

The pea aphid genome consortium (2010) Genome sequence of the pea aphid Acyrthosiphon pisum: PLoS Biol 8:e1000313.

The silkworm genome consortium (2008) The genome of a lepidopteran model insect, the silkworm Bombyx mori: Insect Biochem Mol Biol 38:1036-1045.

Abascal F, Posada D, Knight RD, Zardoya, R (2006) Parallel evolution of the genetic code in arthropod mitochondrial genomes: PLoS Biol 4:e127.

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

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ,rews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, et al (2000) The genome sequence of Drosophila melanogaster: Science 287:2185-2195.

Bernt M, Braband A, Middendorf M, Misof B, Rota-Stabelli O, Stadler PF (2012) Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences: Mol Phylogenet Evol Published on-line ahead of press.

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement: Mol Biol Evol 25:842-858.

Boore JL (1999) Animal mitochondrial genomes: Nucleic Acids Res 27:1767-1780.

Boore JL, Lavrov DV, Brown WM (1998) Gene translocation links insects and crustaceans: Nature 392:667-668.

Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D (2011) MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda: Proc Natl Acad Sci U S A 108:15920-15924.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Rews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL (2011) The ecoresponsive genome of Daphnia pulex: Science 331:555-561.

Domazet-Loso T, Brajkovic J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages: Trends Genet 23:533-539.

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families: Nucleic Acids Res 30:1575-1584.

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals: Science 334:1091-1097.

Foster PG (2004) Modeling compositional heterogeneity: Syst Biol 53:485-495.

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions: J Mol Evol 48:284-290.

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria: J Mol Evol 44:282-288.

Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouze P, Grbic V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, Hernandez-Crespo P, Diaz I, Martinez M, Navajas M, Sucena E, Magalhaes S, Nagy L, Pace RM, Djuranovic S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi SV, Zeng J, Pires-daSilva A, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Gregory TR, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van de Peer Y (2011) The genome of Tetranychus urticae reveals herbivorous pest adaptations: Nature 479:487-492.

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

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, et al (2002) The genome sequence of the malaria mosquito Anopheles gambiae: Science 298:129-149.

Holton TA, Pisani D (2010) Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm: Genome Biol Evol 2:310-324.

Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W (2001) Mitochondrial protein phylogeny joins myriapods with chelicerates: Nature 413:154-157.

Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, Gerlach D, Kriventseva EV, Elsik CG, Graur D, Hill CA,

Veenstra JA, Walenz B, Tubio JM, Ribeiro JM, Rozas J, Johnston JS, Reese JT, Popadic A, Tojo M, Raoult D, Reed DL, Tomoyasu Y, Kraus E, Mittapalli O, Margam M, Li HM, Meyer JM, Johnson RM, Romero-Severson J, Vanzee JP, Alvarez-Ponce D, Vieira FG, Aguade M, Guirao-Rico S, Anzola JM, Yoon KS, Strycharz JP, Unger MF, Christley S, Lobo NF, Seufferheld MJ, Wang N, Dasch GA, Struchiner CJ, Madey G, Hannick LI, Bidwell S, Joardar V, Caler E, Shao R, Barker SC, Cameron S, Bruggner RV, Regier A, Johnson J, Viswanathan L, Utterback TR, Sutton GG, Lawson D, Waterhouse RM, Venter JC, Strausberg RL, Berenbaum MR, Collins FH, Zdobnov EM, Pittendrigh BR (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle: Proc Natl Acad Sci U S A 107:12168-12173.

Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating: Bioinformatics 25:2286-2288.

Lavrov DV, Boore JL, Brown WM (2000) The complete mitochondrial DNA sequence of the horseshoe crab Limulus polyphemus: Mol Biol Evol 17:813-824.

Lavrov DV, Brown WM, Boore JL (2004) Phylogenetic position of the Pentastomida and (pan)crustacean relationships: Proc Biol Sci 271:537-544.

Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kuck P, Ebersberger I, Walzl M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wagele JW, Misof B (2010) A phylogenomic approach to resolve the arthropod tree of life: Mol Biol Evol 27:2451-2464.

Nabholz B, Ellegren H, Wolf JB (2012) High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes: Mol Biol Evol Published on line ahead of press.

Negrisolo E, Minelli A, Valle G (2004) The mitochondrial genome of the house centipede scutigera and the monophyly versus paraphyly of myriapods: Mol Biol Evol 21:770-780.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'Leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW (2007) Genome sequence of Aedes aegypti, a major arbovirus vector: Science 316:1718-1723.

Pisani D (2004) Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda: Syst Biol 53:978-989.

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

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

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences: Nature 463:1079-1083.

Rendon-Anaya M, Delaye L, Possani LD, Herrera-Estrella A (2012) Global transcriptome analysis of the scorpion Centruroides noxius: new toxin families and evolutionary insights from an ancestral scorpion species: PLoS One 7:e43331.

Reyes A, Gissi C, Pesole G, Saccone C (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals: Mol Biol Evol 15:957-966.

Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJ, Klingler M, Lorenzen M, Roth S, Schroder R, Tautz D, Zdobnov EM, Muzny D, Attaway T, Bell S, Buhay CJ, Chandrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan

ZM, Jackson L, Kovar C, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, Nguyen N, Okwuonu G, Parker D, Ruiz SJ, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Vattahil S, Villasana D, White CS, Wright R, Park Y, Lord J, Oppert B, Brown S, Wang L, Weinstock G, Liu Y, Worley K, Elsik CG, Reese JT, Elhaik E, Landan G, Graur D, Arensburger P, Atkinson P, Beidler J, Demuth JP, Drury DW, Du YZ, Fujiwara H, Maselli V, Osanai M, Robertson HM, Tu Z, Wang JJ, Wang S, Song H, Zhang L, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen HC, Ermolaeva O, Hlavina W, Kapustin Y, et al (2008) The genome of the model beetle and pest Tribolium castaneum: Nature 452:949-955.

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator): Mol Phylogenet Evol 53:826-834.

Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata: Proc Biol Sci 278:298-306.

Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV (2010) Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda: Genome Biol Evol 2:425-440.

Rota-Stabelli O, Telford MJ (2008) A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics: Mol Phylogenet Evol 48:103-111.

Rota-Stabelli O, Yang Z, Telford MJ (2009) MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies: Mol Phylogenet Evol 52:268-272.

Rota-Stabelli O, Lartillot N, Philippe H, Pisani D (2012) Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study: Syst Biol Published on line ahead of press.

Shao R, Barker SC (2003) The highly rearranged mitochondrial genome of the plague thrips, Thrips imaginis (Insecta: Thysanoptera): convergence of two novel gene boundaries and an extraordinary arrangement of rRNA genes: Mol Biol Evol 20:362-370.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data: Genome Res 19:1117-1123.

Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Fave MJ, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkampf M, Hu H, Johnson BR, Kim J, Marsh SE, Moeller JA, Munoz-Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S, Tsutsui ND, Viljakainen L, Wissler L, Yandell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR (2011) The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle: PLoS Genet 7:e1002007.

Thorley JL, Wilkinson M (1999) Testing the phylogenetic stability of early tetrapods: J Theor Biol 200:343-344.

Vieira FG, Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system: Genome Biol Evol 3:476-490.

Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution: J Hered 92:371-373.

Zhang DX, Hewitt GM (1997) Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies: Biochemical Systematics and Ecology 25:99-120.

Table 3.1. The most important among the available arthropod genomes

Classification	Species	Genome Size (Mb)	GC (%)	Chromosomes	Genes	Proteins
Chelicerata (Acari – Acariformes)	*Tetranychus urticae*	89.6	32.3	N/A	N/A	18,414
Chelicerata (Acari – Parasitiformes)	*Ixodes scapularis*	1,896.32	45.5	15	7,112	5,867
Myriapoda (Chilopoda)	*Strigamia maritime*	173.61	35.7	N/A	N/A	N/A
Crustacea (Branchiopoda)	*Daphnia pulex*	158.62	40.8	N/A	30,613	30,611
Hexapoda (Anoplura)	*Pediculus humanus*	108.37	27.5	N/A	10,993	10,775
Hexapoda (Coleoptera)	*Tribolium castaneum*	210.27	38.4	10	10,132	9,833
Hexapoda (Hemiptera)	*Acyrthosiphon pisum*	464	29.6	4	N/A	11,089
Hexapoda (Hymenoptera)	*Apis mellifera*	250.29		16	N/A	N/A
Hexapoda (Lepidoptera)	*Bombyx mori*	431.75	37.7	28	N/A	N/A
Hexapoda (Lepidoptera)	*Heliconius melpomene*	269		21	12669	N/A
Hexapoda (Diptera)	*Drosophila melanogaster*	139.73	42.2	6	15,431	24,113
Hexapoda (Diptera)	*Aedes aegypti*	1,310.11	38.3	3	16,684	16,785
Hexapoda (Diptera)	*Anopheles gambiae*	265.03	44.5	5	13,24	14,099

Table legend: N/A not available. All the values in the table were obtained either from the NCBI website or from the original genome paper.

Figure 3.1

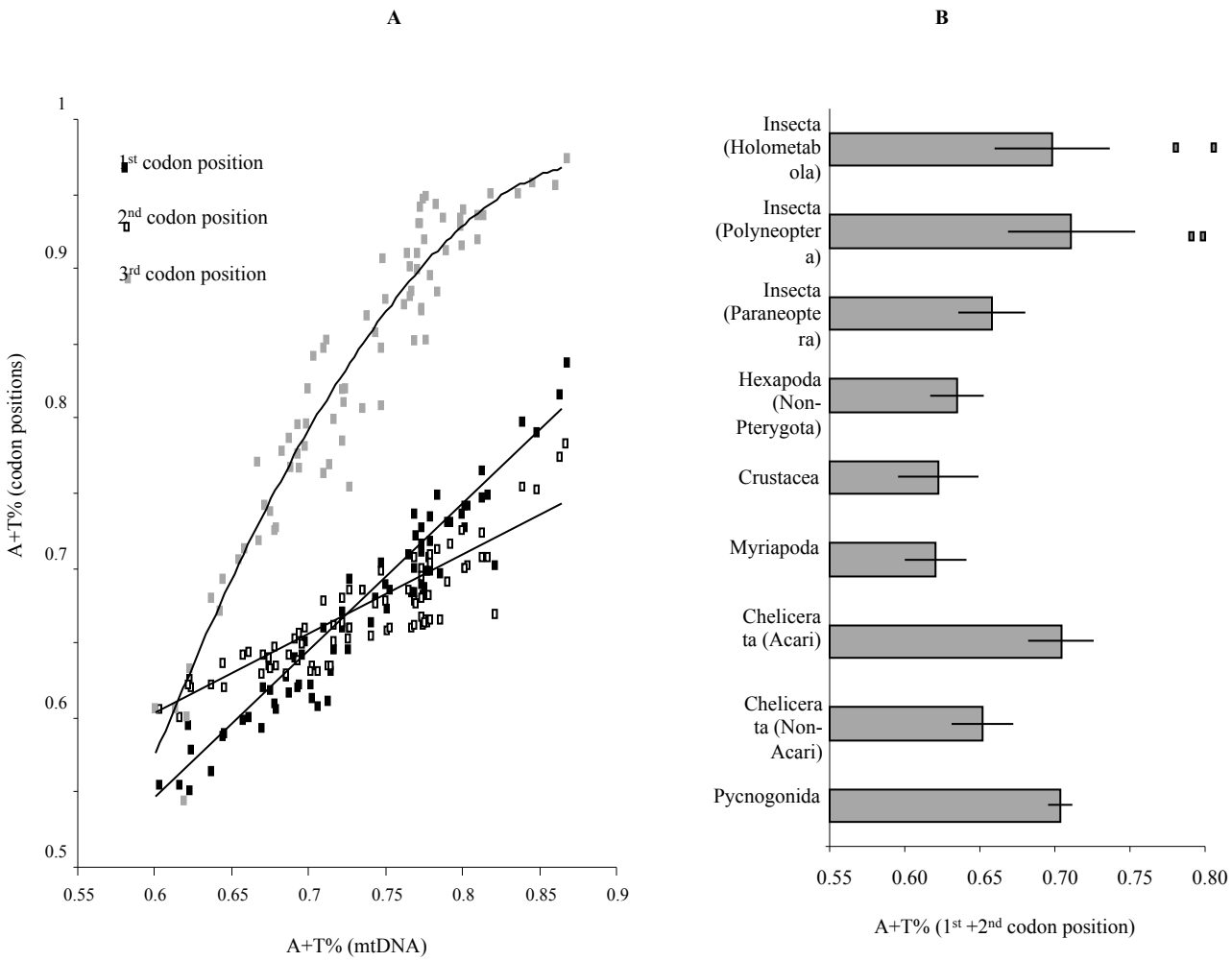
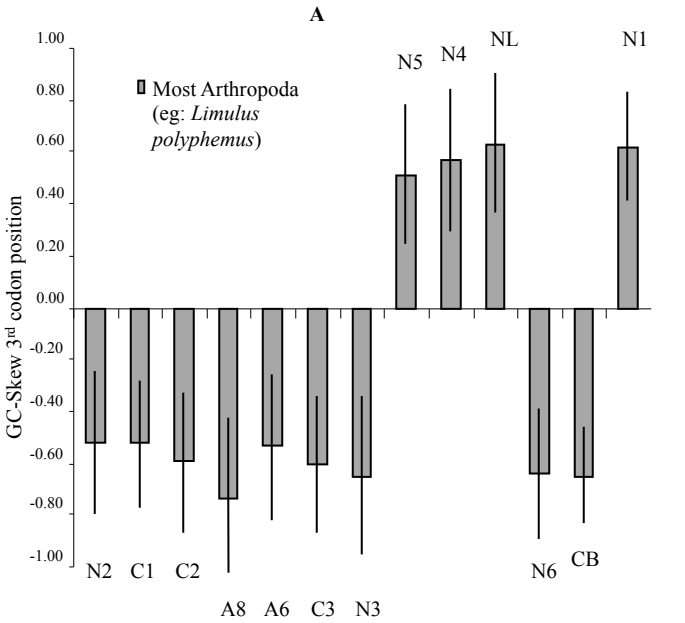
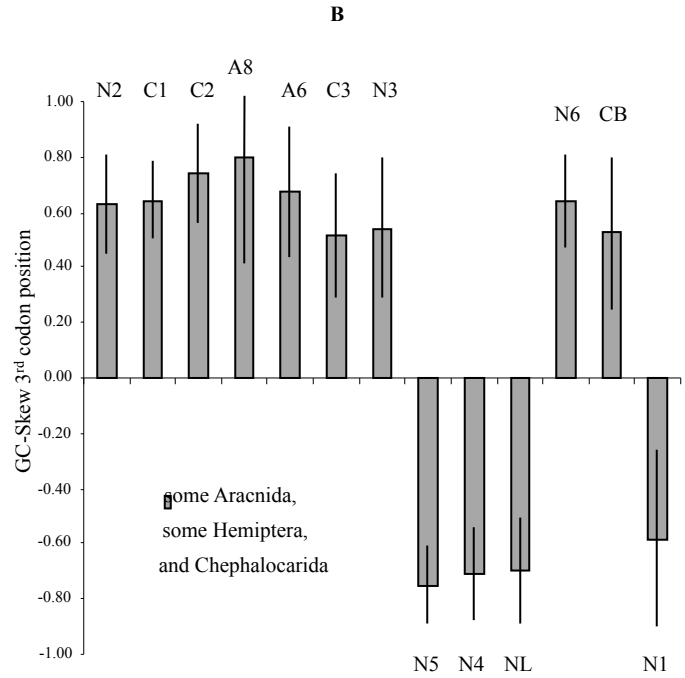
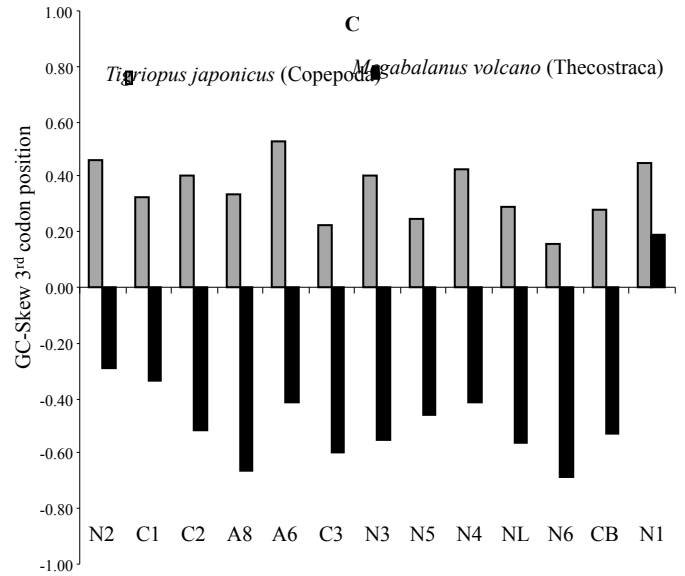
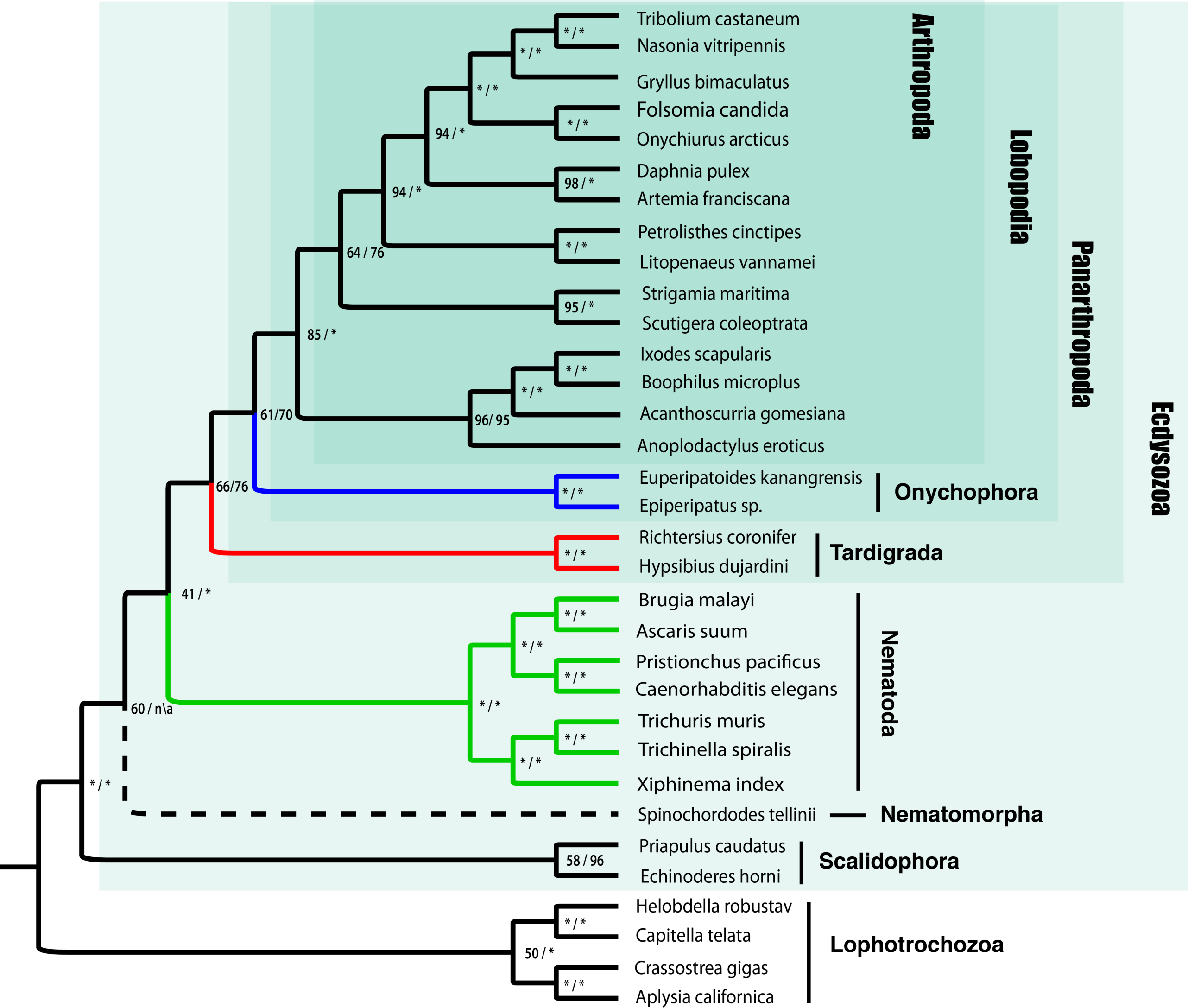

A
B
1st codon position
2nd codon position
3rd codon position
A+T% (codon positions)
A+T% (mtDNA)
1
0.9
0.8
0.7
0.6
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
Insecta (Holometabola)
Insecta (Polyneoptera)
Insecta (Paraneoptera)
Hexapoda (Non-Pterygota)
Crustacea
Myriapoda
Chelicerata (Acari)
Chelicerata (Non-Acari)
Pycnogonida
0.55
0.60
0.65
0.70
0.75
0.80
A+T% (1st +2nd codon position)

Figure 3.2


A
GC-Skew 3rd codon position
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
-1.00
Most Arthropoda
(eg: Limulus
polyphemus)
N5
N4
NL
N1
N2
C1
C2
A8
A6
C3
N3
N6
CB


B
GC-Skew 3rd codon position
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
-1.00
N2
C1
C2
A8
A6
C3
N3
N6
CB
some Aracnida,
some Hemiptera,
and Chephalocarida
N5
N4
NL
N1


C
GC-Skew 3rd codon position
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
-1.00
Tigriopus japonicus (Copepoda)
Megabalanus volcano (Thecostraca)
N2
C1
C2
A8
A6
C3
N3
N5
N4
NL
N6
CB
N1


Tribolium castaneum
Nasonia vitripennis
Gryllus bimaculatus
Folsomia candida
Onychiurus arcticus
Daphnia pulex
Artemia franciscana
Petrolisthes cinctipes
Litopenaeus vannamei
Strigamia maritima
Scutigera coleoptrata
Ixodes scapularis
Boophilus microplus
Acanthoscurria gomesiana
Anoplodactylus eroticus
Euperipatoides kanangrensis
Epiperipatus sp.
Richtersius coronifer
Hypsibius dujardini
Brugia malayi
Ascaris suum
Pristionchus pacificus
Caenorhabditis elegans
Trichuris muris
Trichinella spiralis
Xiphinema index
Spinochordodes tellinii
Priapulus caudatus
Echinoderes horni
Helobdella robustav
Capitella telata
Crassostrea gigas
Aplysia californica
Arthropoda
Lobopodia
Panarthropoda
Ecdysozoa
Onychophora
Tardigrada
Nematoda
Nematomorpha
Scalidophora
Lophotrochozoa
* / *
94 / *
64 / 76
85 / *
61/70
96/ 95
66/76
41 / *
60 / n\a
98 / *
95 / *
58 / 96
50 / *

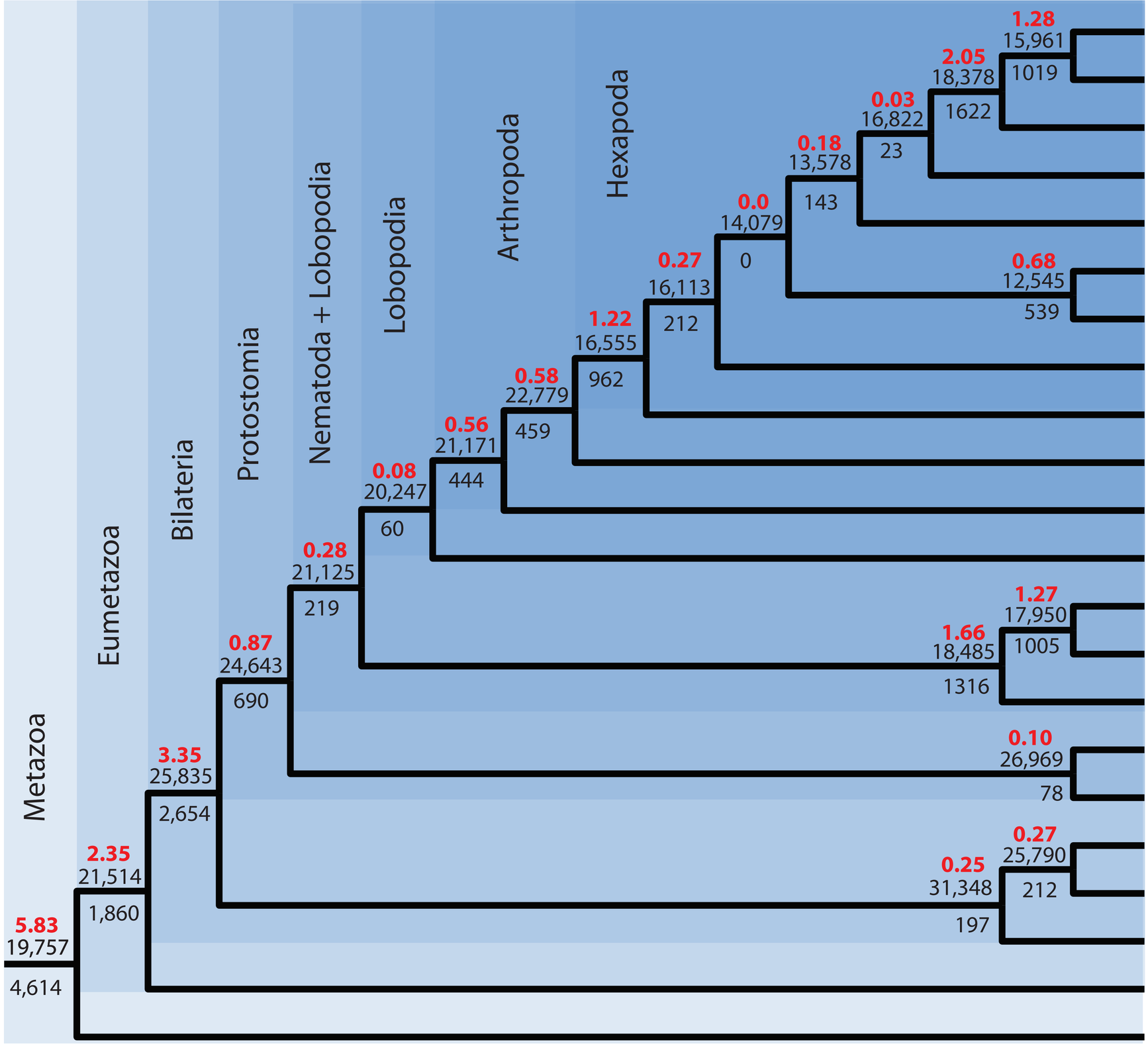

Metazoa
Eumetazoa
Bilateria
Protostomia
Nematoda + Lobopodia
Lobopodia
Arthropoda
Hexapoda
5.83
19,757
4,614
2.35
21,514
1,860
3.35
25,835
2,654
0.87
24,643
690
0.28
21,125
219
0.08
20,247
60
0.56
21,171
444
0.58
22,779
459
1.22
16,555
962
0.27
16,113
212
0.0
14,079
0
0.18
13,578
143
0.03
16,822
23
2.05
18,378
1622
1.28
15,961
1019
1.66
18,485
1316
1.27
17,950
1005
0.25
31,348
197
0.27
25,790
212
0.10
26,969
78
0.68
12,545
539
Amphimedon queenslandica (252 | 18,693)
Hydra magnipapillata (1,100 | 18,950)
Strongylocentrotus purpuratus (4,599 | 42,420)
Homo sapien (2,220 | 32,130)
Ciona intestinalis (621 | 13,892)
Capitella telata (2,305 | 32,415)
Lottia Giganta (132 | 23,851)
Trichinella spiralis (1,391 | 16,380)
Brugia malayi (363 | 11,472)
Caenorhabditis elegans (2,086 | 23,894)
Peripatoides novaezealandiae (444 | 18,447)
Ixodes scapularis (666 | 20,486)
Daphnia Pulex (2,526 | 30,611)
Pediculus humanus (57 | 10,775)
Acyrthosiphon pisum (1,065 | 17,705)
Nasonia vitripennis (272 | 12,988)
Apis melifera (103 | 10,570)
Tribolium castaneum (99 | 9,833)
Bombyx mori (43 | 18,510)
Drosophila melanogaster (1,119 | 22,352)
Aedes aegypti (426 | 15,419)
Anopheles gambiae (254 | 14,086)

Fig. 3.5

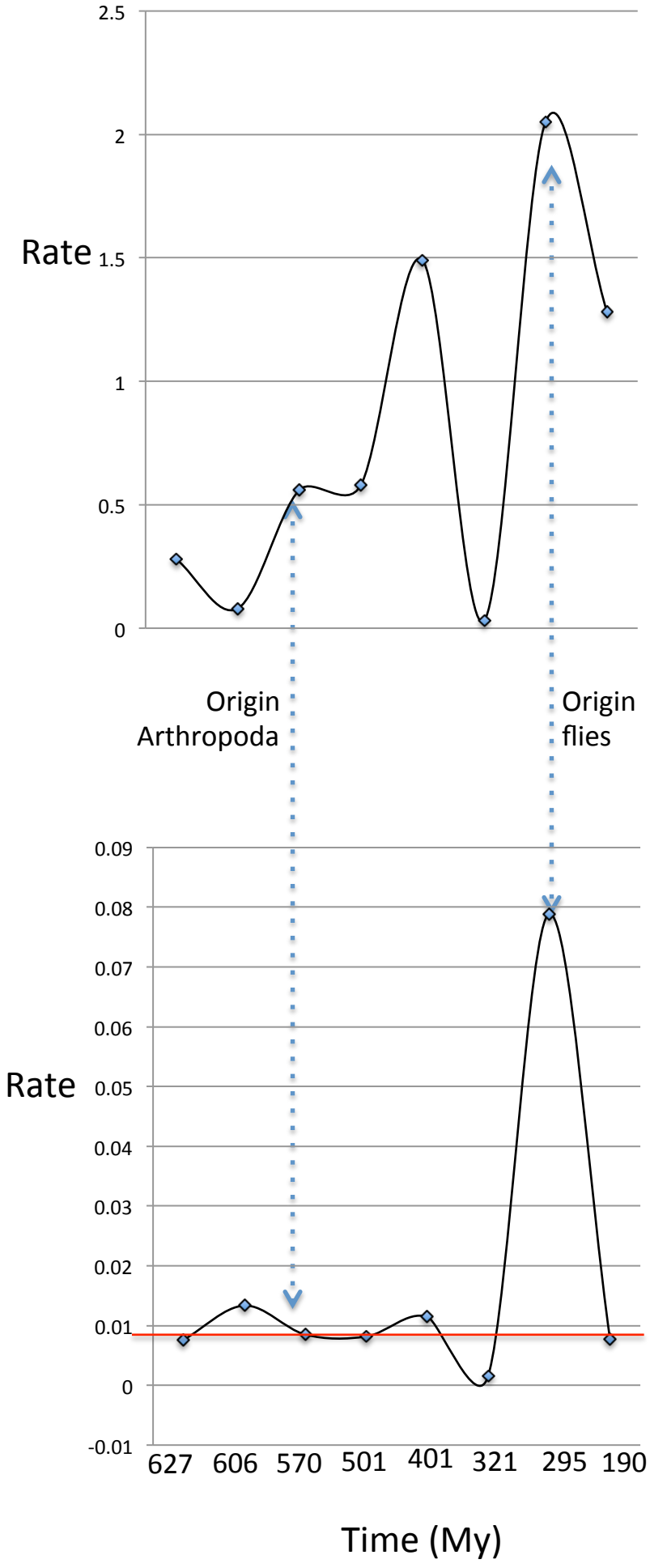
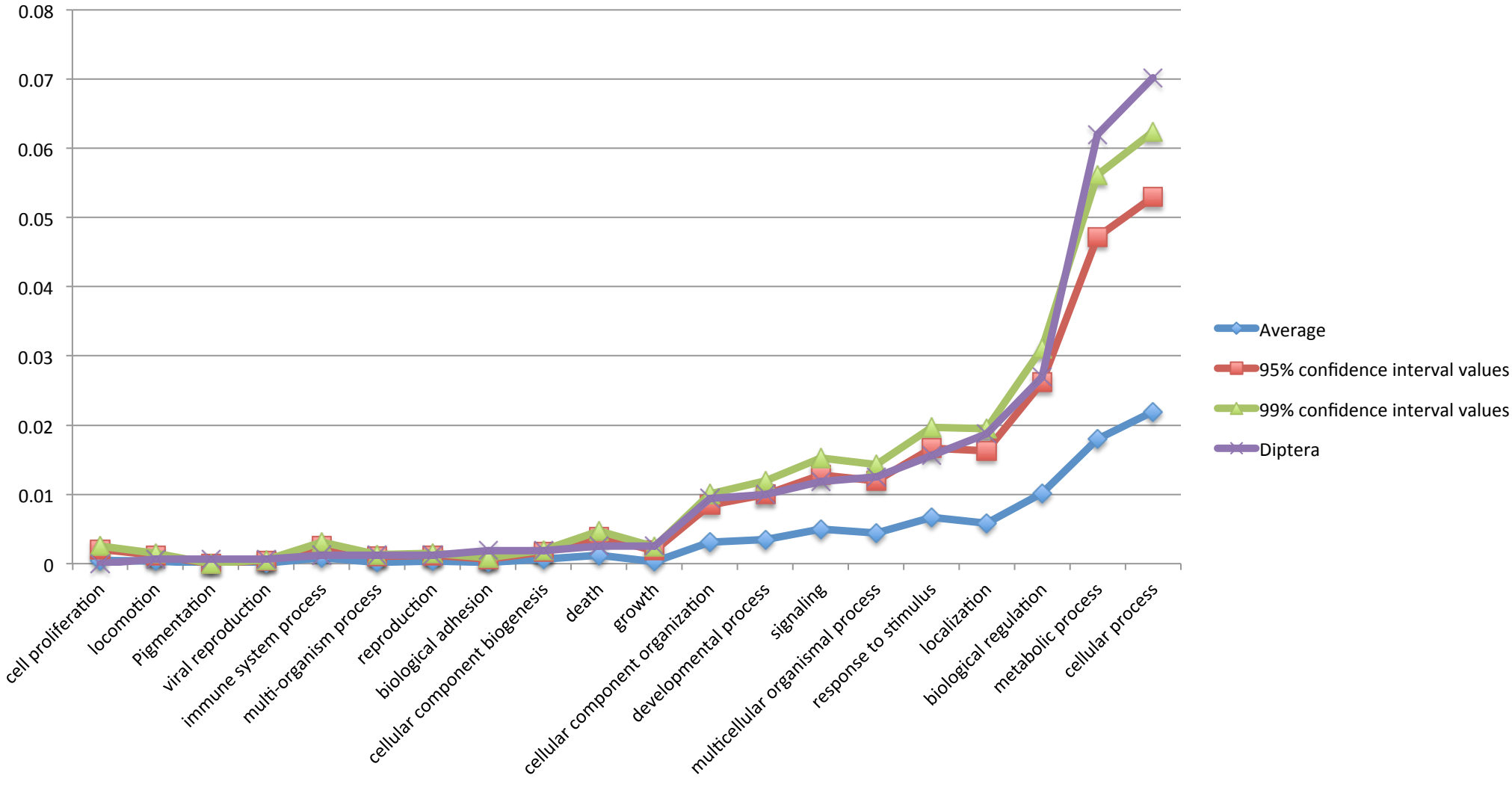

(a)
Rate
2.5
2
1.5
1
0.5
0
Origin
Arthropoda
Origin
flies
(b)
Rate
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
627
606
570
501
401
321
295
190
Time (My)

Fig. 3.6


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
cell proliferation
locomotion
Pigmentation
viral reproduction
immune system process
multi-organism process
reproduction
biological adhesion
cellular component biogenesis
death
growth
cellular component organization
developmental process
signaling
multicellular organismal process
response to stimulus
localization
biological regulation
metabolic process
cellular process
Average
95% confidence interval values
99% confidence interval values
Diptera