universität wien

DISSERTATION

Titel der Dissertation

Postprocessing Phylogenies: Tree Distances and Supertrees

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr. rer.nat.)

Verfasserin:	Anne Kupczok
Matrikel-Nummer:	0649109
Dissertationsgebiet (lt. Studienblatt):	Molekulare Biologie
Betreuer:	Univ.-Prof. Dr. Arndt von Haeseler

Wien, im Januar 2010

Acknowledgements

First, I thank my supervisor, Professor Arndt von Haeseler, who found a well-working balance between advice and motivation on the one side and allowing for space to develop own ideas on the other side. The CIBIV environment also contributed to this thesis. I thank the co-authors of my papers, Heiko Schmidt, in particular for bringing up and persuing the supertree topic, and Steffen Klaere, in particular for his inspiring mathematical assistance. In addition, I acknowledge interesting discussions and support with written English from Greg Ewing and Mareike Fischer. Andrea Führer prepared the tree shape graphics, but our personal talks, especially at the end of her and at the beginning of my PhD, were much more valuable. This thesis is written using the personal pronoun 'we' in the scientific sense. This also highlights that this work could only be accomplished in interaction.

Furthermore, I thank Allen Rodrigo for bringing the geodesic distance back into my mind and thus starting to work on its algorithm and Karen Vogtmann and Megan Owen for helpful discussions and the kind provision of unpublished material.

Last but not least, I thank my friends in Vienna; with their aid, I found a new home. I deeply appreciate Stan Gorkiewicz's support and patience in the final stage of my thesis. Finally, I want to dedicate this thesis to my friend Rainer Bühler, from whom I learned about the meaning of a home, individuality and innumerable more things.

"Selbst du als Wissenschaftlerin wirst keine Hohepriesterin sein, die alles unterrichten und berichtigen kann, sondern du wirst genauso vor den Menschen dich verneigen, die anderes Wissen als du beherrschen." (R. B.)

Abstract

More and more phylogenetic trees are generated, and it frequently occurs that the inferred relationships contradict each other. In this case, tools are necessary which evaluate the amount of difference between two trees, extract the congruencies of two trees, and combine multiple trees by minimizing the incongruencies. These tools are summarized by the term "phylogenetic postprocessing". In this thesis, two aspects of phylogenetic postprocessing are investigated in detail.

First, tree distance computations evaluate the amount of difference between two trees. Most measures only take the topological information into account. There are a few measures that additionally focus on the branch lengths of the trees. One of these is the length of the shortest path in the space of weighted trees, also known as the geodesic distance. Here, an exact, but exponential-time, algorithm to compute the geodesic distance is presented. Comparisons with its approximations show that there is a particular path that approximates the geodesic distance well and that can be computed in linear time.

Phylogenetic trees can also be tested for being statistically similar or different. Then a topological distance measure can be used as a test statistic where the associated p-value is computed under a null distribution of trees. Discrete tests must ensure that the size of the test is conservative, i. e. the size must not exceed the significance level. We present one example where a test has to be modified to ensure this property.

Second, gene trees on overlapping taxon sets can be combined into a so-called supertree. Another possibility is to combine the gene alignments directly, namely, to concatenate the gene alignments into a superalignment and to reconstruct a phylogeny from this long alignment. There is also the possibility to combine the data at a level between superalignment and supertree methods. Simulations of gene alignments along model gene trees allow for the comparison of methods from all three levels. We investigate different settings, e. g. complete or overlapping taxon sets, equal or different substitution parameters or different gene topologies. The results show a good performance of matrix representation methods compared to other su-

pertree and medium-level methods. Furthermore, superalignment is well applicable in the case of differing parameters between genes but is problematic when a high level of incongruence is present among the true gene trees.

Additionally to the practical evaluation of supertree methods, theoretical and algorithmic aspects are of interest. Therefore we study different null models underlying supertree reconstruction. We find only the distribution of equally likely splits to behave in an appropriate way if little information is present. In contrast, the distribution of equally likely trees inserts a tree shape bias in split-based supertree methods. This bias can be traced back to the unequal split distribution in the null model.

Finally, a supertree can also be defined by minimizing the total distance to the trees in the set, i. e. as a median tree. The majority-rule consensus is described as a median tree method for trees on the same taxon set. For trees on overlapping taxon sets, however, different specifications can be used, namely MR(-)supertrees and MR(+)supertrees. We present algorithms to compute the respective distances in the matrix representation framework. Applying their implementation to simulated data sets shows a clearly better performance of MR(-) compared to MR(+). This discrepancy is likely to trace back to a tree shape bias in MR(+).

To conclude, we see that the two aspect of phylogenetic postprocessing, tree distances and tree combination methods, are not independent. Instead, they are linked by the definition of the median tree. Thus our understanding of tree distances influences data combination methods and vice versa.

In summary, the space of trees, weighted or unweighted, plays an important role in the different tasks. Postprocessing phylogenies must also consider the complex space of trees to get unbiased results. We show that from six taxa on, tree topologies follow a complex structure: They possess different kinds of splits and different kinds of shapes. To highlight this, the topologies for six taxa will be presented on one page each.

Zusammenfassung

Es werden immer mehr phylogenetische Bäume berechnet. Die berechneten Verwandtschaften zwischen den Arten können sich allerdings widersprechen. In diesem Fall sind Werkzeuge notwendig, welche die Höhe des Unterschiedes berechnen, die Gemeinsamkeiten zweier Bäume extrahieren und mehrere Bäume zusammenfassen indem sie die Unterschiede minimieren. Diese Werkzeuge werden unter dem Begriff "Phylogenetic Postprocessing" zusammengefasst. In dieser Arbeit werden zwei Aspekte des Phylogenetischen Postprocessings im Detail untersucht.

Zuerst werden Baumdistanzen untersucht. Diese evaluieren den Unterschied zweier Bäume. Die meisten Maße berücksichtigen dabei nur die topologische Information. Allerdings tragen auch die Kantenlängen der Bäume Informationen, da sie z. B. eine Schätzung der Menge an Unterschied zwischen zwei Sequenzen sind. Ein Maß, welches sowohl die Topologie als auch die Kantenlängen berücksichtigt, ist die Länge des kürzesten Weges durch den Raum aller Bäume mit Kantenlängen. Dies ist die geodätische Distanz. Hier präsentieren wir einen exakten Algorithmus um die geodätische Distanz zu berechnen, der in exponentieller Zeit läuft. Vergleiche mit ihren Approximationen zeigen, dass es einen bestimmten Weg gibt, der die geodätische Distanz gut annähert und in linearer Zeit berechnet werden kann.

Phylogenetische Bäume können auch daraufhin untersucht werden, ob sie statistisch ähnlich oder unterschiedlich sind. Dabei kann ein topologisches Distanzmaß als Teststatistik verwendet und die assoziierten p-Werte werden unter einer Nullverteilung der Bäume berechnet werden. Bei diskreten Testverfahren, muss allerdings die Testgröße konservativ gewählt werden, d. h. sie darf das Signifikanzniveau nicht überschreiten. Wir zeigen ein Beispiel auf, bei dem ein Test abgeändert werden muss um dies zu gewährleisten.

Der zweite Aspekt ist die Kombination von Bäumen oder allgemein phylogenetischen Datensätzen. Genbäume mit sich überschneidenden Artenmengen können zu einem sogenannten Supertree zusammengefügt werden. Eine andere Möglichkeit ist bereits die Genalignments zu kombinieren. Dabei werden die Genalignments aneinan-

dergehangen, d.h. zu einem sogenannten Superalignment kombiniert. Anschließend wird eine Phylogenie aus diesem langen Alignment berechnet. Es gibt auch die dritte Möglichkeit, die Daten auf einer Stufe zwischen Superalignment und Supertree zu kombinieren. Mit Hilfe von Simulationen von Genalignments entlang Modellbäumen können Methoden von diesen drei Stufen verglichen werden. Wir untersuchen verschiedene Parameter, z.B. vollständige oder sich überschneidende Artenmengen, gleiche oder unterschiedliche Substitutionsparameter oder unterschiedliche Gentopologien. Die Simulationen zeigen gute Ergebnisse der Matrix-Representation-Methoden im Vergleich zu anderen Supertreemethoden. Weiterhin ist Superalignment gut geeignet bei unterschiedlichen Parametern zwischen den Genen, aber problematisch wenn es viele Unterschiede zwischen den wahren Genbäumen gibt.

Zusätzlich zu diesem praktischen Vergleich von Supertreemethoden sind auch theoretische und praktische Aspekte von Interesse. Daher untersuchen wir die Nullmodelle, die der Supertreerekonstruktion zugrunde liegen. Ein solches Nullmodell ist die Gleichverteilung der Splits, also jeder möglichen Unterteilung der Arten in zwei Mengen. Es stellt sich heraus, dass nur diese Verteilung angemessene Eigenschaften hat, wenn wenig Information vorhanden ist. Ein zweites Nullmodell ist die Gleichverteilung der Bäume. Diese fügt allerdings eine Verzerrung zugunsten bestimmter Baumstrukturen in splitbasierte Supertreemethoden ein. Diese Verzerrung kann auf die ungleiche Verteilung der Splits in diesem Nullmodell zurückgeführt werden.

Schließlich kann ein Supertree auch als Median-Tree definiert werden, also als Baum, der die totale Distanz zu allen Bäumen in der Menge minimiert. Der Majority-Rule Consensus wurde als Median-Tree-Methode für Bäume mit gleichen Artenmengen beschrieben. Für Bäume mit sich überschneidenden Artenmengen gibt als allerdings unterschiedliche Ausprägungen, und zwar MR(-)supertrees und MR(+)supertrees. Wir präsentieren Algorithmen um die entsprechenden Distanzen im Matrix-Representation-Framework zu berechnen. Durch die Anwendung ihrer Implementierungen auf simulierte Datensätze sehen wir deutlich bessere Ergebnisse für MR(-) im Vergleich zu MR(+). Es ist naheliegend diesen Unterschied auf eine Verzerrung zugunsten bestimmter Baumstrukturen in MR(+) zurückzuführen.

Zusammenfassend sehen wir, dass die zwei Aspekte des Phylogenetischen Postprocessings, also Baumdistanzen und Baumkombinationsmethoden, nicht unabhängig sind, sondern durch die Definition des Median-Trees verbunden. Daher wird unser Verständnis von Baumdistanzen auch die Kombination von Bäumen beeinflussen und umgekehrt.

Parts of this thesis have been published in the following articles:

1. Anne Kupczok, Arndt von Haeseler, and Steffen Klaere (2008) An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees. *J. Comput. Biol.*, **15**(6):577–591.

2. Anne Kupczok and Arndt von Haeseler (2009) Comment on 'A congruence index for testing topological similarity between trees'. *Bioinformatics*, **25**(1):147–149.

3. Anne Kupczok, Heiko A. Schmidt, and Arndt von Haeseler (2009) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Submitted to *BMC Evol. Biol.*

4. Anne Kupczok (2009) Consequences of different null models on the tree shape bias of supertree methods. Submitted to *Syst. Biol.*

Contents

Chapter 1

Introduction into Phylogenies

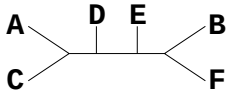
1.1 Introduction

A phylogenetic tree or *phylogeny* is a graphical representation of putative gene or species relationships. More and more phylogenetic trees are generated (about 15 published per day; Rokas, 2006), and it frequently occurs that the represented relationships contradict each other. In this case, tools for phylogenetic postprocessing are of particular importance. Phylogenetic postprocessing mainly comprises two different tasks: (1) Computing the distance between two (or more) phylogenies and (2) combining multiple phylogenies in an appropriate way into a new phylogeny that best uses the information present and resolves conflicting information. The latter task also comprises the combination of trees on different, but overlapping, sets of species into a larger phylogeny displaying relationships between all the species studied. Before these topics are investigated in detail in Chapters 2 and 3, respectively, we will broadly introduce the phylogeny problem (Section 1.2). For a detailed introduction into phylogenetic trees, see Section 1.3.

1.2 Reconstructing Trees from Sequences

1.2.1 Gene Trees and Species Trees

Phylogenies are usually computed from aligned gene sequences (Section 1.2.2). The assumption, that the gene phylogeny equals the species tree, may be invalid due to the following problems:



Biological processes The gene phylogeny is correctly reconstructed, but may not be equal to the species phylogeny for one of many different reasons, e. g. gene duplication and loss, deep coalescence, or horizontal gene transfer (see e. g. Maddison, 1997, for an overview). In this case, there may not even exist a tree to display the true species relationships, but a phylogenetic network would be the correct model.

Model misspecification and bias Estimating phylogenies necessarily involves the specification of an explicit or implicit model which never fits the true biological process completely. These processes may bias the phylogeny estimation (e. g. Ho and Jermiin, 2004).

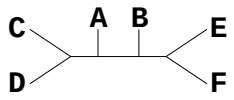
Stochastic processes Even if the *true* gene history equals the species history, a different phylogeny is inferred from the limited amount of data. This may be caused by randomly occuring mutations which better fit another tree.

Some of the problems can be reduced by sampling data from multiple genes and concatenating their alignments. First, by looking at various gene histories, the analysis is less biased by particular gene histories, and second, using more information also improves parameter estimation and reduces stochastic noise (e. g. Rokas *et al.*, 2003). However, using a large data set can also intensify the bias in the data (e. g. Phillips *et al.*, 2004; Rodríguez-Ezpeleta *et al.*, 2007). This could be averted by computing each gene phylogeny independently and using a consensus or supertree approach to combine the trees. These data combination methods will be discussed in Chapter 3.

1.2.2 Methods for Phylogeny Reconstruction

The trees for postprocessing must have been inferred from data at some point in time. The classical phylogeny reconstruction problem is the reconstruction from a biological data set, namely morphological or molecular information. We will concentrate on molecular sequences although including morphological sequences may improve resolution and support of a phylogeny (Wortley and Scotland, 2006).

The collection of data sets from multiple genes follows two general strategies: (1) using only genes that provide full information, i. e. cover all taxa of interest (e. g. Ciccarelli *et al.*, 2006) or (2) using all available genes that are present in some taxa and fulfill special overlap conditions (e. g. Driskell *et al.*, 2004; McMahon and Sanderson, 2006; Schmidt, 2003). The latter approach is able to use many more genes and taxa,



since it allows for missing data. It can also be applied for phylogeny reconstruction from expressed sequence tags (ESTs, e. g. Philippe and Telford, 2006).

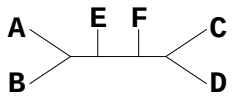
Before the gene alignments are obtained, two important steps can influence the phylogeny result: First, orthologs must be assigned correctly (see e. g. Chen *et al.*, 2007; Dutilh *et al.*, 2007, for method comparisons). Second, these orthologs need to be aligned with sufficient accuracy (see e. g. Edgar and Batzoglou (2006) for a review, and Landan and Graur (2007) for an example of the impact of alignment accuracy on phylogeny reconstruction).

Then the phylogeny reconstruction problem can be formulated as follows: Given a set of aligned orthologous sequences, reconstruct a weighted tree which best describes the observed sequence data. This simple description has two implications: First, the phylogeny reconstruction problem is usually independent from the alignment problem, i. e. it takes the orthology relationships and the alignment as given. Second, we need a formal way to evaluate which tree is better than an other. Only then, we can find a *best* tree. There are different objective functions for this problem:

Maximum Parsimony (MP) Minimize the number of substitutions the sequences need to evolve along a tree (Camin and Sokal, 1965). Thereby, each alignment column is considered independently, and its number of substitutions needed along a particular tree is called *parsimony length*. The *total parsimony length* (PL) is the sum of parsimony lengths over every alignment column, i. e. the total number of substitutions the sequences need to evolve along a particular tree.

Maximum Likelihood (ML) Maximize the conditional probability of the sequence data given a tree and an evolutionary model (Felsenstein, 1981). For nucleotide data, not only the phylogeny, but also the parameters of the evolutionary model are estimated. Thereby, the type of evolutionary model must be given, e. g. Jukes-Cantor model (JC, Jukes and Cantor, 1969) where each substitution is equally likely, HKY model (Hasegawa *et al.*, 1985) where transitions and transversions have different probabilities, or general time-reversible model (GTR, Lanave *et al.*, 1984) where each type of substitution has a different probability. In contrast, the substitution model for protein sequences is usually given by an empirical model, e. g. JTT (Jones *et al.*, 1992).

Distance-based methods Compute a distance matrix from the alignment by estimating the pairwise distances for each pair of taxa. Subsequently, fit the distance matrix to a tree. Thereby, the branch lengths connecting any two taxa should



get close to the pairwise distances (the least squares method, Fitch and Margoliash, 1967). Neighbor joining (Saitou and Nei, 1987) is another algorithm to compute trees from pairwise distances.

This is an incomplete list of phylogeny algorithms, for which an even larger number of programs is available (some of which are listed at `http://evolution.genetics.washington.edu/phylip/software.html`). That means, already at the tree reconstruction step, a decision is necessary how the tree shall be computed. The different reconstruction algorithms and programs are described in detail elsewhere (e. g. Lemey *et al.*, 2009).

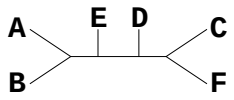
In principle, the origin of the trees is not important for the subsequent postprocessing. Only in the case where branch lengths are considered, e. g. for distance computations, the scales of the trees should be comparable.

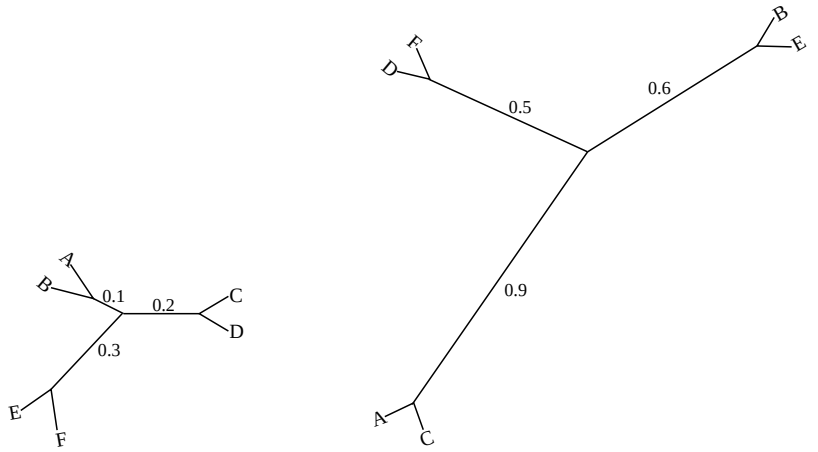
1.3 Mathematical Description of Phylogenetic Trees and Tree Spaces

1.3.1 Introduction into Trees

Here we will give a detailed description about the mathematical structure of trees and tree spaces which is needed in the following chapters. The terminology will follow Kupczok *et al.* (2008).

Phylogenetic trees are leaf-labeled trees, where the leaves are called *taxa*. One distinguishes between rooted or unrooted phylogenetic trees. In case of rooted trees, we treat the root as an additional taxon of an unrooted tree. The usual phylogenetic inference methods (Section 1.2.2) can only reconstruct unrooted trees, thus we will concentrate on unrooted trees. The term phylogenetic tree can stand for a topology only, or a weighted tree. A topology is the branching pattern of the taxa, whereas a weighted tree adds branch lengths to such a topology. Topologies and weighted trees are introduced in detail in Sections 1.3.2 and 1.3.3, respectively. We will use the two terms if the discrimination is important or the term (phylogenetic) tree if the meaning is unambiguous in the context.





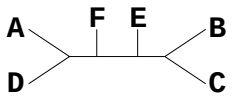
(a) Tree with $\mathcal{T}^1$ **(b)** Tree with $\mathcal{T}^2$

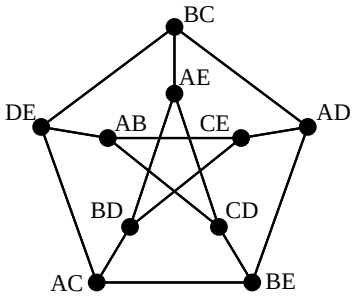
Figure 1.1: Examples for phylogenetic trees on taxon set $X = \{A, B, C, D, E, F\}$: $\mathcal{T}^1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{A, B\}, \{C, D\}, \{E, F\}\}$ and $\mathcal{T}^2 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{A, C\}, \{B, E\}, \{D, F\}\}$. We will use the following notation: $\mathcal{T}^1 = \{A, B, C, D, E, F, AB, CD, EF\}$ and $\mathcal{T}^2 = \{A, B, C, D, E, F, AC, BE, DF\}$.

1.3.2 The Discrete Topology Space

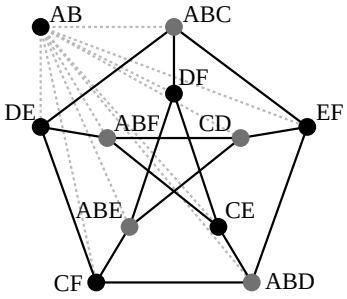
Topologies

A *topology* $\mathcal{T}$ is identified by its taxon set X and its edge set, where *terminal* edges connect a leaf with an inner node and *interior* edges connect two inner nodes. In unrooted trees, there is no node of degree two. If an edge of a phylogenetic tree is deleted, the tree decomposes into two connected components. Thus, the taxon set is then partitioned into two sets (X_1 and X_2), one for each component. Such a bipartition is called a *split*, and is identified by $X_1|X_2$. If the underlying taxon set $X = X_1 \cup X_2$ is clearly stated, we also identify a split with one set, X_1 or X_2, which does not have more elements than the other. A k-split refers to a partition into k and $n - k$ taxa, i. e. $k = min(|X_1|, |X_2|)$. Since each edge in a topology corresponds to a split, we will identify a topology on taxon set X by the corresponding split set (see example in Figure 1.1). In this thesis, the example taxon set will usually contain only one-letter taxa. Then the taxon sets in a split can be shortly written as a string of concatenated taxa (Figure 1.1).

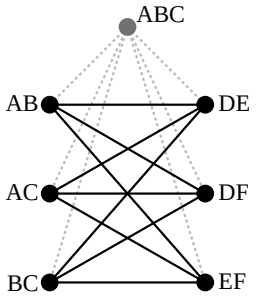




(a) 5 taxa



(b) 6 taxa, subgraph induced by split AB



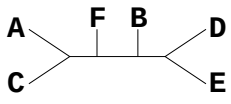
(c) 6 taxa, subgraph induced by split ABC

Figure 1.2: Compatibility graphs for five and six taxa. Nodes denote interior splits and edges indicate compatibility between the connected splits. (a) Interior splits of five taxa ($X = \{A, B, C, D, E\}$). Here, an edge also depicts a bifurcating topology identified by two compatible interior splits. There are 15 edges and 15 topologies. The graph is the well-known Petersen graph. (b) and (c) Compatibility subgraphs for six taxa ($X = \{A, B, C, D, E, F\}$) induced by the splits AB resp. ABC. The splits AB and ABC are representatives for all 2-splits resp. 3-splits, since compatibility graphs induced by other splits are isomorphic to one of the graphs. The full graph for six taxa would consist of 25 nodes and 105 edges forming 105 3-cliques. Thus, there are 105 different bifurcating topologies for six taxa.

Compatibility Relationships

For $n = |X|$ taxa, there are $m = 2^{n-1} - 1$ possible splits. We will denote the set of all splits for n taxa by $\mathbb{S}_n$. Analogously to the edges, we will distinguish between the n *terminal splits* (*trivial splits*) and the $m - n$ *interior splits*. Two splits are called *compatible* if there is a phylogenetic tree containing both splits. This holds for two splits $X_1|X_2$ and $Y_1|Y_2$ if at least one of the following taxon sets is empty: $X_1 \cap Y_1$, $X_1 \cap Y_2$, $X_2 \cap Y_1$ or $X_2 \cap Y_2$. Note that terminal splits are compatible to any other split. The *compatibility graph* for a set of splits is a graph whose nodes represent the splits, and edges in the graph indicate compatibility between two splits. Figure 1.2 shows the compatibility graph for the interior splits for five and six taxa. For six taxa, only the subgraphs induced by $AB|CDEF$ (Figure 1.2b) and by $ABC|DEF$ (Figure 1.2c) are shown. The *subgraph* of the compatibility graph *induced* by a split $\mathcal{S}$ consists of all splits compatible with $\mathcal{S}$. The observations for compatibility relationships on six taxa can be extended to compatibility graphs for an arbitrary number of taxa:

1. The compatibility graph induced by a 2-split is isomorphic to a complete com-



patibility graph for splits of $n-1$ taxa. E. g. the compatibility graph of all splits compatible to the 2-split $AB|CDEF$ (Figure 1.2b) is isomorphic to a compatibility graph for five taxa (Figure 1.2a).

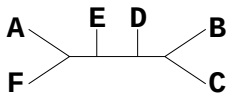
2. The compatibility graph induced by a k-split ($k > 2$) consists of two types of nodes: Type-1-nodes correspond to splits for $k+1$ taxa and are connected according to the complete compatibility graph for $k+1$ taxa, and type-2-nodes correspond to $n-k+1$ taxa and are connected accordingly. Furthermore, all edges between the nodes of the two types exist, since all the splits in the independent subtrees are compatible. In Figure 1.2c, there are two subgraphs isomorphic to two compatibility graphs for four taxa. These are simply three disconnected nodes. Both classes of nodes are completely connected with one another.

An unrooted phylogenetic tree of n taxa contains at most $n-3$ interior splits. If it contains exactly $n-3$ interior splits, all inner nodes have degree three, and the tree is called *bifurcating*, and *multifurcating* or *unresolved* otherwise. It is well-known, that $T_n = (2n-5)!! = 1 \times 3 \times \cdots \times (2n-5)$ distinct bifurcating topologies exist for $n \geq 3$ taxa (Felsenstein, 2004, Chapter 3). In the compatibility graphs, the bifurcating trees are given as cliques of $n-3$ nodes, i. e. 2-cliques for five taxa and 3-cliques for six taxa (Figure 1.2). Thus, the discrete topology space is enumerated by the maximal cliques in the compatibility graphs. Due to the compatibility restriction, the number of possible bifurcating trees in which a k-split can occur depends on k (Table 1.1).

Null Models

Different distributions can be defined on the discrete topology space. We are particularly interested in a null model containing no phylogenetic information. This is the well known distribution that each bifurcating tree for a particular number of taxa is equally likely (Proportional to Distinguishable Arrangements, PDA, e. g. Semple and Steel, 2003), i. e. each tree has a probability of $1/T_n$. With *perfect PDA* we denote the data set which contains each tree exactly once.

Second, we introduce the model that each split is equally likely (Proportional to Distinguishable Splits, PDS). Analogously, we denote the data set which contains each possible split exactly once as *perfect PDS*. The PDS model does not directly correspond to a tree distribution. But it is possible to relate each split with a multifurcating tree with only one inner branch. The PDS corresponds to an equal distribution of



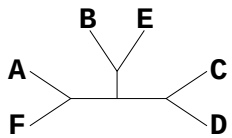
n	k	Number of splits	Number of trees per split
5	2	10	3
6	2	15	15
	3	10	9
	total	25	
7	2	21	105
	3	35	45
	total	56	
8	2	28	945
	3	56	315
	4	35	225
	total	119	
9	2	36	10395
	3	84	2835
	4	126	1575
	total	246	

Table 1.1: Number of different splits for each k and the number of different trees containing a particular k-split.

those multifurcating trees. Note that there is no distribution of bifurcating trees that corresponds to the PDS model (Steel and Pickett, 2006). The PDS and PDA models are distinct since some splits occur in more trees than others (Table 1.1). For a k-split, the number of trees containing this splits is $T_{k+1} \times T_{n-k+1}$. E. g. for $n = 6$, there are 15 different 2-splits and one particular 2-split is present in 15 different bifurcating trees. But there are 10 different 3-splits, and each is present in only 9 trees. This gap increases for larger n. E. g. for $n = 9$, the number of trees containing a particular 2-split is more than six times higher than the number of trees containing a particular 4-split.

Tree shapes

A *(tree) shape* can be obtained from a bifurcating topology by ignoring the labels. Thus a shape is an unlabeled bifurcating tree. From six taxa on, there is more than one tree shape. Note that the probabilities of shapes are different under the PDA model (Table 1.2). There are unbalanced and balanced tree shapes. We define shapes with exactly two 2-splits as *unbalanced*. (shapes $S_{6,1}$, $S_{7,1}$, $S_{8,1}$ and $S_{9,1}$). *Balanced* shapes are not clearly defined for all n but always maximize the number of 2-splits.



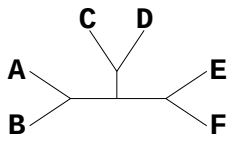
n	Shape	Number of trees	Probability
5	S_5 	15	1
6	$S_{6,1}$ 	90	0.857
	$S_{6,2}$ 	15	0.143
	total	105	
7	$S_{7,1}$ 	630	0.667
	$S_{7,2}$ 	315	0.333
	total	945	
8	$S_{8,1}$ 	5040	0.485
	$S_{8,2}$ 	2520	0.242
	$S_{8,3}$ 	315	0.030
	$S_{8,4}$ 	2520	0.242
	total	10395	
9	$S_{9,1}$ 	45360	0.336
	$S_{9,2}$ 	22680	0.168
	$S_{9,3}$ 	45360	0.336
	$S_{9,4}$ 	7560	0.056
	$S_{9,5}$ 	11340	0.084
	$S_{9,6}$ 	2835	0.021
	total	135135	

Table 1.2: Shape probabilities under the uniform tree model (PDA).

1.3.3 The Continuous Tree Space

The space is more complex when not only topologies but *weighted trees*, i. e. trees with branch lengths, are considered. The tree space $\mathbb{T}_n$ is the space of all weighted trees on n taxa. $\mathbb{T}_n$ is defined as follows (Billera *et al.*, 2001). Each split is identified with a different orthogonal unit vector $\mathbf{e}^{\mathcal{S}}$ ($\mathcal{S} \in \mathbb{S}_n$) in the m-dimensional space. Recall, that $m = 2^{n-1} - 1$ is the number of possible splits for n taxa. These unit vectors are the axes of $\mathbb{T}_n$. Thus, $\mathbb{T}_n$ is a subspace of $\mathbb{R}^m$.

For each topology $\mathcal{T}$, the unit vectors associated with its splits span a $|\mathcal{T}|$-dimensional subspace. Recall, that $|\mathcal{T}|$ is the number of splits in $\mathcal{T}$ and that $n \leq |\mathcal{T}| \leq 2n - 3$ because each topology consists at least of n terminal splits and at most $n-3$ pairwise compatible interior splits.



A weighted tree $\mathbf{p}$ with topology $\mathcal{T}$ is a point in $\mathbb{T}_n$ given by

$$\mathbf{p} = \sum_{\mathcal{S} \in \mathcal{T}} p_{\mathcal{S}} \ \mathbf{e}^{\mathcal{S}},$$

where $p_{\mathcal{S}}$ denotes the split weight of split $\mathcal{S}$ from topology $\mathcal{T}$. With this, every weighted tree $\mathbf{p}$ defines a split weight function $\lambda_{\mathbf{p}} : \mathbb{S}_n \to \mathbb{R}$ with $\lambda_{\mathbf{p}}(\mathcal{S}) = p_{\mathcal{S}}$ if $\mathcal{S} \in \mathcal{T}$ and 0 otherwise. In other words, $\lambda_{\mathbf{p}}$ assigns to each split $\mathcal{S} \in \mathcal{T}$ its weight for tree $\mathbf{p}$. Therefore, the weight function also identifies the tree (and implicitly also its topology), and we use the convention $\lambda_{\mathbf{p}}(\mathcal{T}) = \mathbf{p}$. We can apply $\lambda_{\mathbf{p}}$ to any collection of splits $\mathcal{A}$ and get a point in $\mathbb{T}_n$ with

$$\lambda_{\mathbf{p}}(\mathcal{A}) = \sum_{\mathcal{S} \in \mathcal{A}} \lambda_{\mathbf{p}}(\mathcal{S}) \ \mathbf{e}^{\mathcal{S}}.$$

In particular, $\lambda_{\mathbf{p}}$ assigns 0 to each split not in $\mathcal{T}$, thus the point $\lambda_{\mathbf{p}}(\mathcal{A})$ lies on the subspace spanned by the splits in $\mathcal{T} \cap \mathcal{A}$. We are mainly concerned with either one or two trees and thus will use λ resp. λ_i, $i = 1, 2$ to identify the trees.

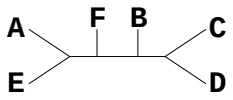
The union of weighted trees (analogously, topologies and weight functions) forms $\mathbb{T}_n$. Unresolved topologies are also included in this space. More precisely, an unresolved topology lies on the boundary of more resolved topologies. Thus, unresolved topologies connect the bifurcating topologies. An example is shown in Figure 1.3a, where the unresolved topology corresponds to the single axis $CD|ABE$ and connects the two bifurcating topologies.

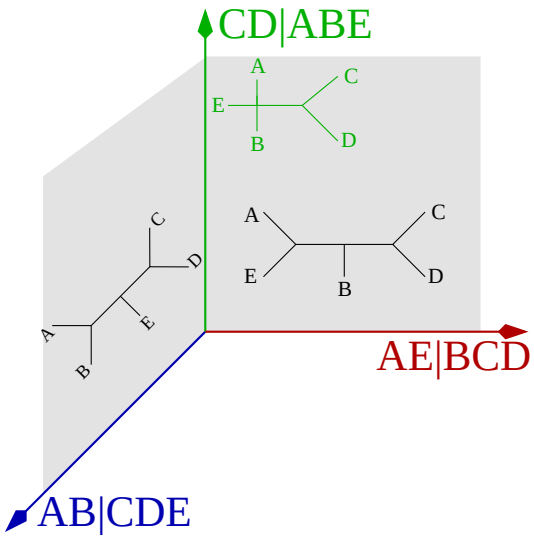
Figure 1.4 shows a visualization of $\mathbb{T}_4$ and $\mathbb{T}_5$, where only the interior splits are illustrated. The previous considerations about compatibilities of splits help to understand how to extend these figures for higher dimensions:

1. By deleting all splits incompatible with a 2-split, one reduces the dimension of $\mathbb{T}_n$. In particular, one projects $\mathbb{T}_n$ on the subspace spanned by the compatible splits. This results in a space isomorphic to $\mathbb{T}_{n-1}$ with two extra dimensions, one for the 2-split and one for the additional terminal split.

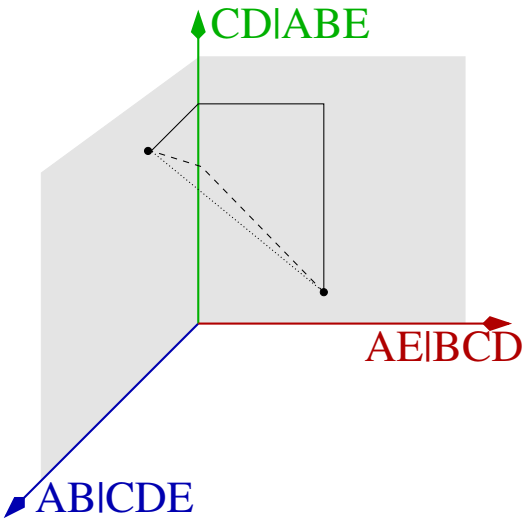
2. When projecting $\mathbb{T}_n$ onto the vector space spanned by the splits compatible to a k-split ($k > 2$), the resulting space has the following structure: $\mathbb{T}_{k+1} \times \mathbb{T}'_{n-k+1}$, where $\mathbb{T}'_{n-k+1}$ is $\mathbb{T}_{n-k+1}$ with one terminal split missing. An example is shown in Figure 1.5.

Furthermore, $\mathbb{T}_n$ is a true subspace of $\mathbb{R}^m$. Already in $\mathbb{T}_4$ we see that the tree





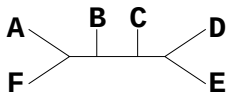
(a) Example of an unresolved topology (green) connecting two resolved topologies

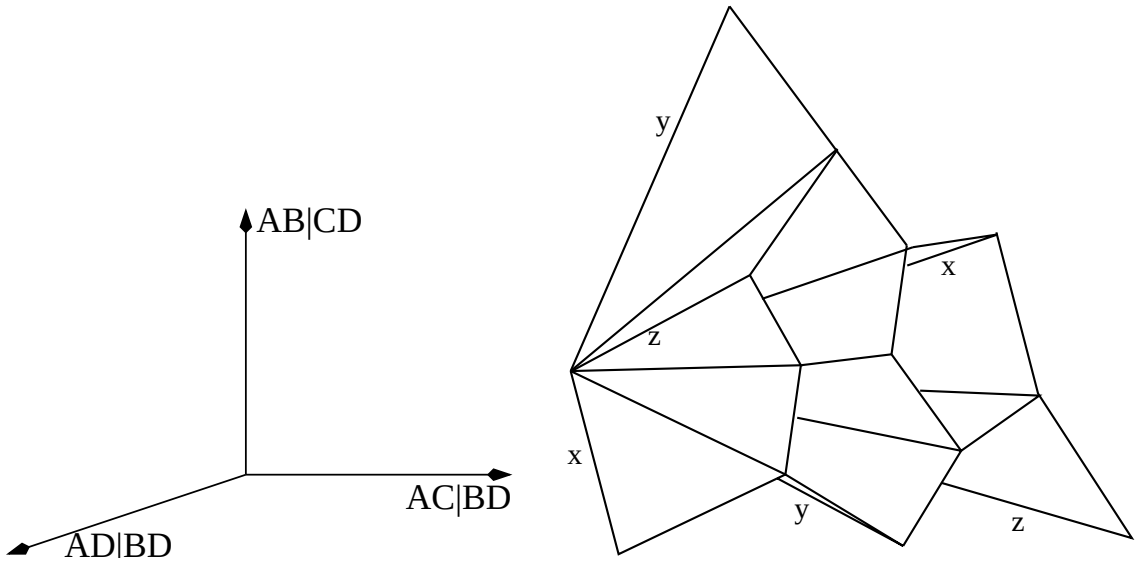


(b) Example of paths in the tree space connecting two weighted trees (points): Manhattan path (line), Euclidean path (dotted) and geodesic path (dashed).

Figure 1.3: Subspace of the tree space for five taxa.

space is sparse. Although the dimension of the space is $m = 3$, each tree has only one internal split. For $\mathbb{T}_5$, the dimension is $m = 15$ but each tree lies on a 2D-plane. This disparity increases for higher dimensions, since the number of splits for one tree increases linearly with n, but the number of possible splits, m, increases exponentially with n.





(a) $\mathbb{T}_4$ **(b)** $\mathbb{T}_5$

Figure 1.4: Tree space for four and five taxa, where only the interior splits are shown. (a) $\mathbb{T}_4$: the split corresponding to each axis is given. Only points on the axes lie in $\mathbb{T}_4$. (b) $\mathbb{T}_5$: Billera *et al.* (2001) introduced this two-dimensional description of the space spanned by the ten nontrivial splits for five taxa. Here, each topology is a 2D-plane. Note that the figure is entangled as some splits (x, y, z) are shown twice at the boundary of the figure.

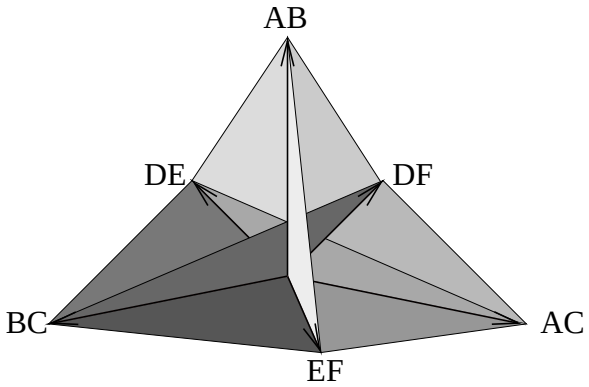
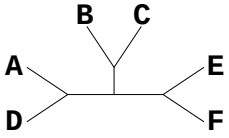


Figure 1.5: Subspace of $\mathbb{T}_6$ showing only the interior splits compatible to the split ABC. This corresponds to a cross-product of the two $\mathbb{T}_4$-spaces with axes $\{AB, AC, BC\}$ and $\{DE, DF, EF\}$, respectively. See Figure 1.2c for the compatibility relationships.



Chapter 2

Distances between Phylogenetic Trees

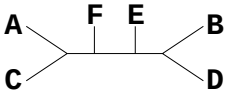
2.1 Overview of Distance Measures

2.1.1 Introduction

Comparing phylogenetic trees is a major task in phylogenetic research. Comparisons are necessary when trees derived from different genes are incongruent (e. g. Rokas and Carroll, 2005), when the outcomes of different reconstruction methods disagree (e. g. Dutilh *et al.*, 2007), or when one compares the outcome of different tree reconstruction methods by simulation (e. g. Gadagkar *et al.*, 2005).

A natural way to compare pairs of trees is to apply a distance measure. Most measures only take the topological information into account, e. g. the Robinson-Foulds distance (Section 2.1.3; Robinson and Foulds, 1981), the nearest neighbor interchange distance (Waterman and Smith, 1978), the subtree prune and regraft distance (Hein, 1990), or the quartet distance (Estabrook *et al.*, 1985). On the other hand, there are a few measures that focus on the branch lengths of the trees, e. g. the weighted Robinson-Foulds distance (Section 2.1.4; Robinson and Foulds, 1978), the branch score distance (Section 2.1.5; Kuhner and Felsenstein, 1994) or the geodesic distance (Section 2.2; Billera *et al.*, 2001). An advantage of distance measures that consider branch lengths information is that they yield continuous values. This increases the distinguishability between different comparisons and allows for applications in the clustering and visualization of trees (Stockham *et al.*, 2002; Hillis *et al.*, 2005; Smythe *et al.*, 2006).

Phylogenetic trees can also be tested for being statistically similar or different. Then two basic settings are distinguished: (1) The trees are tested with respect to



an underlying data set or (2) the trees are tested solely based on their topology and eventually branch lengths. The first setting is extensively used in combination with Maximum Likelihood to test for the difference in likelihood of two topologies (for reviews see Goldman *et al.*, 2000; Schmidt, 2009). Here we will concentrate on the second setting (Section 2.3).

In the following, we present some existing distance measures. All compare trees on the same taxon set of size n.

2.1.2 Size of the Maximum Agreement Subtree

The maximum agreement subtree of two trees is the largest possible subtree identical in both input phylogenies (Finden and Gordon, 1985). Thereby the size is measured by the number of taxa. This can be transformed into the *MAST-based distance* MD:

$$\mathrm{MD}(\mathcal{T}^1, \mathcal{T}^2) = n - \text{MAST size}(\mathcal{T}^1, \mathcal{T}^2)$$

2.1.3 Robinson-Foulds Distance

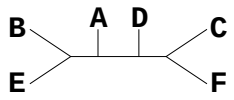
The most common distance measure for the topological difference between two trees is the *Robinson-Foulds distance* (RF, Robinson and Foulds, 1981). It corresponds to the number of splits in each one but not in both topologies. In the set-theoretical sense, the Robinson-Foulds distance between two topologies $\mathcal{T}^1$ and $\mathcal{T}^2$ is given by the size of the symmetric difference:

$$\mathrm{RF}(\mathcal{T}^1, \mathcal{T}^2) = |\mathcal{T}^1 \Delta \mathcal{T}^2| = |(\mathcal{T}^1 \cup \mathcal{T}^2) \setminus (\mathcal{T}^1 \cap \mathcal{T}^2)|.$$

The example topologies in Figure 1.1 (page 5) have $\mathrm{RF}(\mathcal{T}^1, \mathcal{T}^2) = 6$ because each topology has three interior splits and the two trees have no interior split in common.

2.1.4 Weighted Robinson-Foulds Distance

One measure to compare two trees with respect to both topology and branch lengths is the *weighted Robinson-Foulds distance* (RF_w, Robinson and Foulds, 1978). It is the sum of the absolute branch lengths differences for all splits in the trees. Thus, for



two weighted trees λ_1 and λ_2 with topology $\mathcal{T}^1$ and $\mathcal{T}^2$, respectively, the weighted Robinson-Foulds distance is given as:

$$\mathrm{RF}_w(\lambda_1, \lambda_2) = \sum_{\mathcal{S} \in \mathbb{S}_n} |\lambda_1(\mathcal{S}) - \lambda_2(\mathcal{S})|.$$

This measure corresponds to the L^1 norm or the length of the Manhattan path in $\mathbb{T}_n$. An example of the Manhattan path in $\mathbb{T}_5$ is shown in Figure 1.3b (page 11). For the weighted trees in Figure 1.1, the weighted Robinson-Foulds distance is equal to 2.6.

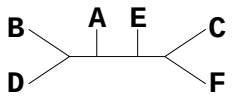
2.1.5 Branch-score Distance

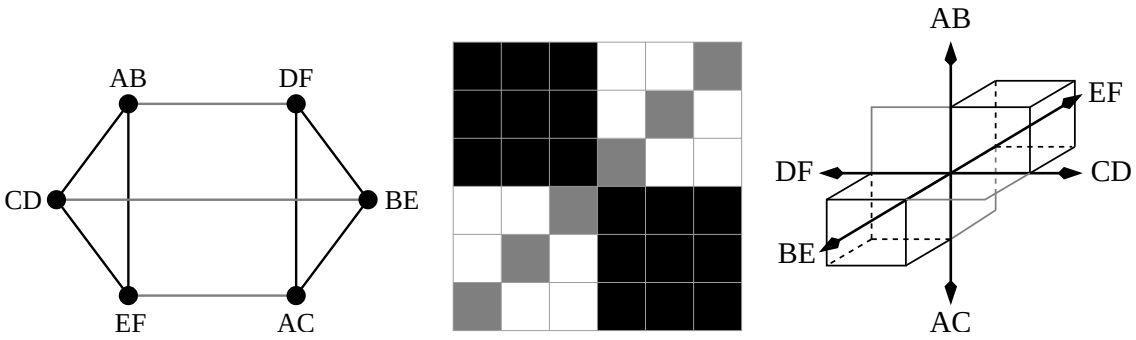
Another measure that respects branch lengths is the *branch-score distance* (BS, Kuhner and Felsenstein, 1994; Felsenstein, 2005). It corresponds to the Euclidean distance between the branch lengths of all splits in $\mathbb{R}^m$:

$$\mathrm{BS}(\lambda_1, \lambda_2) = \|\lambda_1(\mathcal{T}^1) - \lambda_2(\mathcal{T}^2)\| = \sqrt{\sum_{\mathcal{S} \in \mathbb{S}_n} (\lambda_1(\mathcal{S}) - \lambda_2(\mathcal{S}))^2}.$$

Topological information is not incorporated explicitly in this measure, but it is considered implicitly since branch lengths of non-existing splits are zero.

For two different topologies, the corresponding path in $\mathbb{R}^m$ (the Euclidean path) is not a path in $\mathbb{T}_n$ (see example in Figure 1.3b). This implies that the Euclidean distance does not correspond to the L^2-norm on tree space. But Billera *et al.* (2001) have shown that an L^2-norm in $\mathbb{T}_n$ exists by proving that $\mathbb{T}_n$ is a CAT(0)-space (Bridson and Haefliger, 1999). In CAT(0)-spaces, a unique shortest path exists between any two points. These paths are called *geodesics* and their length, the *geodesic distance*, is a metric on $\mathbb{T}_n$ which corresponds to the L^2-norm (see also Section 2.2).





(a) Compatibility graph **(b)** Adjacency matrix **(c)** Space visualization

Figure 2.1: Three visualizations for the example topologies $\mathcal{T}^1$ and $\mathcal{T}^2$ from Figure 1.1. (a) Compatibility map of the six interior splits from the two topologies. Compatibilities between splits of the two topologies are highlighted in gray. We see that no bifurcating topology other than $\mathcal{T}^1$ and $\mathcal{T}^2$ can be formed from these splits. (b) Associated adjacency matrix with the same color code, where white entries denote incompatibility. (c) Part of tree space $\mathbb{T}_6$ spanned by the interior splits in $\mathcal{T}^1$ and $\mathcal{T}^2$ with the same color code. The three gray planes correspond to unresolved topologies spanned by splits from both trees.

2.2 Geodesic Distance

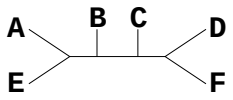
2.2.1 Algorithm to Compute the Geodesic Distance

Preliminary Considerations

In the following, we will provide an algorithm to determine the geodesic path between two trees with the same n taxa. The length of the geodesic path is the L^2-norm in $\mathbb{T}_n$ (Section 2.1.5, Billera *et al.*, 2001). The algorithm was first presented in Kupczok *et al.* (2008).

The dimension of the tree space, m, increases exponentially with n (Section 1.3.2). But given two topologies $\mathcal{T}^1$ and $\mathcal{T}^2$, we only need to consider the splits in these two topologies since the geodesic path will never pass other splits (Vogtmann, 2003). Figure 2.1 depicts the implications of this statement for the trees in Figure 1.1 (page 5).

For trees containing common internal splits, the problem of finding the geodesic path can be simplified further: for $\mathcal{S}_c \in \mathcal{T}^1 \cap \mathcal{T}^2$ with $\mathcal{S}_c = X_A|X_B$, each internal split in $\mathcal{T}^1$ further resolves either taxon set X_A or X_B. We then call the corresponding split sets $\mathcal{T}^1_A$ and $\mathcal{T}^1_B$, respectively. These two split sets are subtopologies on the taxon sets



X_A resp. X_B (analogous for $\mathcal{T}^2$). Since all splits from subtopology $\mathcal{T}_A^i$ are compatible with all splits from subtopology $\mathcal{T}_B^j$, $i, j = 1, 2$, paths through these subtopologies are independent. Therefore, the geodesic for all splits can be found by looking separately at taxon set X_A (using the subtopologies $\mathcal{T}_A^1$ and $\mathcal{T}_A^2$) and taxon set X_B (using $\mathcal{T}_B^1$ and $\mathcal{T}_B^2$) and assembling the paths afterwards (Vogtmann, 2003). As a consequence, we will assume in the following that the topologies are fully decomposed and contain no common splits. This involves both reducing the topologies and setting the other split weights (including the terminal splits) to zero.

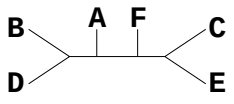
Another useful property of geodesic paths is that they are *piecewise linear*. In Figure 1.3b (page 11) the geodesic path between two trees with only one different split is shown. This path is linear between one tree and the intersection point with the axis. Therefore, the idea of the algorithm presented here is to enumerate all possible intersections efficiently, to compute the length of a path given special intersections and to find the shortest among these paths.

The Split Set

We present how to compute the geodesic path between two weighted trees λ_1 and λ_2 with topologies $\mathcal{T}^1$ and $\mathcal{T}^2$, respectively. As explained, we already decomposed the topologies such that $\mathcal{T}^1$ and $\mathcal{T}^2$ contain no common splits. We first assume that the respective topologies are bifurcating and discuss multifurcating trees at the end of this section. In the bifurcating case both topologies contain the same number of splits, i. e. $d = |\mathcal{T}^1| = |\mathcal{T}^2|$. Thus, we reduced $\mathbb{T}_n$ to a $2d$-dimensional subspace by deleting all splits not in $\mathcal{T}^1 \cup \mathcal{T}^2$. The remaining split set is $\mathbb{S}'$ with $|\mathbb{S}'| = 2d$. Note that the splits in $\mathbb{S}'$ are exactly the splits contributing to the Robinson-Foulds distance, and that $2d$ corresponds to the RF distance for the reduced topologies.

Example: The trees in Figure 1.1 on taxon set $X = \{A, B, C, D, E, F\}$ are each composed of three interior splits. They do not have interior splits in common, and therefore $d = 3$, $\mathbb{S}' = \{AB, CD, EF, AC, BE, DF\}$ and

$$\mathcal{T}^1 = \{AB, CD, EF\}, \quad \mathcal{T}^2 = \{AC, BE, DF\}$$



with the following weights for the topologies:

$$\lambda_1(\mathcal{T}^1) = (0.1, 0.2, 0.3, 0, 0, 0), \quad \lambda_1(\mathcal{T}^2) = (0, 0, 0, 0, 0, 0),$$
$$\lambda_2(\mathcal{T}^1) = (0, 0, 0, 0, 0, 0), \quad \lambda_2(\mathcal{T}^2) = (0, 0, 0, 0.9, 0.6, 0.5).$$

Legal Topologies

We construct *legal topologies* $\mathcal{A}$ that are formed by the splits in $\mathbb{S}'$ and fulfill the compatibility condition. A legal topology $\mathcal{A}$ is required to be *maximal*, i. e. adding splits $\mathcal{S} \in \mathbb{S}'$ to $\mathcal{A}$ will violate the compatibility within $\mathcal{A}$. This corresponds to extracting the maximal cliques from the compatibility graph with node set $\mathbb{S}'$. Non-maximal topologies are by definition composed of fewer splits and are therefore subtopologies of a maximal topology. This implies that any path through legal subtopologies runs along the boundary of a legal topology. Thus, paths through subtopologies are already contained in the possible paths through other legal topologies by setting the corresponding split weights to 0.

Let $\mathbb{A}$ be the set of all legal topologies. By this definition, $\mathcal{T}^1 \in \mathbb{A}$ and $\mathcal{T}^2 \in \mathbb{A}$.

Example: For the trees in Figure 1.1, the compatibility graph is shown in Figure 2.1a. The set $\mathbb{A}$ of legal topologies comprises the maximal cliques of the graph, thus

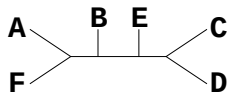
$$\mathbb{A} = \{\mathcal{T}^1, \quad \mathcal{T}^2, \quad \{AB, DF\}, \quad \{CD, BE\}, \quad \{EF, AC\}\}$$

Sequences of Legal Topologies

A *sequence of legal topologies* connects $\mathcal{T}^1$ with $\mathcal{T}^2$ by passing through legal topologies from $\mathbb{A}$ in such a manner that in each step at least one split from $\mathcal{T}^1$ is replaced by at least one split from $\mathcal{T}^2$.

Accordingly, a sequence of legal topologies $(\mathcal{A}^j)_{j=0}^k$ with $k \leq d$ and $\mathcal{A}^0 = \mathcal{T}^1$, $\mathcal{A}^k = \mathcal{T}^2$ must fulfill the following two conditions for all $j = 0, \ldots, k-1$:

$$\mathcal{A}^{j+1} \cap \mathcal{T}^1 \subset \mathcal{A}^j \cap \mathcal{T}^1 \text{ and } \mathcal{A}^j \cap \mathcal{T}^2 \subset \mathcal{A}^{j+1} \cap \mathcal{T}^2 \tag{2.1}$$



i. e. no split from $\mathcal{T}^1$ can re-emerge in the sequence and no split from $\mathcal{T}^2$ can be lost.

From the adjacent topologies $\mathcal{A}^j$ and $\mathcal{A}^{j+1}$ we are interested in the splits that change their membership between these sets. These *transitions* I^j ($j = 0, \ldots, k-1$) are given as the symmetric difference between these two sets, i. e. $I^j = \mathcal{A}^j \Delta \mathcal{A}^{j+1}$. Note that the series of transitions $(I^j)_{j=0}^{k-1}$ form a partition of the split set $\mathbb{S}'$.

From this sequence of topologies, we generate a piecewise linear path. The path is linear while passing through a topology $\mathcal{A}$. Two adjacent topologies are connected by a *transition point* where all splits in I^j have weight 0.

Legal Paths

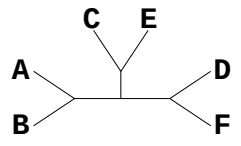
A path is parameterized with constant speed by a piecewise linear function $g : [0, 1] \to \mathbb{R}^{2d}$ with $g(0) = \lambda_1(\mathcal{T}^1)$ and $g(1) = \lambda_2(\mathcal{T}^2)$ (Vogtmann, 2003; Bridson and Haefliger, 1999). For each transition from topology $\mathcal{A}^j$ to $\mathcal{A}^{j+1}$, there exists a time t_j at which $g(t_j)_i$ is zero for the splits $i \in I^j$. In other words, t_j is the *transition time* in which the splits $i \in \mathcal{T}^1 \cap I^j$ are reduced to length zero and $1 - t_j$ is the transition time in which the splits $i \in \mathcal{T}^2 \cap I^j$ are expanded from zero to their weight in λ_2. Due to the constant speed condition of the path (Vogtmann, 2003), the transition times t_j are calculated from the transitions by

$$t_j = \frac{\|\lambda_1(I^j)\|}{\|\lambda_1(I^j)\| + \|\lambda_2(I^j)\|}.$$

For a sequence of topologies to contain a *legal path*, the transition times $(t_j)_{j=0}^{k-1}$ must be increasing with k. This condition ensures that the sequence of topologies is visited in the proposed order, and thus, only legal topologies are traversed. For a legal path g, the entries of g for an arbitrary time $t \in [0, 1]$ are given by

$$g_i(t) = \left\{ \begin{array}{ll} -\frac{\lambda_1(i)}{t_j}(t - t_j), & i \in \mathcal{T}^1 \cap I^j \text{ and } t < t_j, \\ \frac{\lambda_2(i)}{1-t_j}(t - t_j), & i \in \mathcal{T}^2 \cap I^j \text{ and } t > t_j, \\ 0, & \text{otherwise} \end{array} \right\}.$$

Herewith, the function g describes the path between the two weighted trees, which



changes direction at $(t_j)_{j=0}^{k-1}$. Its length is computed by:

$$\|g\| = \sum_{j=1}^{k} \|g(t_j) - g(t_{j-1})\|, \quad \text{with } t_0 = 0 \text{ and } t_k = 1.$$

There is always a legal path defined by a single transition at time t^* with $I^* = \mathbb{S}'$. This path simultaneously replaces all splits in $\mathcal{T}^1$ by all splits in $\mathcal{T}^2$ at time t^* and is called *cone path*, because it passes through the origin of the $2d$-dimensional subspace of $\mathbb{T}_n$. If decompostion can be applied there is always another legal path, the *decomposed cone path*. It passes independently through the origins of each subspace originating from the independent decompositions. The cone path always has only one transition point, whereas the decomposed cone path has one transition point for each decomposition. Therefore, its length lies between the cone path length and the geodesic path length.

Example: The example set $\mathbb{A}$ from Figure 1.1 suggests four sequences of legal topologies with the following transitions and transition times:

Path 1: $\mathcal{T}^1 \xrightarrow[t_1=0.42]{\{CD,EF,AC\}} \{AB, DF\} \xrightarrow[t_2=0.09]{\{AB,AC,BE\}} \mathcal{T}^2$

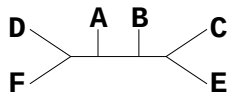
Path 2: $\mathcal{T}^1 \xrightarrow[t_1=0.35]{\{AB,EF,BE\}} \{CD, BE\} \xrightarrow[t_2=0.16]{\{CD,AC,DF\}} \mathcal{T}^2$

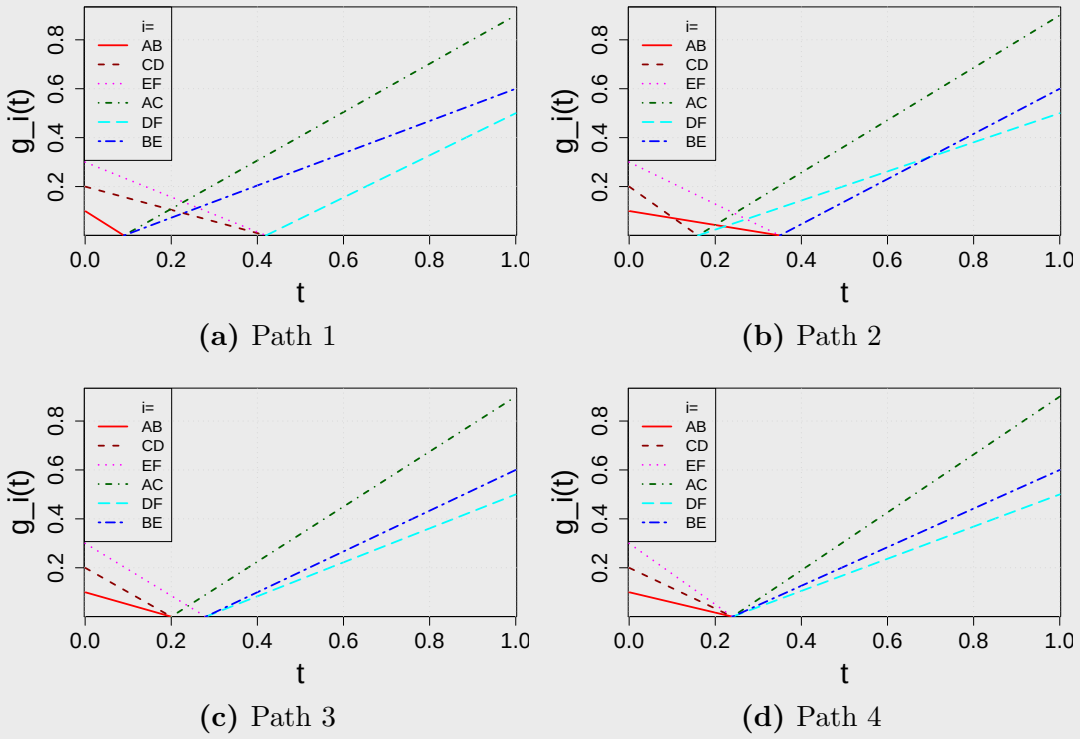
Path 3: $\mathcal{T}^1 \xrightarrow[t_1=0.2]{\{AB,CD,AC\}} \{EF, AC\} \xrightarrow[t_2=0.28]{\{EF,BE,DF\}} \mathcal{T}^2$

Path 4: $\mathcal{T}^1 \xrightarrow[t^*=0.24]{\{AB,CD,EF,AC,BE,DF\}} \mathcal{T}^2$

Because $t_1 > t_2$ for the first sequence, it does not yield a legal path. This is visualized in Figure 2.2a: The topologies suggested by the transition times are $\mathcal{T}^1 \xrightarrow[t_2=0.09]{\{AB,AC,BE\}} \{CD, EF, AC, BE\} \xrightarrow[t_1=0.42]{\{CD,EF,AC\}} \mathcal{T}^2$. So four splits coexist between time 0.09 and 0.42, where the splits from the first tree are incompatible with the splits from the second tree. Thus, the condition of a legal path can be checked by testing whether the times in a sequence of topologies are increasing.

Since $t_1 \leq t_2$ does also not hold for the second sequence (Figure 2.2b), only path 3 (Figure 2.2c, length 1.5592) and path 4 (Figure 2.2d, the cone path, length 1.5658) correspond to legal paths. In this example, there is no decomposed cone path since decomposition is not applicable. Path 3 is the geodesic path between the two trees.





(a) Path 1 **(b)** Path 2

(c) Path 3 **(d)** Path 4

Figure 2.2: Parameterizations for the four possible paths of the example trees in Figure 1.1.

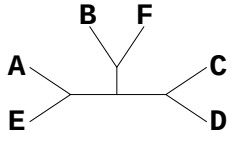
Multifurcating Trees

Only one extension of our findings is necessary to include trees with less than $n-3$ interior splits, i. e. multifurcating trees. If $\mathcal{T}^1$ is multifurcating, there may be splits in $\mathcal{T}^2$ that are compatible to each split in $\mathcal{T}^1$ and vice versa. The length of a split in $\mathcal{T}^2$ with this property will be extended immediately from the beginning, while the length of a split in $\mathcal{T}^1$ with this property will be reduced until the end. Thus, only the remaining topologies $\hat{\mathcal{T}}^1$ and $\hat{\mathcal{T}}^2$ without these splits are relevant for the calculation of the transition times and contribute to the topologies in $\mathbb{A}$. Note that $\hat{\mathcal{T}}^1$ and $\hat{\mathcal{T}}^2$ do not necessarily contain the same number of splits.

Implementation Details

The presented algorithm for the geodesic path comprises several steps:

1. Decomposing the topologies into the sets $\mathcal{T}^1$ and $\mathcal{T}^2$,



2. building the legal topologies $\mathbb{A}$,

3. arranging the legal topologies in legal sequences with transitions,

4. extracting the legal paths from the legal sequences (these are the sequences where the transition times are in the correct order) and

5. finding the shortest among these legal paths, this is the geodesic path.

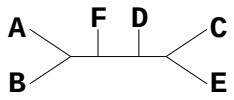
Our implementation does not follow these steps, but computes the legal topologies together with the transitions and their respective times. The computation starts with $\mathcal{T}^1$ and generates all possible transitions I, which lead to a maximal legal topology. A directed acyclic graph (DAG) is thereby generated whose node set is $\mathbb{A}$ and an edge is inserted for each generated transition and labeled with its time. A legal sequence is a directed path in the DAG which connects $\mathcal{T}^1$ and $\mathcal{T}^2$, and a legal path is a legal sequence with increasing edge weights on the path through the DAG. Not all sequences have to be enumerated until the end. The transition times are computed co-instantaneously and tested for an ascending sequence. Illegal paths are identified and terminated before reaching $\mathcal{T}^2$.

The time-limiting step is the generation of all transitions I, which is done for each topology in $\mathbb{A}$. First, for topology $\mathcal{T}^1$ all possible I leading to a maximal legal topology are generated. There are not more than 2^d of these transitions. Then the complete set $\mathbb{A}$ is already known, and for each element all possible transitions are generated again. This yields

$$\underbrace{2^d}_{\text{Generation of } \mathbb{A} \text{ from } \mathcal{T}^1} + \underbrace{2^d \times 2^d}_{\text{Generation of all } I \text{ for the other topologies}} = \mathcal{O}(2^{2d})$$

for building the graph. But note that because of incompatibilities there are much less then 2^d legal topologies and the size of $\mathbb{A}$ is much smaller than 2^d. Furthermore, the algorithm is exponential in d, which is small for topologically similar trees and which can be decreased by decomposing the trees.

The algorithm is implemented in a python program called `GeoMeTree` (Geodesic Metric on Trees) available from `www.cibiv.at/software/geometree`. The program has been used for the calculations in the next section.



2.2.2 Simulation Study

The Data Set

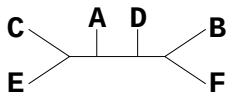
A data set was generated from 216 alignments of 20 metazoa species and yeast as an outgroup (Ewing *et al.*, 2008; Ebersberger, 2007). The orthologs were extracted from the Inparanoid database (O'Brien *et al.*, 2005), where pairwise orthologs of eukaryotes are stored. Orthology is expanded to cover all 21 species by taking an arbitrary order of the 21 species and determining the orthologous pairs between neighboring species. If a chain occurs where the protein in the first and last species are also orthologs this protein is added to the data set. This resulted in 216 alignments for 21 species, where each alignment consists of putative orthologs. For these, alignments were produced with T-coffee (Notredame *et al.*, 2000). Maximum likelihood gene trees were reconstructed for each of the gene alignments with phyML (Guindon and Gascuel, 2003). A widely accepted species tree had been found for these gene trees with different methods (Ewing *et al.*, 2008).

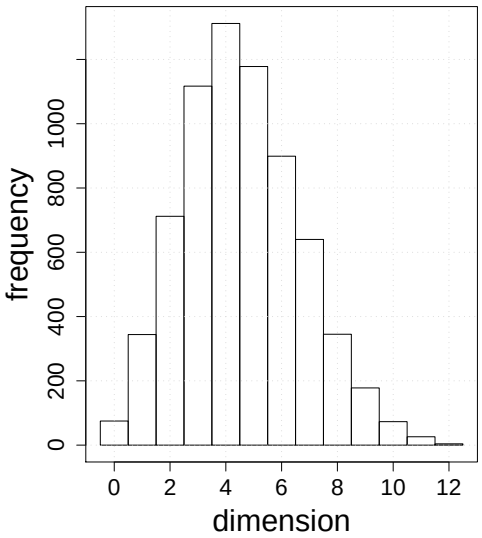
From the resulting set of 216 trees, we used the 118 strictly bifurcating trees (6903 pairs) for the distance computations. The resulting weights for each tree are normalized (i. e. each has a Euclidean norm of 1). Otherwise the branch lengths are expected to dominate the distance between two trees, while differences between their topologies have less influence on the measure.

Dimension and Number of Paths

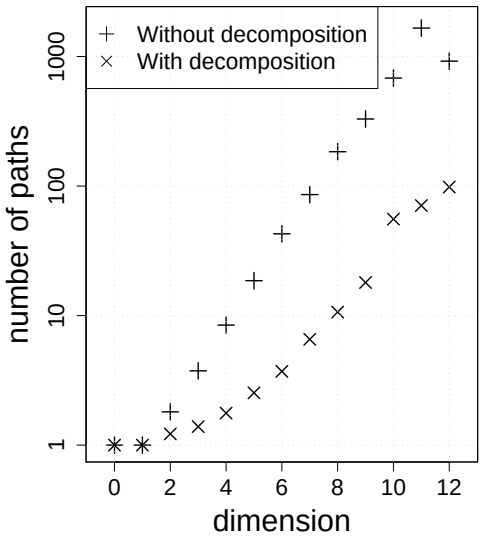
The computations were first done without decomposition. As stated earlier, the dimension d is the number of splits in one topology but not in the other. For pairs of bifurcating trees, d corresponds to half of the Robinson-Foulds distance (Section 2.1.3). For 21 taxa, the maximal possible value of d is 18, but the observed dimension d in the data ranged from 0 to 12 (Figure 2.3a). The few pairs with high dimension d are mainly caused by a few gene trees with many incongruencies to the species tree.

Without decomposition, the mean number of legal paths in tree space increases exponentially with the dimension (Figure 2.3b). This is due to the fact that the number of legal topologies increases exponentially with d and more legal paths are expected for a larger set of legal topologies. A substantially smaller number of paths is explored when the topologies are decomposed: The numbers with decomposition





(a) Frequency (number of pairs of trees) of the observed dimension d



(b) Mean number of paths (log-scale) for the computation without decomposition and the product of the paths through independent decompositions

Figure 2.3: Frequence of observed dimensions and mean number of paths

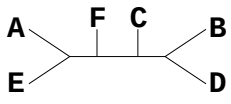
in Figure 2.3b refer to the product of the number of paths through the independent decompositions, which is smaller than the complete number of paths.

Computing Time

Although the algorithm is exponential in d, the program is reasonably fast. The mean runtime without decomposition was 0.4 s. However, the time to compute the distance for one pair of trees strongly depends on the number of paths evaluated. This is reflected by the longer runtime for high dimensions and without decomposition (Table 2.1, left part). The runtime is highly improved when decomposition of the trees is applied (Table 2.1, right part). For the two runs which then show a runtime of > 30 s, the trees could not be decomposed.

Approximations of the Geodesic Distance

A lower and an upper bound is known for the geodesic distance (Amenta *et al.*, 2007): The lower bound is the branch score distance (Section 2.1.5) and the upper bound is



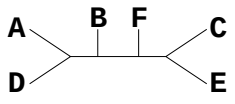
	Without Decomposition		With Decomposition			
	Time				Dec.	
d	mean	max	mean	max	mean	max
0	0 s	0 s	0 s	0 s	1	1
1	0 s	0 s	0 s	0 s	1	1
2	0 s	0 s	0 s	0 s	1.6	2
3	0 s	0 s	0 s	0 s	2.2	3
4	0 s	0.1 s	0 s	0.1 s	2.6	4
5	0.2 s	1 s	0 s	0.2 s	2.9	5
6	0.9 s	9.9 s	0 s	0.7 s	3	5
7	3.7 s	102 s	0 s	0.8 s	2.7	6
8	15.8 s	11.4 m	0.1 s	7.7 s	2.7	6
9	56.1 s	33.7 m	0.2 s	22.6 s	2.4	5
10	8.6 m	9.9 h	2.1 s	46 s	2.1	4
11	43.7 m	17.9 h	2.7 s	44.3 s	2	4
12	2.3 m	5.4 m	3.9 s	13.9 s	1.5	2

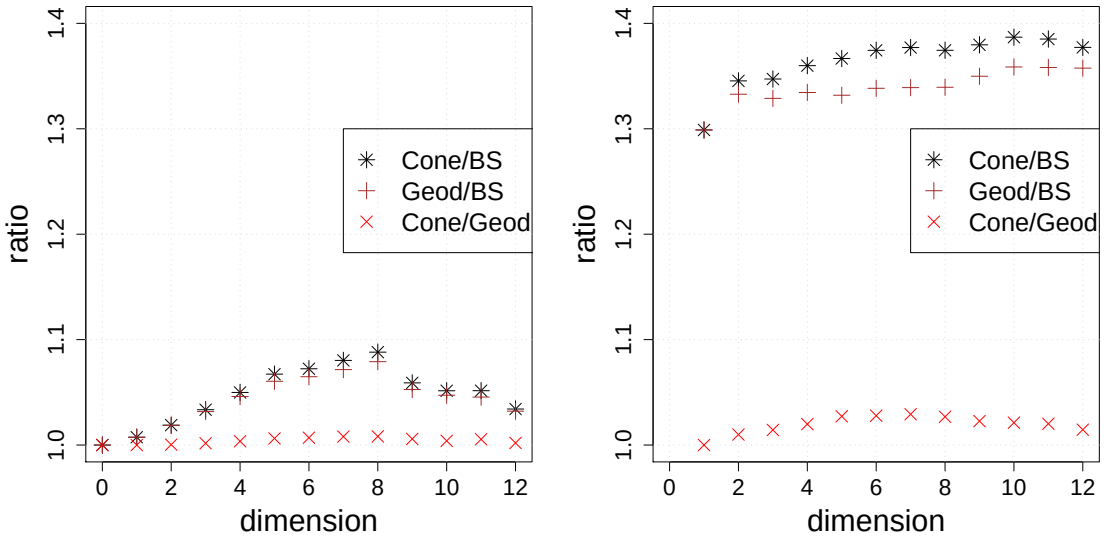
Table 2.1: Time: Mean and maximal time consumption for the different dimensions d without and with decomposition. Dec.: Mean and maximal number of decompositions for a dimension d.

the length of the cone path. Amenta *et al.* (2007) showed that these two lengths differ at most by a factor of $\sqrt{2}$.

We observed that the mean ratio of the cone path to the branch score distance is close to 1.4, when the distance is computed only from the differing splits (Figure 2.4b). In contrast, the ratio is smaller than 1.1 when all splits are considered (Figure 2.4a). Thus, both distances give a tight interval for the geodesic distance. Figure 2.4 also shows, that the geodesic distance is better approximated by the cone path than by the branch score distance. This is expected, since the cone path is already a path in tree space. In contrast, the branch score distance measures the length of the Euclidean path between two trees, which is not a path in tree space for trees with at least one different split.

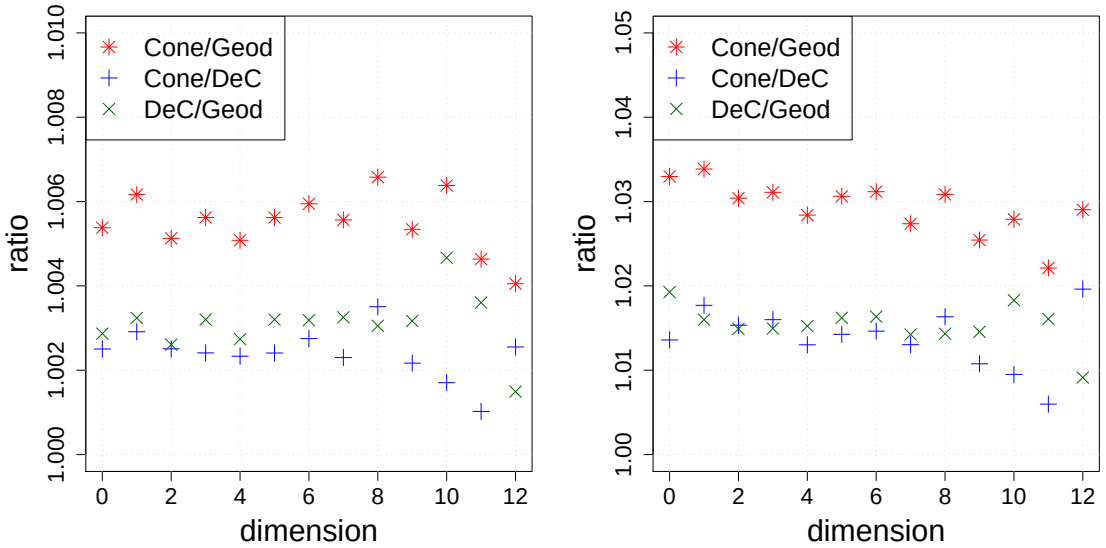
The approximation with the cone path can be further improved by applying the decomposed cone path (Figure 2.5). Both for different and for all splits the approximation is on average improved by about a factor of two.





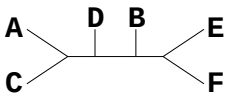
(a) All splits **(b)** Different splits

Figure 2.4: Means of the ratios of the distance measures. "Cone": cone path length; "BS": branch score distance; "Geod": geodesic distance.



(a) All splits **(b)** Different splits

Figure 2.5: Means of the ratios of the distance measures. "Cone": cone path length; "DeC": decomposed cone path length; "Geod": geodesic distance. Note the different scaling of the y-axes.



Relationship to the Robinson-Foulds distance

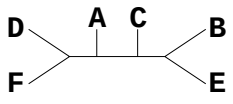
The Robinson-Foulds distance (Section 2.1.3) is a pure topological and discrete distance measure, which counts the number of different splits between two trees and thus corresponds to $2d$ in our notation. Since it is the prevalent distance measure for phylogenetic trees, it would be preferable if continuous distance measures also displayed this topological information and extended it with the branch lengths information.

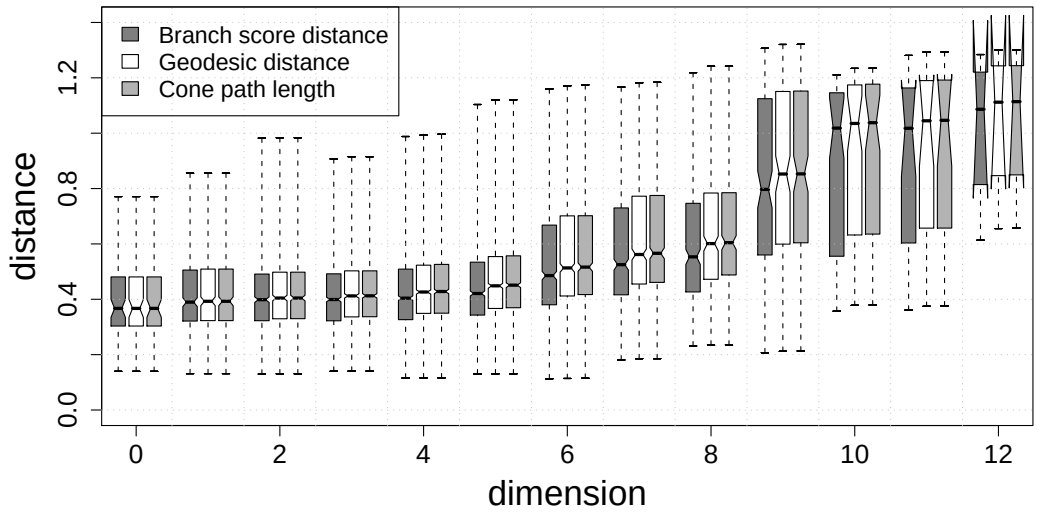
As Figure 2.6a indicates, comparing pairs of trees over all their splits results in a distance range which appears to be hardly related to the Robinson-Foulds distance especially for small values of d. The notches of the boxplots of different dimensions are overlapping for the geodesic distance (and also for its approximations). This indicates that the medians do not differ significantly. Under these circumstances the lengths of the branches have a much higher influence on the distance than the topological features. If one intends to make the comparison more sensitive to topological differences, we suggest reducing the study to the $2d$ different splits (see Figure 2.6b). Here, the notches for each distance method do not overlap up to a dimension of nine, indicating that the median geodesic distance is increasing with increasing dimension. However, the broad distributions show that the branch lengths do still have a substantial impact.

2.2.3 Conclusions

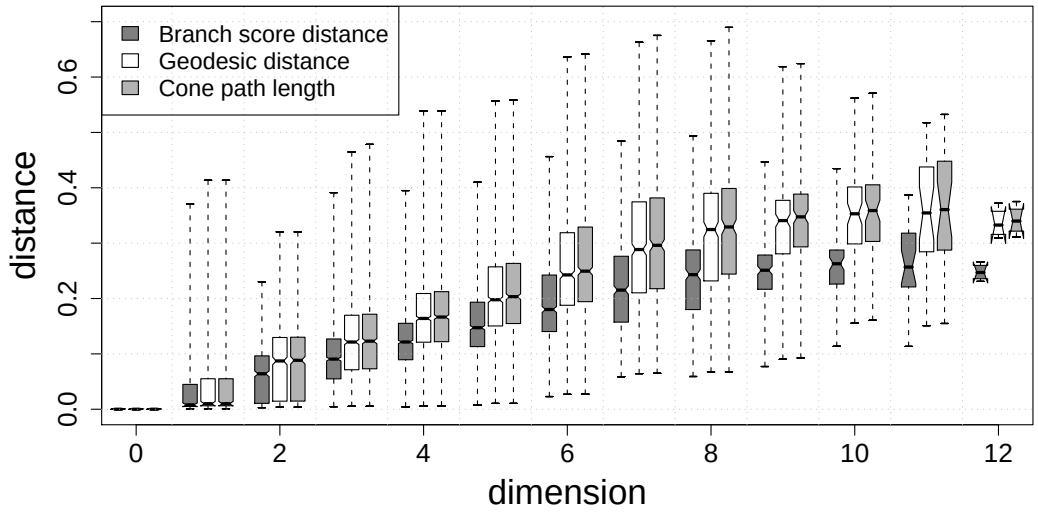
We presented an exact algorithm for computing the geodesic distance and showed its applicability for phylogenetic trees. For a pair of trees, the algorithm constructs legal topologies formed from splits of the input trees. From these topologies it enumerates the legal paths leading from one tree through a sequence of legal topologies to the other tree. We employed computational techniques to reduce the number of paths enumerated. This facilitated the calculation of the geodesic path between two trees in reasonable time, although the algorithm is still exponential in the number of different splits.

Megan Owen developed another algorithm for that problem independently (Owen, 2008). Furthermore, a polynomial time algorithm appeared recently (Owen and Provan, 2010). Our algorithm, however, constitutes the first available implementation. Furthermore, we used this implementation to emprically compare the geodesic distance with its approximations.



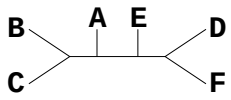


(a) All splits



(b) Different splits

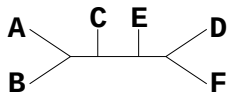
Figure 2.6: Relation between the Robinson-Foulds distance and the three continuous distances: The Robinson-Foulds distance is two times the dimension d of the pair. For each category of d, a boxplot of the distribution of each of the three distances is shown.



We showed the applicability of the presented algorithm on a metazoa dataset which was generated from 118 alignments of 21 species (Ewing *et al.*, 2008; Ebersberger, 2007). In this example, the contribution of the branch lengths outweights the influence of the topologies (Figure 2.6a) on the distance. To incorporate the topological signals, we suggest as an appropriate distance measure the length of the geodesic path through the splits exclusive to one of the trees considered (Figure 2.6b).

Another notable observation is the small factor by which the cone path and the geodesic distance differ (Figure 2.4). Thus, the length of the cone path is a useful approximation of the geodesic distance, especially since it incorporates topological differences in a similar way. The previous finding of only using splits exclusive to one of the trees as a distance can also be applied to the cone distance (Figure 2.6b). The approximation is further improved by using the length of the decomposed cone path (Figure 2.5).

The availability of a distance metric in tree space allows us to address further issues. These include clustering or visualizing trees (e. g. Hillis *et al.*, 2005), or finding the center of a set of trees (Billera *et al.*, 2001), which can be interpreted as a consensus method. One possibility to define a consensus method for a given distance metric are median trees. For the Robinson-Foulds distance, the median tree corresponds to the majority-rule consensus tree (Barthélemy and McMorris, 1986), which is one of the prevalent consensus methods (Bryant, 2003). For the weighted Robinson-Foulds distance, the median tree is given by the majority-rule consensus tree in which each branch length is the minimum of the lengths of the respective split in the different trees (Pattengale, 2005). The median tree for a given set of trees under the geodesic distance corresponds to the tree with the smallest total geodesic distance to all trees in the regarded set. A closed formula to determine the median tree under the geodesic distance is unknown to date, thus searching over the whole tree space would be necessary. For simplification, one could assume that the median tree is among the observed trees and thus determine the tree with the smallest total distance from all pairwise distances.



2.3 Testing Phylogenetic Trees based on Distances

2.3.1 Introduction

Testing trees only based on their topology is closely related to tree distances: Every distance measure can be used as a test statistic where the associated p-value is computed under a null distribution of trees. de Vienne *et al.* (2007) suggest the corresponding test using the MAST-based distance (Section 2.1.2) and the null model of equally likely topologies (see also Section 1.3.2). Since computing the associated p-values involves intensive resampling from the topology space, they also introduce approximative formulas. Kupczok and von Haeseler (2009) criticize this approximation for not being statistically sound for some numbers of taxa. However, de Vienne *et al.* (2009) argue that the method is fast and provides approximate p-values close to the correct ones.

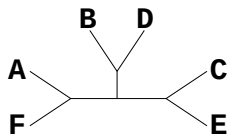
Here, we will reproduce the arguments of Kupczok and von Haeseler (2009) in detail. We will refer to the test in de Vienne *et al.* (2007) as the *MAST test* and outline it first.

2.3.2 The MAST Test

Remember that the size of the maximum agreement subtree (the *MAST size*) is the number of taxa in the largest possible subtree identical in both input phylogenies (Section 2.1.2). Its null distribution can be obtained by generating pairs of trees, where trees are assumed to be equally likely, and evaluating their MAST size. From this null distribution, de Vienne *et al.* (2007) estimated functions for the mean and standard deviation of the MAST size depending on the number of taxa n. Thereby, they confirmed the results of Bryant *et al.* (2003) that the mean MAST size grows proportionally to the square root of n (Figure 2.7). E. g. for $n = 50$, the mean MAST size is 10, thus on average 40 taxa (4/5 of all taxa) are pruned from random trees.

The test statistic of the MAST test for two trees of n taxa is the MAST size centered by the mean for n and rescaled by the standard deviation for n. The resulting standardized distribution for $7 \leq n \leq 50$ is then used to fit an analytical curve to the left tail of the distribution. With this curve, p-values up to 0.05 can be estimated.

Using the centered and rescaled MAST size as a test statistic causes several problems. First, only the taxa in the MAST contribute to the significance while the topo-



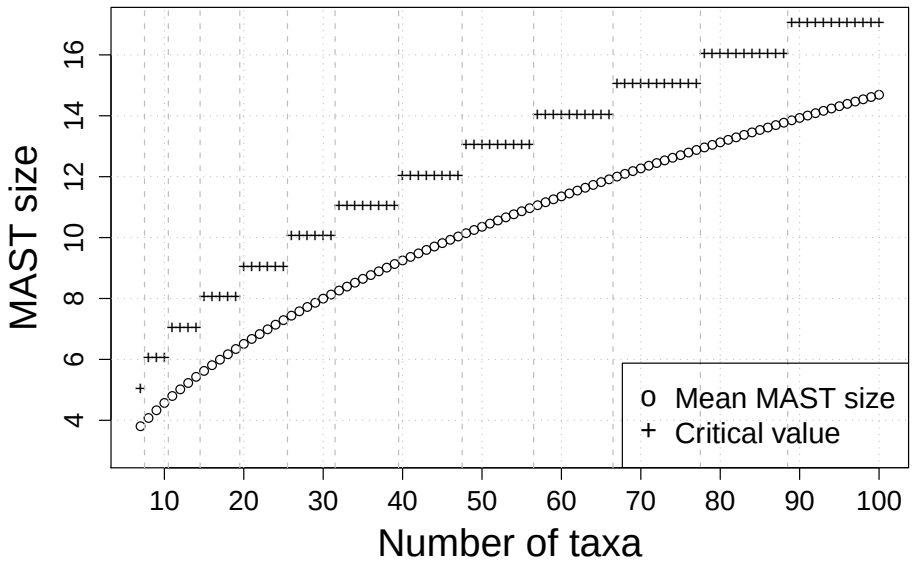
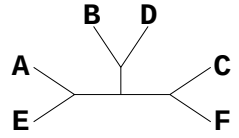


Figure 2.7: Mean MAST size and critical value of the MAST size for $\alpha = 0.05$: The mean is given by equation (1) in de Vienne *et al.* (2007) and the critical value is computed with equation (6) in de Vienne *et al.* (2007). The vertical lines indicate the steps in the critical value.

logical information of the others is ignored. In a biological framework, however, this may not use all the information present in the topologies. Second, the mean MAST size increases only with the square root of the number of taxa n. Hence the mean *relative* MAST size (the MAST size divided by n) approximates zero with increasing n. That means, for two large trees on average a high proportion of taxa is pruned to obtain the MAST. Our main concern when appplying the MAST test is, however, a statistical one. Therefore we first introduce the terminology for discrete testing.

2.3.3 Discrete Testing

When applying a statistical test, a *significance level* α has to be chosen in advance (usually 5 %). This implies that not more than 5 % of the tests are rejected under the null hypothesis. The actual fraction of significant results for a predefined α is known as the *size of a test*. In the ideal case, the significance level equals the test size. For discrete tests, however, the size will rarely match α exactly since the sum of probabilites for the extreme cases grows in discrete steps. The *critical value* c is the cutoff value for a predefined significance level which marks the border of the rejected hypotheses. For the critical value and all more extreme values, the null hypothesis



is rejected. If it corresponds to the last value where the cumulative density function is $\leq \alpha$, the test statistic is said to be *conservative*. On the other hand, if there are more significant results than the predefined significance level, the test statistic is *liberal*. Only if the test statistic is conservative, we can be sure that a significant result or a more extreme case occurs under the null hypothesis not more often than the significance level.

We will demonstrate these terms using the well-known binomial distribution. This distribution describes the number of successes in a sequence of n independent Bernoulli experiments, each of which yields success with probability p and failure with probability $1 - p$. E. g. for $n = 7$ and $p = 0.5$, 7 successes occur with a probability of 0.78 %, whereas at least 6 successes occur with a probability of 6.25 % (Figure 2.8a). Thus, for a cutoff value of 6, the cumulative density function already exceeds 5 %. For $\alpha = 5$ %, the conservative critical value is 7, and thus the size is only 0.78 %.

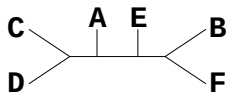
The significance could also be computed in analogy to the MAST test. Then, mean and standard deviation are computed for $7 \leq n \leq 50$ using the common formulas. For each n, 1000 random values are generated and significance is assigned to the values in the 5 %-quantile of the distribution of the centered and rescaled values combined for all n (Figure 2.8b). Then, the cutoff value is -1.67 (red line in Figure 2.8b). For $n = 7$ this results in a critical value of 6, thus the resulting test size of 6.25 % is liberal.

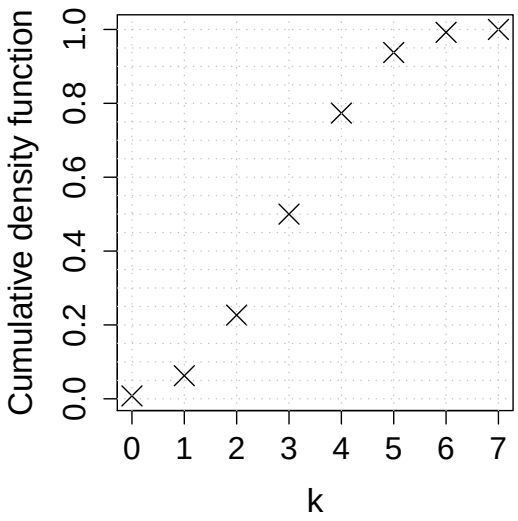
2.3.4 Investigating the MAST Test

Random Trees

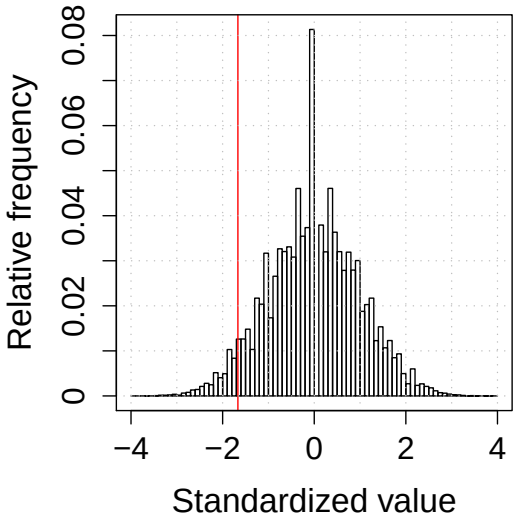
As described in Section 2.3.3, the size of a conservative test must not exceed the significance level α. To determine the size of the MAST test, we first compute the critical value of the MAST size for $\alpha = 0.05$ (Figure 2.7). E. g. for $n = 50$, the critical value is 13, thus when pruning 37 ($\approx 3/4$) or less of the taxa, the trees are considered to be congruent. This high proportion is counterintuitive, but results from the vast number of trees for large n and the fact already shown that the mean MAST size grows slower than n (Section 2.3.2).

The size of the MAST test cannot be computed analytically as for the binomial distribution. We approximate it by simulating 10,000 pairs of random trees, where all trees are equally likely (the PDA model, Section 1.3.2). For the resulting pairs of





(a) Cumulative density function for $n = 7$.



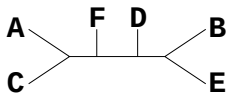
(b) Histogramm of standardized random values (1000 random values for each $7 \leq n \leq 50$). The red line marks the 5 %-quantile.

Figure 2.8: Binomial distribution for $p = 0.5$.

trees, the MAST size is constructed with the algorithm of Goddard *et al.* (1994) as implemented in PAUP* (Swofford, 2002).

The average size of the MAST test for all n and for $7 \leq n \leq 50$ is 0.043 and 0.048, respectively, and thus below the significance level of 0.05. However, in Figure 2.9 we see that the size exceeds the significance level for some n. In these cases, the estimate of the conservative size is much smaller but the corresponding critical value is only one taxon less than the critical value obtained from the MAST test (Figure 2.7). E. g. for $n = 7$, the critical value of the MAST test is 5, but the corresponding size is 12.7 %, whereas 6 or 7 has only been observed in 0.8 % of the cases. Thus 0.8 % is the correct size of the conservative test with a critical value of 6. We suggest to modify the MAST test for the cases where the conservative size is smaller than the original size in Figure 2.9. In these cases the critical value has to be incremented by one to result in the *conservative critical value*. We will refer to this test as the *modified MAST test*. For the random trees, the modified MAST test results in an average size of 0.024 and 0.021 for all n and for $7 \leq n \leq 50$, respectively.

We observe that the original size exceeds the significance level when the number of taxa approaches the right boundary of an area delimited by vertical lines (Figure 2.9).



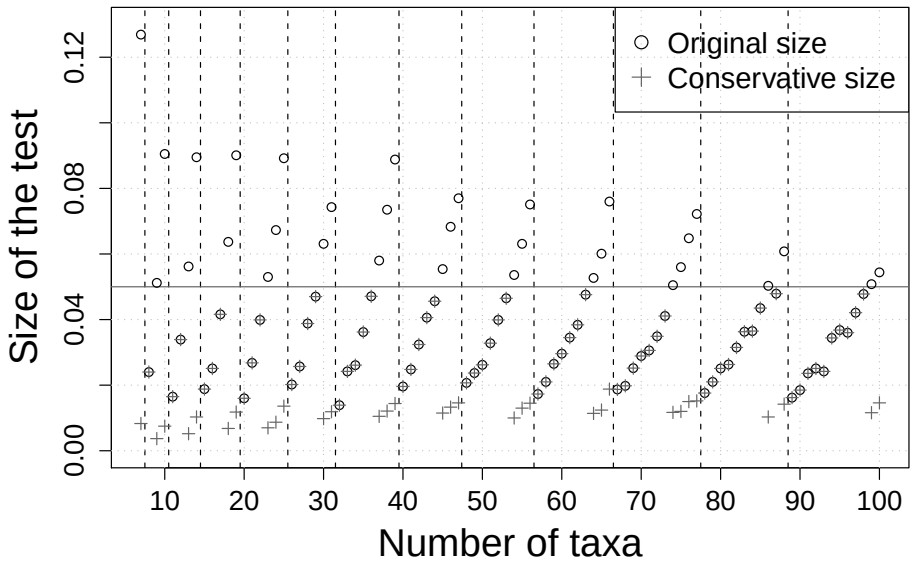
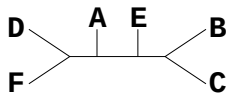


Figure 2.9: Size of the test: Evaluation of the test statistic for 10,000 random pairs of trees for each number of taxa. "Original size" is obtained by using the critical values of the MAST test (Figure 2.7). "Conservative size" is obtained by using the largest critical value which yields a size of 5 % or smaller. The vertical lines are the same as in Figure 2.7. The horizontal line displays the significance level 0.05.

Within these areas, all n have the same critical value (cf. Figure 2.7). For instance with $40 \leq n \leq 47$, the critical value is 12, thus a maximum of 28 ($n = 40$) to 35 ($n = 47$) taxa can be pruned while the pairs of trees are still significant. While the critical value remains constant between two lines, the number of taxa allowed to be pruned increases, thus more pairs show significance.

TreeBASE Data Set

To evaluate the behavior of the MAST test in a more realistic setting, we used real trees but random taxon mappings. To this end, we downloaded all 5023 trees from TreeBASE (`http://www.treebase.org/treebase/data/Tree.txt`, April 2008). Thereof, we investigated the 4610 trees comprising between 7 and 100 taxa. Unfortunately, the number of available trees varies strongly for the numbers of taxa. Especially for each $n \geq 94$ there are less than 10 trees available, but for each $n \leq 50$ there are more than 30 trees present. Two different trees with the same number of taxa are drawn randomly. If a tree contains multifurcations, these are randomly resolved each time



the tree is drawn, where each resolution is equally likely. The resulting bifurcating trees are relabeled randomly with the same taxon set.

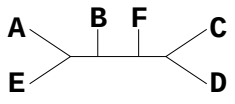
In Figure 2.10a, the fraction of significant results for the MAST test is shown for each number of taxa. Thereby results are considered significant if $p < 0.05$. On average, the MAST test is slightly too liberal with an average fraction of significant results of 0.064 and 0.058 for all n and for $7 \leq n \leq 50$, respectively.

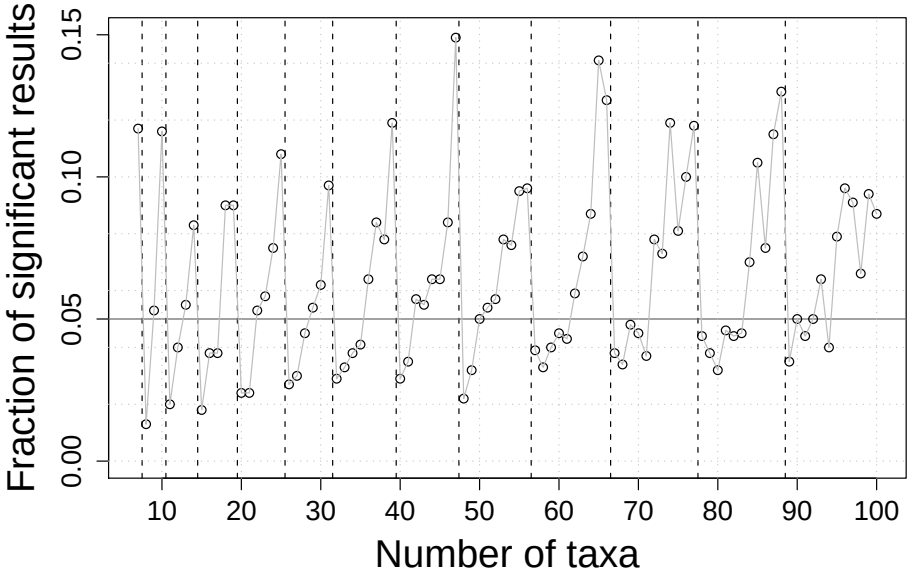
The modified MAST test uses the conservative critical value as explained in the previous section. Applying this test to the TreeBASE data set results in an average fraction of significant results of and 0.038 and 0.027 for all n and for $7 \leq n \leq 50$, respectively. The modified test is still liberal for some large values of n on this data set (Figure 2.10b).

The higher fraction of significant results in Figure 2.10 compared to Figure 2.9 may be due to the fact that the assumption of the null hypothesis of equally likely trees is not true (see e. g. Blum and François, 2006, for a study about tree shape distributions on a similar data set). Note that we weakened this fact by randomly resolving the multifurcations in the trees. We observe that the fraction of significant results depends strongly on the number of taxa, as already observed for random trees.

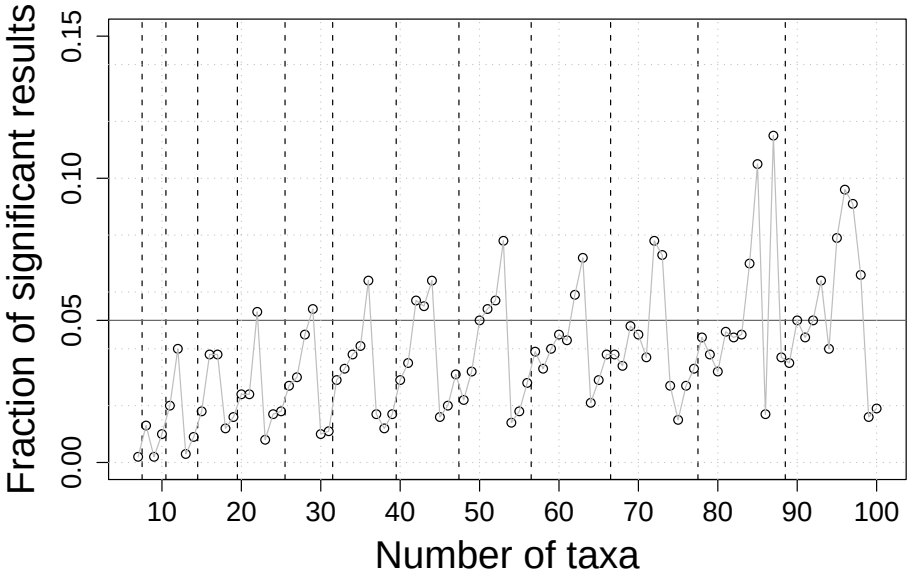
2.3.5 Conclusions

We have shown that a number of pitfalls exist when using the MAST test introduced by de Vienne *et al.* (2007) to test whether two phylogenetic trees are congruent. First, by using the MAST size as the basis of the test statistic, the positions of the taxa pruned from the trees are completely ignored and any positional information e. g. whether they were in the same subtrees is discarded. When applying the test in a biological framework, the taxa in the maximum agreement subtree should be regarded, not only their number. Second, a high number of taxa can be pruned from the phylogenies while the pair remains significant. Our third and major concern is that tree topologies are discrete as is the MAST size of two trees. One pitfall of the discreteness of the MAST size is the strongly varying size of the test for different numbers of taxa. The MAST test is too liberal for quite some values of n. Therefore, we recommend to adjust the critical value such that the test is conservative for all n. Finally, the test is more liberal using random phylogenies from TreeBASE which indicates that the assumption of equally likely trees may not be an appropriate null model.



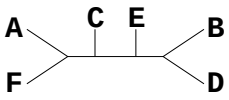


(a) MAST test



(b) modified MAST test

Figure 2.10: Results with trees from TreeBASE (1000 repetitions for each number of taxa). The vertical lines are the same as in Figure 2.7. The horizontal line displays the significance level 0.05.



Chapter 3

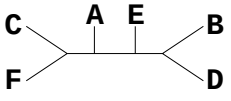
Combining Phylogenetic Trees

3.1 Introduction into the Combination of Overlapping Gene Data Sets

Combination at Different Levels

This chapter deals with combining trees, or more generally, combining phylogenetic data sets. A phylogenetic data set can be, for example, a gene alignment or morphological characters. As stated in Section 1.2.2, we focus on gene alignment data.

Different methods are available to combine the original data at different points along the way from the underlying sequences to the final tree (Schmidt, 2003; Ebersberger *et al.*, 2006): First, superalignment methods combine the data at an *early level* by directly concatenating the gene alignments without any intermediate computations (early-level combination; also called "supermatrix", "concatenation" or "total evidence", Kluge, 1989; de Queiroz and Gatesy, 2007). Superalignment methods have been used to infer phylogenies for eukaryotes (Philippe *et al.*, 2004), Metazoa and green plants (Driskell *et al.*, 2004), legumes (McMahon and Sanderson, 2006) or species from all three domains of life (Ciccarelli *et al.*, 2006). Second, *medium level* combination methods first compute intermediate results from the gene alignments, e. g. pairwise distances (Lapointe and Cucumel, 1997; Criscuolo *et al.*, 2006) or quartets (Schmidt, 2003), and subsequently reconstruct a phylogeny by combining this information. Third, consensus and supertree methods combine the data at the *late level* of gene trees (late-level combination; e. g. Bininda-Emonds, 2004). Supertree methods are used especially if only published trees but not the original data are available,

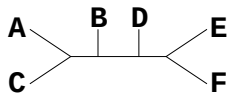


or if data of different kinds are combined. The prevalent method for reconstructing supertrees is matrix representation with parsimony (MRP, Baum, 1992; Ragan, 1992), MRP has been applied to many different kinds of species data, for instance to mammals (Bininda-Emonds *et al.*, 2007) or bacteria (Daubin *et al.*, 2002).

Each of these approaches has general advantages and disadvantages. Superalignment methods use all character information but assume the same underlying topology and often the same parameters for all genes. These model violations may lead to biased phylogeny estimates (Section 1.2.1). Consensus and supertree approaches account for differing topologies and parameters between genes. On the other hand, they are more susceptible to stochastic errors since estimating substitution parameters and a topology for each gene independently may lead to overfitting. Furthermore, they try to minimize the amount of missing data when constructing the gene trees. Medium-level approaches also allow for gene-specific parameters, but they use quartet likelihoods or distances, not gene trees, when building the final tree.

Combination in the Consensus Setting

The differences between concatenated alignments and tree combination have been extensively discussed in the consensus setting, i. e. where all data sets contain the same taxa (e. g. Barrett *et al.*, 1991; Bull *et al.*, 1993; de Queiroz *et al.*, 1995; Page, 1996; Gadagkar *et al.*, 2005). Consensus methods have the following advantages: They can be easily applied when data of different kinds are combined since the trees are reconstructed individually. Furthermore, it is possible to test these reconstructed trees for incongruencies before their combination, thereby the combination of heterogeneous data is prevented (Bull *et al.*, 1993). It has been observed that the certainty for a particular phylogeny increases with the data size and therefore, concatenation approaches may be misleading (Bull *et al.*, 1993). On the other hand, also the support for the true phylogeny may increase with more data and trees suboptimal for single data sets may emerge when the data is combined into one superalignment (Barrett *et al.*, 1991). In addition, consensus trees are usually less resolved. Therefore they are more conservative, but superalignment trees may have greater explanatory power (de Queiroz *et al.*, 1995).



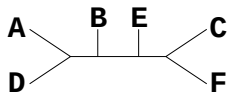
Properties of Supertree Methods

Supertree methods can be viewed as an extension of consensus methods as suggested by their first reference as "consensus supertrees" (Gordon, 1986). The so-called source trees can have overlapping taxon sets, but need not have identical ones. They are first computed for each gene, or are obtained from the literature, and are subsequently combined into a supertree. Supertree methods can be broadly characterized by the way they handle conflicting data. Supertrees generated by a *veto method* do not contain clades that any source tree would vote against (e. g. Goloboff and Pol, 2002; Ranwez *et al.*, 2007). Most supertree methods, including the ones we consider here, fall in the category of *liberal methods*. These methods resolve conflict by a voting procedure. This feature is intended since some amount of conflict is expected to be present among the gene trees (see also Section 1.2.1). For these liberal supertree methods, there are several desirable properties (Wilkinson *et al.*, 2004). E. g. to acquire accurate and unbiased results, the methods should not be biased by the number of taxa in the input trees, should be independent of the order of input trees, and should not prefer particular tree shapes. Some desirable properties, however, cannot be satisfied in conjunction (Steel *et al.*, 2000). Investigating different supertree methods in the consensus setting also helps to understand the properties of the respective methods since a property that does not hold in the consensus setting does not hold in general (Wilkinson *et al.*, 2007).

Furthermore, there is an inherent difference between supertree construction in the rooted and unrooted setting. For rooted trees, there is a polynomial-time algorithm to find a parent tree if it exists, i. e. a tree that contains all the relationships also present in the input trees (Aho *et al.*, 1981). On the other hand, to determine whether a parent tree for a set of unrooted trees with only overlapping taxon sets exists, is NP-complete (Steel, 1992). As mentioned in our definition of phylogenetic trees (Section 1.3.1), we are mainly interested in unrooted trees. However, in the simulation study (Section 3.3), we will also consider methods for rooted trees.

Practical Investigations of Supertree Methods

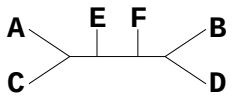
In addition to theoretical properties, practical investigations using real data sets or simulated data are of interest to compare different methods. Various authors used real data sets to compare superalignment and supertree approaches (Salamin *et al.*,



2002; Gatesy *et al.*, 2004; Fitzpatrick *et al.*, 2006; Dutilh *et al.*, 2007; Baker *et al.*, 2009). These real data sets have the advantage of a realistic setting, however, the true tree is usually unknown. Then only well-established clades can be used for assessing the performance (e. g. Dutilh *et al.*, 2007) or methods can be compared with one another (e. g. Baker *et al.*, 2009). In contrast, with simulations the results are compared to a model tree, and thus their performance can be measured at an absolute scale. Several studies investigating supertree methods using simulations were carried out (Bininda-Emonds and Sanderson, 2001; Bininda-Emonds, 2003; Eulenstein *et al.*, 2004; Levasseur and Lapointe, 2006). They employed the following general scheme: (1) Generation of a model tree assuming a Yule process, (2) generation of alignments along that tree, (3) random deletion of a proportion of taxa, (4) reconstruction of gene trees by MP, (5) construction of the supertree from the inferred gene trees, and (6) compare the supertree to the model tree. Bininda-Emonds and Sanderson (2001) compared superalignment and MRP for different degrees of divergence and observed that, with increasing divergence, the distance of the MRP trees to the superalignment tree increased. Levasseur and Lapointe (2006) compared average consensus, superalignment with distances and MRP for gene trees with complete taxon sets. They found average consensus to perform nearly as well as superalignment, whereas MRP was substantially worse since it ignores gene tree branch lengths.

Simulations can also be used to evaluate the impact of undesired properties for a particular supertree method. For instance, one of these properties is the emergence of "novel clades", i. e. clades contradicted by all gene trees. Bininda-Emonds (2003) found such clades to be very rare. However, note that due to missing taxa and multifurcating trees, it is not straightforward to measure supporting and conflicting relationships between a supertree and the gene trees (an alternative definition is presented by Wilkinson *et al.*, 2005b).

Each of the above simulation studies focused on a special subset of methods for supertree construction. A general performance assessment, however, has not yet been carried out, and the strengths and weaknesses of the different methods are unknown. After presenting data combination methods in detail (Section 3.2), we present an extensive simulation study about combining gene alignments (Section 3.3). There, we compare different data combination methods, including common supertree, superalignment and medium-level methods, to assess their accuracy in biologically reasonable situations.



After this practical evaluation, we investigate theoretical properties of existent and suggested supertree methods. Thereby we focus on majority-rule supertrees (Section 3.4) and the null models underlying supertree reconstruction (Section 3.5).

3.2 Methods for Combining Overlapping Gene Data Sets

First, we present some early-, medium-, and late-level methods together with the implementation used for the simulation study (Section 3.3). All methods investigated here, together with the abbreviations used, are listed in Table 3.1.

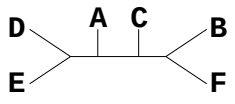
3.2.1 Early-level Combination

A superalignment is generated from single gene alignments by concatenating the different alignments and adding gaps where no sequence information is present for a specific taxon. The superalignment method (**SA**) refers to reconstructing the superalignment tree. Here, we use maximum likelihood (ML) or maximum parsimony (MP), depending on the size of the data set. ML phylogenies are computed with IQPNNI version 3.1 (Vinh and von Haeseler, 2004), assuming the substitution model HKY for DNA sequences (Hasegawa *et al.*, 1985) and JTT for protein sequences (Jones *et al.*, 1992). In both cases, site heterogeneity is modeled with four Γ-distributed rate categories. MP phylogenies are computed with PAUP* 4.0b10 (Swofford, 2002) and the following parameters: heuristic search with TBR branch swapping, random addition of sequences, and a maximum of 10,000 trees in memory.

3.2.2 Late-level Combination

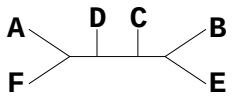
Phylogenetic Reconstruction of Gene Trees

The first step of any late-level combination method is the reconstruction of the gene phylogenies (see also Figure 3.1, page 47), which serve as source trees for the supertree reconstruction. We compute ML gene trees with IQPNNI using the same reconstruction parameters as for the early-level combination. In some simulations,



Abbreviation	Description	Reference
Late-level combination:		
Consensus	Majority-rule consensus	Margush and McMorris (1981)
MRP_BR	Matrix representation with parsimony and Baum/Ragan coding	Baum (1992); Ragan (1992)
MRP_PU	Matrix representation with parsimony and Purvis coding	Purvis (1995)
MRP_I	Matrix representation with irreversible parsimony and Baum/Ragan coding	Camin and Sokal (1965)
MRF_BR	Matrix representation with flipping and Baum/Ragan coding	Chen *et al.* (2003); Burleigh *et al.* (2004)
MRF_PU	Matrix representation with flipping and Purvis coding	-
MRC	Matrix representation with compatibility and Baum/Ragan coding	Rodrigo (1996); Ross and Rodrigo (2004)
MinCut	Minimal cut	Semple and Steel (2000)
ModMinCut	Modified minimal cut	Page (2002)
MaxCut	Maximal cut	Snir and Rao (2006)
QILI	Quartet inference and local inconsistency	Piaggio-Talice *et al.* (2004)
Medium-level combination:		
SuperQP	Super quartet puzzling	Schmidt (2003)
AvCon	Average consensus	Lapointe and Cucumel (1997); Lapointe and Levasseur (2004)
SDM	Super distance matrix	Criscuolo *et al.* (2006)
Early-level combination:		
SA	Superalignment	e. g. Kluge (1989)

Table 3.1: Overview of reconstruction methods compared in the simulation study (Section 3.3) and corresponding abbreviations.



the gene trees are obtained via bootstrapping. In this case, we generate 100 bootstrap replicates of each gene alignment with `seqboot`, i. e. we build alignments of the same length from the original alignment, where columns are drawn with replacement. From these alignments, phylogenies are computed with IQPNNI and subsequently a majority-rule consensus tree (see also next section) is built from the bootstrapped trees with `consense`. Both `seqboot` and `consense` are part of PHYLIP version 3.6 (Felsenstein, 2005).

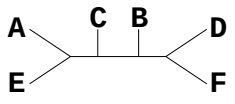
Consensus

Consensus methods combine trees on complete data, i.e. trees on the same taxon set (for a review see Bryant, 2003). For complete data, we compute the majority-rule consensus of the gene trees. The majority-rule consensus tree contains all splits present in more than half of the input trees. We use the implementation in `consense`.

Methods using Matrix Representation

Three methods based on matrix representation (MRep) coding schemes are available: MRep with parsimony (MRP), MRep with flipping (MRF), and MRep with compatibility (MRC). All three aim to optimize an objective function as explained later. If more than one optimal tree is found, we take the strict consensus tree of the optimal trees as the reconstructed tree.

Different coding schemes have been suggested to decompose the gene trees into an MRep: In the Baum-Ragan (**BR**) coding scheme, every gene tree topology is coded as follows (Baum, 1992; Ragan, 1992; Baum and Ragan, 2004): A split in a tree divides the taxa into two disjoint sets (see also Section 1.3.2). For each split, a column is added to the MRep, where '`0`' and '`1`' indicate the taxa on either side of the split and missing taxa are coded as '`?`'. For rooted trees, the root-side is always coded as '`0`'. The Purvis (**PU**) coding scheme can only be applied to rooted trees. Then, sister groups are coded binarily, and the remaining taxa are coded as '`?`' (see Table 3.2 for an example). This aims at removing some redundant information (Purvis, 1995). We generate both matrix representations from the list of gene trees using `r8s` version 1.71 (Sanderson, 2003). We see that MRep-Methods are split-based methods since they first extract the split information of the input trees, code them in the MRep, and subsequently compute a supertree out of this split information.



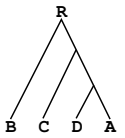
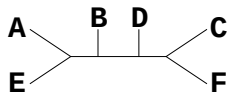
Tree	Baum/Ragan coding	Purvis coding
		R 00
	A 11	A 11
	B 00	B 0?
	C 10	C 10
	D 11	D 11

Table 3.2: Example of coding a gene tree as a matrix representation. The Baum/Ragan coding codes every split independently. We use the unrooted version of the BR coding, i. e. without coding the root explicitly. The Purvis coding codes only sister groups of rooted trees.

MRP trees are reconstructed by searching the most parsimonious tree for the matrix representation (Baum, 1992; Ragan, 1992; Baum and Ragan, 2004). We apply two kinds of parsimony: (1) reversible Fitch parsimony (Fitch, 1971), which assumes the character changes to be undirected, and (2) irreversible Camin-Sokal parsimony, which only allows changes from `0` to `1` and thus uses the root information in the trees (Camin and Sokal, 1965). The most parsimonious tree with the respective criterion is determined by PAUP* 4.0b10 (heuristic search with TBR branch swapping and random addition of sequences, and a maximum of 10,000 trees in memory). Overall, we consider three MRP variants in the simulation setting: MRP_BR (reversible parsimony and BR coding), MRP_I (irreversible parsimony and BR coding) and MRP_PU (reversible parsimony and PU coding).

The objective function of **MRF** is to minimize the number of binary flips (changes from '`0`' to '`1`' and vice versa) necessary to convert the original MRep into an MRep compatible with a tree (Chen *et al.*, 2003; Burleigh *et al.*, 2004). Here, we apply MRF to both coding schemes, BR and PU. To date, MRF has only been applied to matrices with Baum/Ragan-coding. Since MRF, like MRP, is an NP-complete problem, we use the heuristic implemented in `HeuristicMRF2` (`http://genome.cs.iastate.edu/CBL/`; Chen *et al.*, 2006).

With **MRC** the optimizing function is maximizing the number of compatible columns, i. e. the number of splits that can be arranged in a tree without conflict (Rodrigo, 1996; Ross and Rodrigo, 2004). As a heuristic, we use Clann version 3.0.2 to find the MRC tree for a BR coded matrix representation (the `sfit` criterion with default parameters; Creevey and McInerney, 2005).



Variants of the "Build" Algorithm

The "Build" algorithm (Aho *et al.*, 1981) is only able to construct a supertree for a set of compatible and rooted gene trees. In case of compatible gene trees, each gene tree is a subtree of the supertree. "Build" and its variants are graph-based rooted triplet methods, thus, rooted trees are required. To combine incompatible gene trees, different **cut methods** have been developed.

MinCut (minimal cut) is an extension of the "Build" algorithm (Semple and Steel, 2000). In case of a conflict, MinCut introduces an edge in the supertree that conflicts with the fewest possible number of triplets.

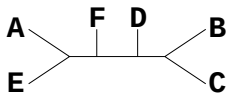
ModMinCut (modified MinCut) improves MinCut by not only considering the contradicting triplets for an edge but, additionally, by trying to keep subtrees that are uncontradicted by the gene trees (Page, 2002).

Both MinCut and the ModMinCut are polynomial-time algorithms implemented in `supertree` by Rod Page. We use a precompiled version of this program taken from Rainbow 1.2 beta (Chen *et al.*, 2004).

MaxCut (Snir and Rao, 2006) considers two types of triplet topologies: bad ones which occur in a gene tree, and good ones for which another possible topology occurs in a gene tree. In case of a conflict, the ratio of these counts is maximized, which is an NP-hard problem. Snir and Rao (2006) suggested a heuristic based on semidefinite programming. We compute the MaxCut tree from a set of triplets with a program provided by Sagi Snir (personal communication). To apply it, we first extract triples from the gene trees using a program provided by Gregory Ewing.

Quartet-based Methods

QILI (Quartet Inference and Local Inconsistency; Piaggio-Talice *et al.*, 2004) is based on quartet topologies extracted from unrooted gene trees. First, a set of weighted quartets is computed, where the weights for each quartet are smaller if they occur in more trees. Missing quartets are inferred by a rectifying process using quintet information. From this collection of quartets, a tree is estimated by minimizing the weighted sum of the quartets represented in a tree using Willson's local inconsistency method (Willson, 1999). QILI is available in the `QuartetSuite` 1.0 package.



3.2.3 Medium-level Combination

Quartet-based Methods

SuperQP combines the sequence data based on the quartet likelihoods (Schmidt, 2003). For each gene, TREE-PUZZLE (Schmidt *et al.*, 2002) computes all quartet tree likelihoods. These likelihoods are combined for every possible quartet topology across all genes containing the respective quartet. The likelihoods are used to combine the data into so-called superquartets, the building blocks for SuperQuartetPuzzling (SuperQP). SuperQP is related to the QP algorithm (Strimmer and von Haeseler, 1996), but it takes also missing data into account, using an overlap-graph guided insertion scheme and a voting procedure that is aware of missing quartets. We compute the SuperQP tree with an upcoming version of the TREE-PUZZLE package.

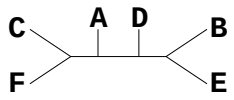
Distance-based Methods

The medium-level information for distance-based methods are pairwise distance matrices computed separately for each gene. Here, we estimate pairwise ML distances with IQPNNI and the same models as for the early-level combination. The distances are combined into one distance matrix for all taxa, which is subsequently fitted to a tree with the least-squares method of Fitch-Margoliash (Fitch and Margoliash, 1967). We use the `fitch` implementation in the PHYLIP package with the `Subreplicates` option, thus allowing for missing data by considering only available entries. Two distance-based medium-level methods, differing only in the combination of the matrices, have been devised so far:

With average consensus (**AvCon**) each entry of the combined distance matrix is computed by averaging over all distances available for the corresponding pair of taxa (Lapointe and Cucumel, 1997; Lapointe and Levasseur, 2004).

Super Distance Matrix (**SDM**; Criscuolo *et al.*, 2006) inserts two types of parameters: (1) weighting factors for each distance matrix, which correspond to a branch lengths scaling for each gene tree, and (2) additive constants for each taxon in each matrix, which correspond to an elongation of terminal branches. Utilizing several contraints, the variance of the scaled and shifted gene distance matrices to the combined distance matrix is minimized.

Both methods are implemented in the SDM program (Criscuolo *et al.*, 2006).



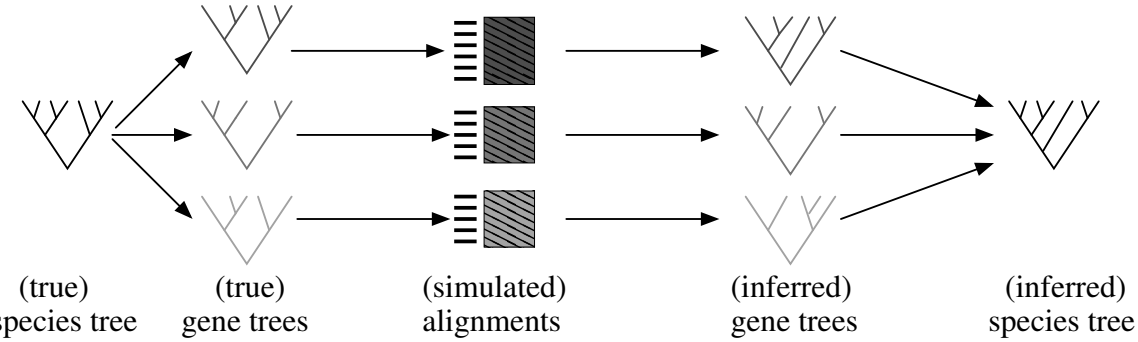


Figure 3.1: Diagram of the simulation setting with supertree reconstruction. The simulation proceeds in several steps: First, gene trees are generated from the given species tree. Alignments are simulated along these gene trees. From these alignments gene trees are inferred. The inferred gene trees are the source trees for supertree reconstruction. With the supertree methods species trees are inferred.

3.3 Simulation Study

The results in this section were first presented in Kupczok *et al.* (2009).

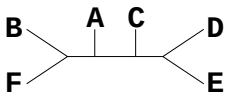
3.3.1 Simulation Setting

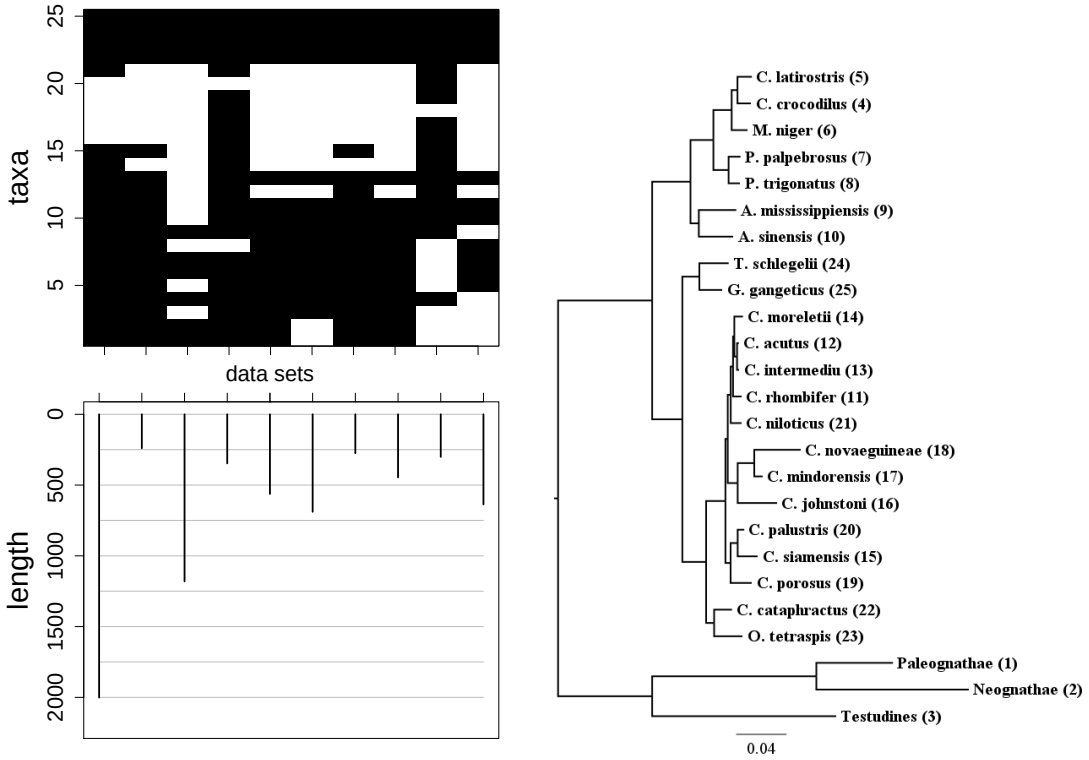
Parameters

Figure 3.1 gives an overview of the simulation setting and notations. We study different parameters involving the underlying data set, the coverage of the sequence data, the topology and parameters of the true gene trees and the sequence lengths (Table 3.3, page 51). The last three parameters will be described in detail along with the results.

Like e. g. Salamin *et al.* (2005) and Gadagkar *et al.* (2005), we simulate according to biologically reasonable assumptions by taking simulation parameters from real data. We use two data sets:

The small data set is given by the parameters of the crocodile data of Gatesy *et al.* (2004). This data consists of 10 DNA alignments, morphological traits, two RFLP matrices, two allozyme data sets, chromosomal morphology and nest type information for a total of 86 recent and extinct crocodile taxa. Here, we only use the DNA data, which reduces the taxon set to 25 recent taxa and a superalignment of 6,681 sites





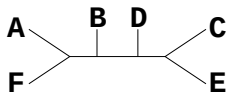
(a) Distribution of taxa and gene lengths in the 10 data sets

(b) Superalignment ML tree, numbers in brackets refer to taxa numbers in (a)

Figure 3.2: Small data set (Crocodile data). On average, 65.2 % of the genes are present in a taxon.

(Figure 3.2a). Our reconstruction of two superalignment ML trees, one with HKY+Γ and one with GTR+Γ, results in the same tree topology but different branch lengths (HKY tree in Figure 3.2b). This topology is more resolved than the one by Gatesy *et al.* (2004), and in addition, there is one resolution conflicting with the superalignment tree computed by Gatesy *et al.* (2004): in our analysis, *C. palustris* and *C. siamensis* form a clade instead of *C. porosus* and *C. palustris*. Note, that there are two main differences between our superalignment analysis and the one of Gatesy *et al.* (2004). First we only use the subset of the data that consists of DNA data, and second we use maximum likelihood, not maximum parsimony, as the tree reconstruction method.

We use the maximum likelihood HKY tree (Figure 3.2b) as the species tree for subsequent simulations. For methods requiring rooted gene trees, we root each tree artificially with a taxon in which all genes are present (*O. tetraspis*, taxon 23). Such a procedure was suggested by Baum (1992). Thus the small data set contains of 25



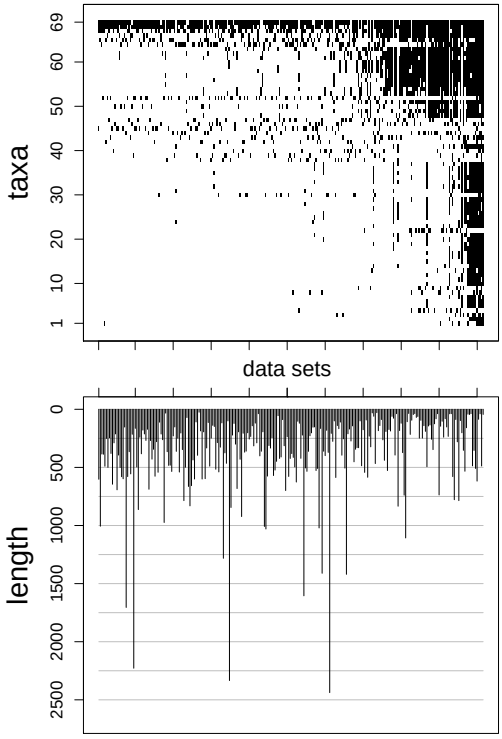
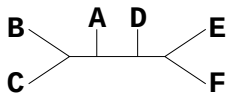


Figure 3.3: Large data set (green plant data): Distribution of taxa and gene length in the 254 data sets. On average, 15.8 % of the genes are present in a taxon.

taxa and 10 genes having different sequence lengths and taxa occurences (Figure 3.2a). Furthermore, the species tree shows a highly non-uniform branch length distribution (Figure 3.2b). These features are typical for real data sets.

The large data set is composed of 254 proteins from 69 green plants with an overall length of 96,698 amino acids (Driskell *et al.*, 2004). Driskell *et al.* (2004) describe this data set as problematic, since their reconstructed tree shows relations not supported by any gene tree and the numbers of supporting genes seem to be barely correlated with the bootstrap support for clades. The data contain a higher fraction of missing data compared to the small data set (Figure 3.3). As species tree we use the superalignment ML tree of the original data, reconstructed with the JTT substitution matrix. Since the data contain no taxon for which all genes are available, every reconstructed gene tree is rooted at the split that best matches the rooting of the model tree. Thereby the model tree is rooted with the taxon suggested by Driskell *et al.* (2004).



Sequence Simulation

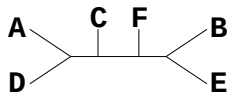
Sequences can be simulated along a given tree with branch lengths by Monte-Carlo simulations (Rambaut and Grassly, 1997). Thereby, random sequences are generated that evolved under the tree and given model parameters. For most simulations, we take the superalignment ML tree for the real data to be the true species tree. Estimated nucleotide and amino acid frequencies, as well as the α-parameter of the Γ-distribution, are used as parameters for the sequence simulations with `seq-gen` (Rambaut and Grassly, 1997). Unless stated otherwise, protein data are generated with JTT and nucleotide data with an HKY model with the transition/transversion ratio taken from the original ML estimation. Sequences are simulated with the same lengths distribution as in the original data.

There is also the possibility to use the gene trees from the original data as the true gene trees (true gene trees gene-specific, G in Table 3.3). In this case there is no true species tree known.

For each simulated data set, at most fifteen different methods are applied to reconstruct a tree (Table 3.1, page 42). Note that not all methods are applicable for all settings. Consensus is only applicable for complete data and the medium- and low-level methods are only applicable if sequence information is present.

Tree Distance Computation

If applicable, we measured the accuracy of the methods by the normalized Robinson-Foulds distance (RF, Section 2.1.3) of the inferred species tree to the true species tree. As explained before, the Robinson-Foulds distance is the number of splits that are present in one tree but not in the other one, and vice versa. Since unrooted n-taxa trees have a maximum of $n-3$ inner branches, the maximal Robinson-Foulds distance is $2(n-3)$. In this section, *RF* denotes the *normalized* Robinson-Foulds distance, where the distances are divided by $2(n-3)$. This yields a value between 0 % and 100 %, which can be interpreted as the percentage of true or missing splits in the inferred tree compared to the true tree.



Parameter	Options
Data set	**S**: small **L**: large
Taxa coverage	**c**: complete **m**: missing
True gene trees	**E**: sub-trees of species tree $\mathbf{R}_\alpha$: rate of evolution assigned randomly from a Γ-distribution with parameter α (i. e. mean 1 and variance $1/\alpha$) **P**: substitution parameters and branch lengths gene-specific **G**: trees gene-specific $\mathbf{T}_\theta$: trees random by coalescent process with parameter θ
Reconstructed gene trees	**e**: equal to true gene trees **n**: normal sequence length and ML estimation **l**: long sequence length (ten times longer) and ML estimation

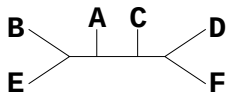
Table 3.3: Parameters varied in the simulations. The setting in each simulation is abbreviated by one of the bold letters given in each of the four categories. Note that not all combinations were tested.

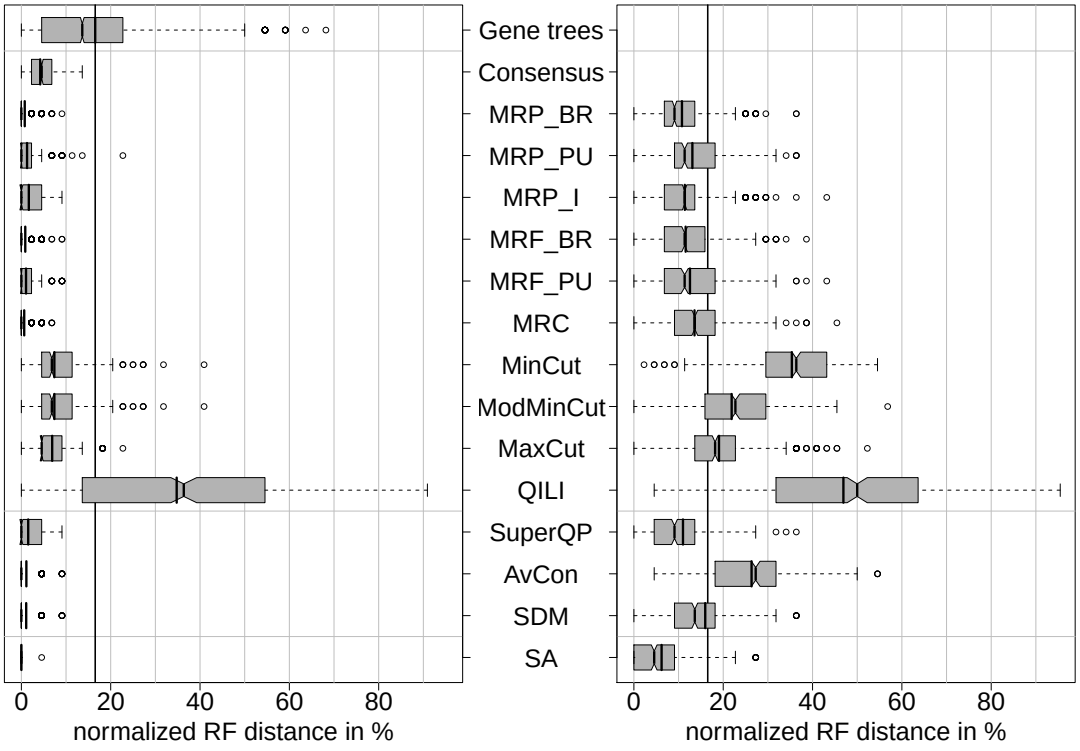
3.3.2 Simulation Results

Each simulation setting is abbreviated by four letters corresponding to values for each of the four categories of simulation parameters (Table 3.3).

Complete Data (S,c,E,n)

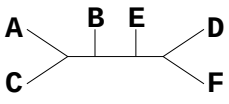
The first and simplest simulation is that the topology and parameters of the species tree equal those of the true gene trees and the length of each gene alignment is taken from the original data set. In 500 replications, SA nearly always reconstructs the true tree, i. e. $RF = 0$ (Figure 3.4a). The MRep-methods and the intermediate methods show mean RF distances of less than 2 %. In contrast, the mean distance of an inferred single gene tree to the true species tree is 16.5 %. This value can be viewed as the mean distance when reconstruction is based on the sequence information of one gene only. Therefore we will call it the *baseline distance*. Surprisingly, QILI shows a mean RF distance of 35 %, which is much larger than 16.5 %. Thus, accuracy is lost by combining gene trees with this method.





(a) S,c,E,n **(b)** S,m,E,n

Figure 3.4: Distribution of normalized *RF* distances (500 simulations) for the simulation settings S,c,E,n and S,m,E,n. The reconstructed trees were compared with the model tree via the *RF* distance. The distributions resulting from 500 repetitions are shown. The boxes mark the 1/4- and 3/4-quantiles, the vertical line with the notches is the median with the 95 % confidence interval for comparing two medians. The vertical line without the notches is the mean of the data. The vertical black line drawn throughout the diagrams is the mean *RF* distance of all complete gene trees, which serves as the baseline distance.



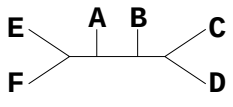
Missing Data (S,m,E,n)

Next, we use the same 500 simulated alignments as before, but delete sequences from the gene alignments according to Figure 3.2a. The resulting distributions of the *RF* distances show that all methods are strongly affected by missing data (Figure 3.4b). With a mean *RF* distance of about 6.2 %, SA is again the most accurate method. Among the remaining methods, MRP_BR (10.8 %) and SuperQP (11 %) show the smallest mean *RF* distances. The cut methods, QILI, and average consensus show mean *RF* distances larger than the baseline distance of 16.5 %. Thus, these methods perform on average worse on incomplete data sets than the ML reconstruction using only one gene present in all taxa. These methods seem to be unable to efficiently utilize the additional information provided by extra, but incomplete, gene data.

Large Data Set (L,m,E,n)

This simulation uses the data set of 254 genes from 69 green plant species. Compared to the small data set, the alignment of the large data set contains more taxa, more genes, but a smaller fraction of genes present per taxon (Figure 3.3). Here, we study the simplest simulation setting with missing data. Although SA trees are reconstructed with parsimony to keep computing time reasonable, they still show the highest accuracy with a mean *RF* distance of 4.8 % (Figure 3.5). Among the MRep-methods, MRP_I (12 %) is no longer as accurate as the other MRep-methods. MRF_BR (5.7 %) and MRF_PU (5.8 %) are the supertree methods with the highest accuracy. MinCut (93.9 %) reconstructs trees that are very distant to the true species tree. A possible reason is the high proportion of missing data. The accuracy of MinCut is improved by ModMinCut (54 %) and MaxCut (31.5 %), but all cut methods show larger distances than the average complete gene tree (the baseline distance, 18.5 %). QILI shows a much better performance compared to the small data set, its mean accuracy (20.4 %) is now comparable to SuperQP (16.1 %) and SDM (20.2 %). These methods show average distance values very close to the baseline distance. But QILI still has a high variance, whereas SuperQP shows good results in most cases and produces unresolved trees in a few cases.

In general, the results of the large data set are similar to those for the small data set: In both settings, the methods that improve the baseline distance are the same, superalignment outperforms the other methods, the MRep-methods are the best supertree methods, and SuperQP is the best medium-level method. Thus, we expect



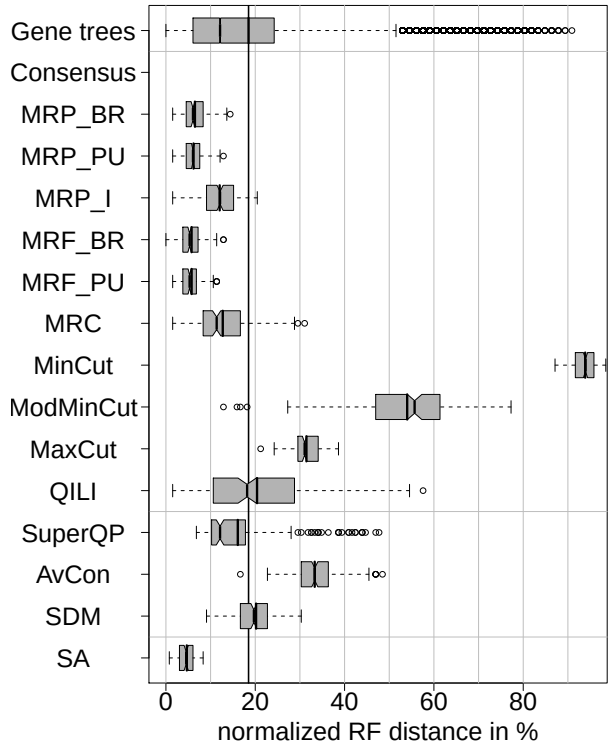
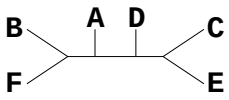


Figure 3.5: Distribution of normalized *RF* distances (200 simulations) for the simulation setting L,m,E,n (large data set with missing data according to Figure 3.3). "Gene Trees" shows the distances of the trees from the complete alignments, not from the pruned alignments, although the latter are used for the data combination methods.

the results also to be similar when introducing deviating settings. In the following, we only present the results for the small data set.

Long Sequences (S,m,E,I)

We also test whether the methods are able to combine highly informative, but incomplete, data sets. Thus, we minimize the effect of erroneous gene tree reconstruction by generating gene sequences ten times longer than the original gene sequences while taxa occurrences are the same as in Figure 3.2a. The accuracy of inferred species trees and gene trees is substantially improved compared to setting S,m,E,n for all methods (Figure A.1 in the appendix). High mean *RF* distances for QILI (30.3 %) and AvCon (8.1 %), however, show that these methods fail to reconstruct reasonable trees from highly informative data sets with missing data. The mean *RF* distances for MinCut, SuperQP and SDM are between 1 % and 2 % and all remaining methods show an



average RF distance of $\leq$1 %.

Bootstrapped Phylogenetic Trees

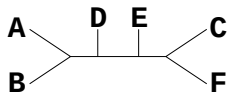
We extended the simulation with missing data (S,m,E,n) by bootstrapping the superalignment and the gene trees. In this case, reconstructed gene trees were the majority-rule consensus of trees reconstructed from bootstrapped alignments. Since branches with low support are discarded from each gene tree, the accuracy of supertree methods is expected to improve. Note that this bootstrap procedure does not affect the medium-level methods. Here, we measured the accuracy of reconstruction for 200 of the alignments that were the basis of the simulations summarized in Figure 3.4b (S,m,E,n). The bootstrapped gene trees lead to an improvement of the accuracy of all supertree methods (Figure A.2) when compared to the results without bootstrapping. The mean RF distance is now 5.6 % for superalignment, between 9 and 10.3 % for all MRep-methods, and between 12 and 22 % for the cut methods.

Gene-specific Evolutionary Rates (S,m,R_α,n)

Now we introduce a more complicated setting where the evolutionary rates vary between genes. The true gene trees were generated from the species tree by stretching or shrinking all branch lengths with a Γ-distributed random factor drawn independently for each gene in each simulation. In two different settings, the shape parameter for the Γ-distribution was $\alpha = 3$ and $\alpha = 1.67$, respectively. As in the previous simulations, the substitution parameters for the sequence simulation were equal for each gene. The gene trees and the SA tree were also obtained by bootstrapping. For each choice of α, we computed 100 simulated alignments. For neither setting do the results differ substantially from the previous simulation with bootstrapping (Figures A.3 and A.4).

Gene-specific Substitution Parameters (S,m,P,n)

Here, as in the previous setting, the true gene trees differ from the species tree by their branch lengths. However, this time the branch lengths were fitted from the original data to obtain the true gene trees. For each alignment, the species tree was pruned to the respective taxon set. Afterwards, GTR parameters and branch lengths were fitted to the pruned tree using the original alignment. If a branch length got down to 10^{-6}, the lower bound in IQPNNI, the respective branch length was set to $1/l$, where



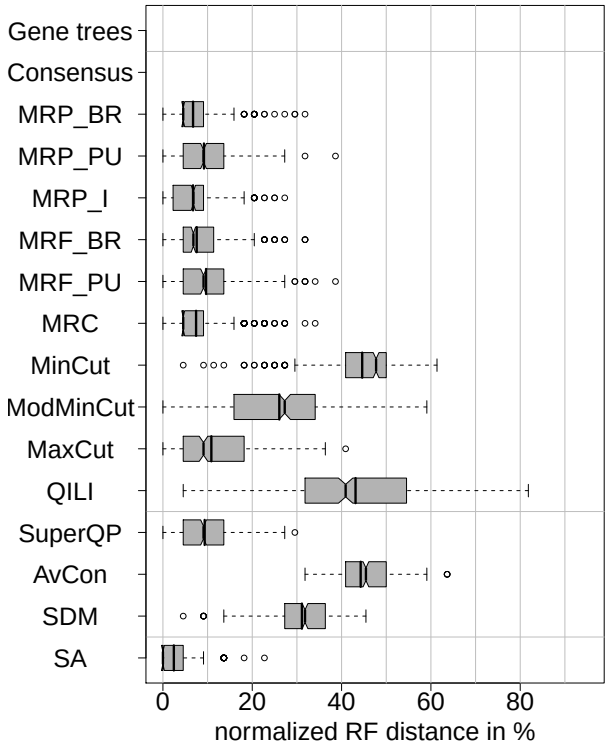
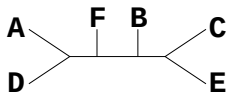


Figure 3.6: Distribution of normalized *RF* distances (500 simulations) for the simulation setting with gene-specific GTR parameters and missing data (S,m,P,n). The baseline distance is not applicable here (see text for details).

l is the length of the corresponding alignment. The trees constructed this way were used as the true gene trees for the simulations. The sequence simulations used the estimated GTR parameters for each gene.

This simulation setting only allows for simulation of pruned data sets. Thus, the baseline distance is not applicable. The results cannot be compared directly to the previous simulations, since the average tree length is now larger, but the ranking of the methods can be compared. Figure 3.6 shows that the superalignment trees remain best (mean *RF* distance of 2.4 %), even if simulation parameters differ between genes. SA, the MRep-methods, MaxCut and SuperQP are clearly better than the distance based methods, MinCut and ModMinCut.



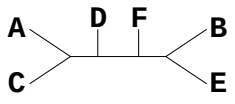
Gene-specific Topologies (S,m,G,n)

Here, the previous setting is extended as follows: Not only branch lengths and substitution parameters are gene-specific but also the topologies. Therefore, the gene trees reconstructed from the original data were used as true gene trees for this simulation. As before, only the setting with missing data can be studied, since the true gene trees already contain missing data. As we do not know the underlying species topology, a more complicated evaluation method is used: the inferred tree from each method is compared to the tree reconstructed from the true gene trees with the same method. E. g. an MRP_BR tree was reconstructed from the true gene trees and was used as a model tree when the distances to MRP_BR are evaluated in Figure 3.7. Also the early- and medium-level trees are reconstructed from the original sequence data and used for the distance computations. With this procedure, we estimate how consistently each method finds its own reconstructed species tree when sequence data is simulated along the gene trees. This is similar to a parametric bootstrap approach. Here, we face the problem that some trees reconstructed from the original data are not fully resolved. Also in these cases, we compute the Robinson-Foulds distances to these trees and normalize it with the same factor of $2(n-3)$, where n is the number of taxa. Thus, the polytomies in these trees are treated as true and the distance increases if a tree reconstructed in the simulation is more resolved. To highlight this problem, we list the number of branches missing in the trees reconstructed from the original data on the right margin of Figure 3.7.

The resulting distances clearly show that SA is the most consistent method, since it has the smallest average distance to the SA tree from the original data (7.8 %). It is followed by MRP_BR with a mean RF distance of 13.2 %.

Incomplete Lineage Sorting (S,c,$\mathbf{T}_\theta$,e and S,m,$\mathbf{T}_\theta$,e)

In this setting, the true gene trees were generated from the true model tree by a coalescent process (for details of the coalescent model used here, see Ewing *et al.*, 2008). This can result in different branch lengths, but also different topologies. The species tree was rooted according to Figure 3.2b. From this rooted species tree, we simulated gene trees with different coalescent parameters. The coalescent parameter θ was used to generate incongruent gene trees with different amounts of incorrect branches. The larger θ, the more incongruence is caused by incomplete lineage sorting.



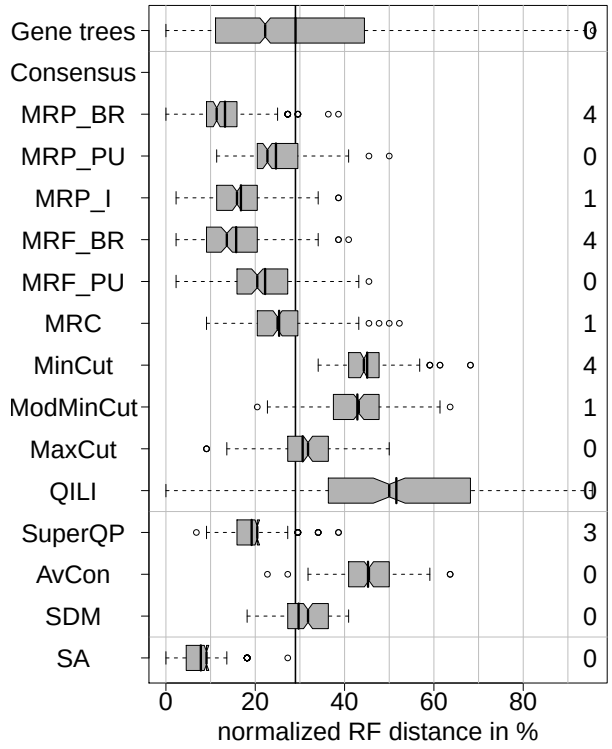
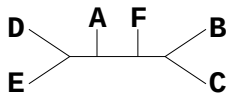
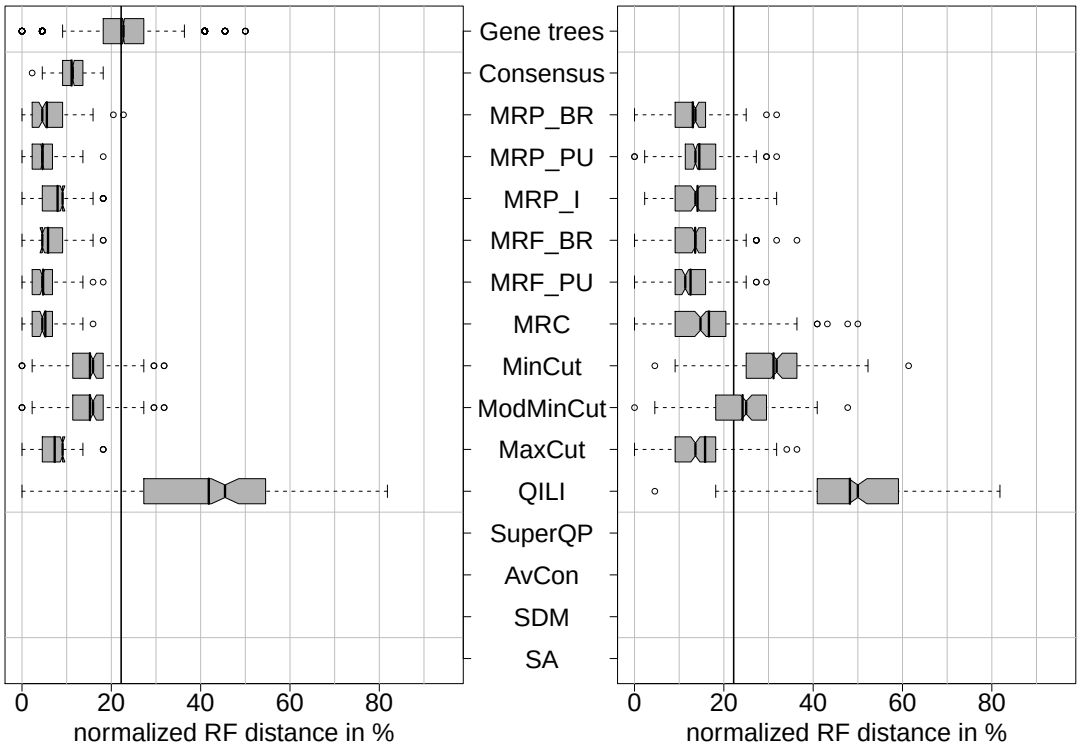


Figure 3.7: Distribution of normalized *RF* distances (200 simulations) for the simulation setting with gene-specific topologies and missing data (S,m,G,n). Note that the baseline distance is defined differently here: the gene tree distances are computed by comparing each reconstructed gene tree to the corresponding true gene tree and normalized with the appropriate number of taxa. The numbers on the right are the numbers of unresolved branches in the tree reconstructed from the original data with the corresponding method.

E. g. $\theta = 0.005$ results in a considerable incongruence among the gene trees: the mean normalized *RF* distance between the true species tree and the true gene trees is 22 % (Figure 3.8a).

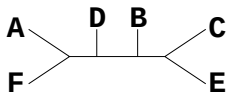
First, we investigate the performance of the supertree methods in the presence of incongruent gene trees without any reconstruction error. In Figure 3.8a, we see that the matrix representation methods can estimate the species tree quite accurately in the presence of complete data; MRP_PU and MRF_PU give the best results with a mean reconstruction error of 4.6 % and 4.7 %, respectively. The matrix representation methods, headed by MRF_PU (12.5 %), are also the best methods when data is missing (Figure 3.8b).

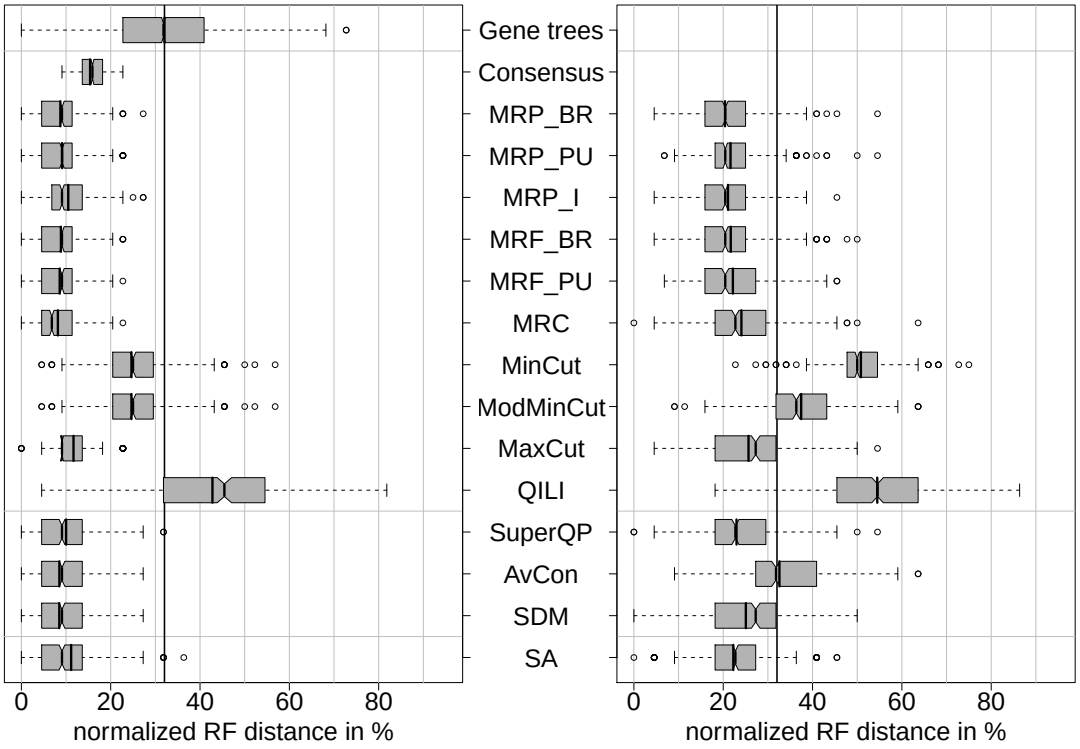




(a) S,c,$T_{0.005}$,e **(b)** S,m,$T_{0.005}$,e

Figure 3.8: Distribution of normalized *RF* distances (500 simulations) for the simulation setting with gene-specific trees generated by a coalescent process ($\theta = 0.005$) without reconstruction error. Early- and medium-level methods cannot be applied since no simulated sequences are available.



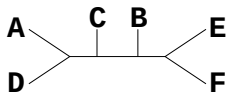


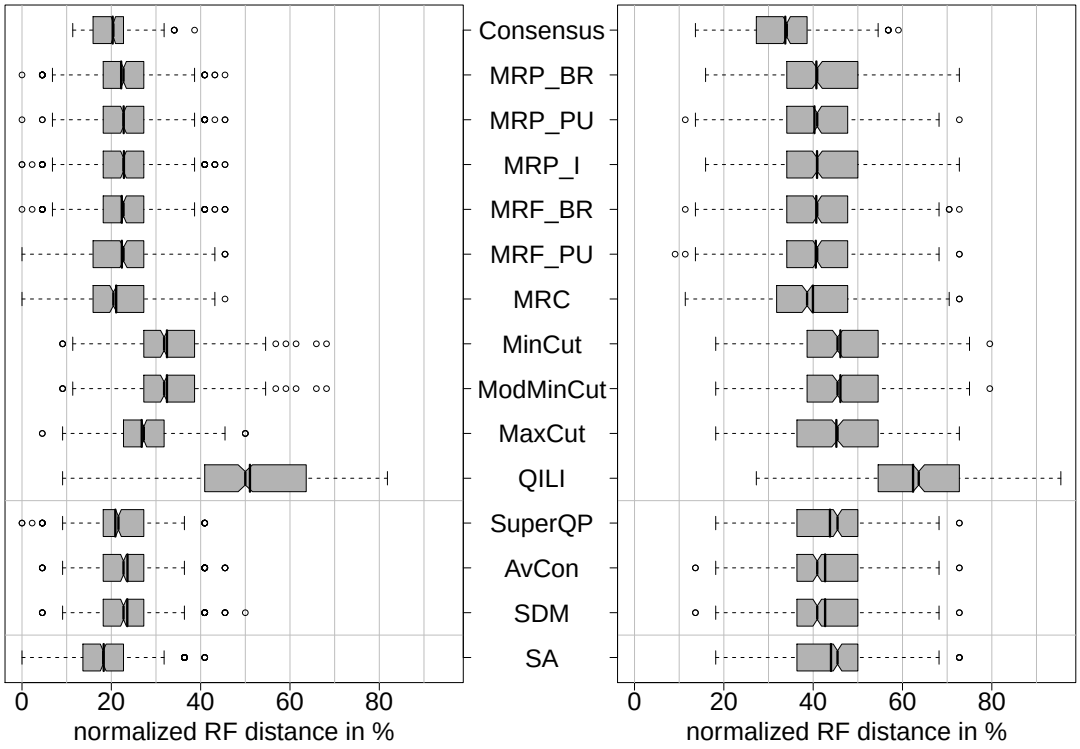
(a) S,c,$T_{0.005}$,n **(b)** S,m,$T_{0.005}$,n

Figure 3.9: Distribution of normalized *RF* distances (500 simulations) for the simulation setting with gene-specific trees generated by a coalescent process ($\theta = 0.005$). The true gene trees are equal to the trees used in Figure 3.8. Now, simulated alignments are obtained by simulating sequences along these true gene trees.

Incomplete Lineage Sorting and Gene Tree Reconstruction (S,c,T_θ,n and S,m,T_θ,n)

The gene trees from the previous section are taken as true gene trees. Along these, sequences are simulated and phylogenies are inferred as before. Thus, reconstruction error is added to the error present due to incomplete lineage sorting. The mean distance of the inferred gene trees to the species tree is 32 % (Figure 3.9a). In the case of complete data, this distance is decreased by all methods except QILI. The distributions and mean distances of MRP_BR (8.7 %), MRP_PU (9.1 %), MRP_I (10.5 %), MRF_BR (8.9 %), MRF_PU (8.6 %), MRC (8.2 %), MaxCut (11.7 %), SuperQP (10 %), AvCon (8.5 %), SDM (8.5 %) and SA (11.1 %) are very similar. Thus, the differences between the methods are less distinct. However, the mean superalignment



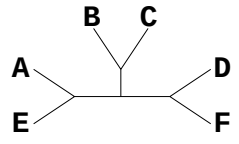


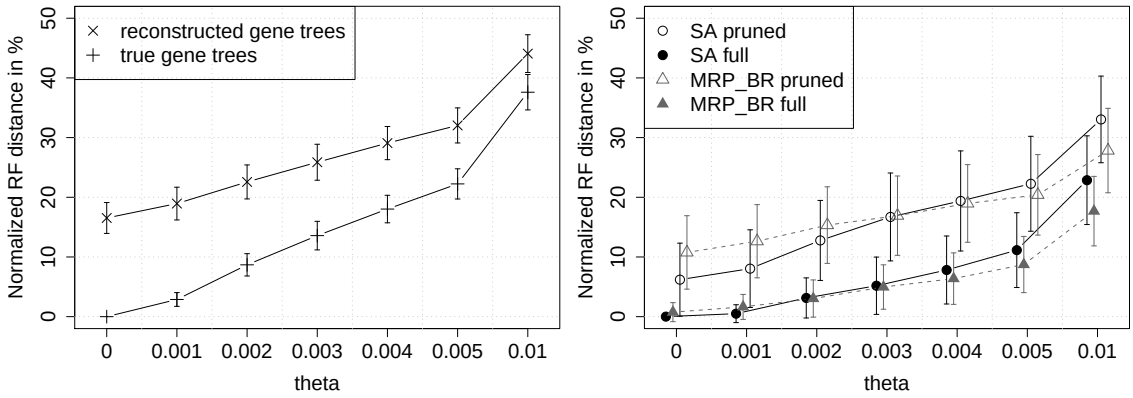
(a) Data set 1 **(b)** Data set 2

Figure 3.10: Distribution of normalized *RF* distances (500 simulations) of the source trees to the respective supertrees for the simulation setting S,c,$T_{0.005}$,n.

distance is now larger than the average distances of most methods. This might be due to the small number of genes (10) and the different sequence lengths (Figure 3.2a). Therefore, we show the distances of the reconstructed gene trees to the supertrees of the respective simulations for genes 1 and 2 (Figure 3.10). Data set 1 has the longest alignment and we observe its gene tree to be more similar to the superalignment tree compared to the other trees (Figure 3.10a). This behaviour is not present for the shorter data set 2 (Figure 3.10b). Thus, long genes mainly drive the superalignment reconstruction. If their gene trees are distant from the true species tree, the superalignment result will also deviate.

We also tested the methods on incongruent gene trees together with missing data (Figure 3.9b). That is, the same alignments were used but the information was pruned according to Figure 3.2a. Several methods show a lower mean accuracy than the phylogeny of a full gene, namely MinCut, ModMinCut, QILI and AvCon. MRP_BR (20.4 %), MRP_PU (21.6 %), MRP_I (21.1 %), MRF_BR (21.7 %) and MRF_PU





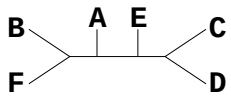
(a) Gene trees **(b)** SA and MRP_BR

Figure 3.11: Mean and standard deviation of the RF distances with different levels of incongruence. The results of the true and reconstructed gene trees are computed from the distribution of mean distances of all simulations. Note that the last step of θ is a doubling. The detailed results for $\theta = 0$ and $\theta = 0.005$ are shown in Figures 3.4 and 3.9, respectively. The remaining detailed results can be found in Appendix A. Different numbers of simulation replicates were used: 500 for $\theta = 0$ and $\theta = 0.005$ and 200 for the remaining settings.

(22.2 %) still outperform superalignment (22.3 %) on average, but the difference is marginal.

However, the above behavior is not representative for all degrees of incomplete lineage sorting. In Figure 3.11a, we see how the mean normalized RF distance of the true gene trees to the true species tree increases with θ. As a consequence, the distances of the reconstructed gene trees increase, too. At low θ (0.001-0.002), the reconstruction error exceeds the error introduced by incomplete lineage sorting. In this parameter area we observe figures similar to Figure 3.4 with SA performing better on average (Figures A.5 and A.6). With very high θ, however, the error introduced by incomplete lineage sorting is larger than the reconstruction error added to the true gene trees (Figure 3.11a). In this parameter area, we observe that MRP_BR slightly outperforms SA (Figure 3.11b). MRP_BR is used here as a representative supertree method, which usually performs well compared to other methods.

Note that in each case, the standard deviations are overlapping with the mean of the competing method (Figure 3.11b). However, we must keep in mind that the data are paired, i. e. for each of the 500 simulations with $\theta = 0$, we get one distance value for SA and one for MRP_BR. Thus, we tested the null hypothesis that the median difference



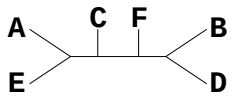
in these paired distances is 0 using the Wilcoxon signed-rank test (Table 3.4). The results shown in Table 3.4 support the conclusion that SA is significantly better in regions where the error introduced by phylogenetic reconstruction is prevalent, whereas MRP_BR is significantly better in regions where true gene trees differ a lot. Thus, if the reconstruction error dominates the error caused by incomplete lineage sorting, SA is the most accurate method by minimizing stochastic error. On the other hand, if incomplete lineage sorting is the prevalent source of gene tree incongruency, reconstructing the trees first and then applying a supertree method is favorable. However, in the case of high incomplete lineage sorting effects, the accuracies of all reconstruction methods are quite low. Figure 3.9 shows that about 8 % of the branches are reconstructed incorrectly with complete data and about 20 % with missing data for the best reconstruction methods.

3.3.3 Conclusions

We presented a detailed simulation study to assess the accuracy of superalignment, supertree and medium-level methods for reconstructing phylogenetic trees from multiple data sets. Although supertrees are often used to combine data of different kinds, our simulations only refer to sequence-based approaches. Morphological characters are not included due to the lack of reasonable probabilistic models to simulate their evolution. This study is first in comparing a broad range of methods for combining incomplete data sets. All conclusions are based on the specific implementation used for these methods as described in Section 3.2.

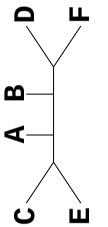
Gene features like sequence lengths and taxon overlap influence the accuracy of the presented methods. Instead of covering many different parameter combinations, we used the parameters of two very different natural data sets for the simulations. Furthermore, the true gene trees were generated from the true species tree in different ways (see also Table 3.3): (a) all gene trees were identical to the species tree, (b) the branch lengths but not the topology were gene-specific, (c) the gene trees from the original data were used as true gene trees and (d) the gene trees showed different topologies modeled by incomplete lineage sorting.

The first main result is that one of the matrix representation methods, which are the most abundant supertree reconstruction methods used in the literature, usually shows the second-best result after superalignment. Especially the MRP and MRF methods with Baum/Ragan-coding result in very accurate trees. Since these methods



θ	Sample size	Complete data			Missing data		
		p-value	Median difference	Confidence interval	p-value	Median difference	Confidence interval
0	500	$\mathbf{< 2.2 \times 10^{-16}}$	**1.5**	[1.5, 1.5]	$\mathbf{< 2.2 \times 10^{-16}}$	**2.5**	[2, 2.5]
0.001	200	$\mathbf{2.2 \times 10^{-9}}$	**1.5**	[1, 1.5]	$\mathbf{< 2.2 \times 10^{-16}}$	**2.5**	[2, 2.5]
0.002	200	0.69	-1.8×10^{-5}	[-0.5, 1.9×10^{-5}]	$\mathbf{2.2 \times 10^{-6}}$	**1**	[0.5, 1.5]
0.003	200	0.6	-2.47×10^{-5}	[-0.5, 0.5]	0.69	3.2×10^{-5}	[-0.5, 0.5]
0.004	200	$\mathbf{1.8 \times 10^{-3}}$	**-1**	[-1.5, -3.2×10^{-5}]	0.47	-6×10^{-5}	[-0.5, 0.5]
0.005	500	$\mathbf{1.6 \times 10^{-14}}$	**-1.5**	[-1.5,1]	$\mathbf{1.1 \times 10^{-8}}$	**-1**	[-1.5, -0.5]
0.01	200	$\mathbf{6.1 \times 10^{-16}}$	**-2.5**	[-3,-2]	$\mathbf{7.4 \times 10^{-15}}$	**-2.5**	[-3, -2]

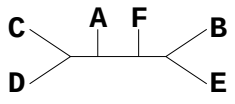
Table 3.4: Paired Wilcoxon signed-rank test. The distances of MRP_BR are compared with SA, thus positive median differences stand for higher distances in MRP_BR and negative differences stand for higher distances in SA. p-values < 0.05 and median differences whose 95 %-confidence interval does not include 0 are marked in bold.



are based on splits, bootstrap-based weighting can be easily incorporated, which is expected to further increase the accuracy of the reconstructed trees (Bininda-Emonds and Sanderson, 2001; Salamin *et al.*, 2002). Among the medium-level methods, SuperQP yields better results than the distance-based approaches, especially when data are missing. The accuracy of SuperQP is often consecutive to or among the accuracies of the matrix representation methods.

Second, in the case of complete gene trees, the majority-rule consensus method is also applicable. In all simulation settings with complete gene trees, some supertree methods perform better on average than the consensus method. In these cases, supertree branches that are supported by less than half of the gene trees are correctly resolved. This shows that, although supertree methods have been criticized for not being majority-rule methods (Goloboff, 2005), the resolution of additional branches can be favorable. Majority-rule supertree methods have also been suggested (Cotton and Wilkinson, 2007; Dong and Fernández-Baca, 2009). Some of these will be studied in detail in Section 3.4.

Third, we introduced the baseline distance as a measure to judge the benefit of the combination methods. The baseline distance for one setting is defined as the mean RF distance between the true species tree and the reconstructed gene trees using complete alignments. We observe that, for most of the simulation parameters studied here, QILI, average consensus, MinCut and modified MinCut show larger mean RF distances than the single gene trees. QILI has already been observed to be slightly worse than MRP (Piaggio-Talice *et al.*, 2004). Average consensus is clearly outperformed by SDM when data is missing. We applied both methods as medium-level methods by taking pairwise distances directly from the alignment distances, not from the reconstructed gene trees. While average consensus was suggested as a late-level method (Lapointe and Cucumel, 1997), SDM has been explicitly designed as a medium level method (Criscuolo *et al.*, 2006). Thus, average consensus may not be able to resolve the conflicts in the non-treelike distances. The behavior of MinCut can be partly explained by the fact that it resembles Adams consensus (Semple and Steel, 2000). This means that uncertain taxa will be placed at the root of subtrees, which can disturb quite a few splits, leading to high RF distances. The cut methods presented here implement a heuristic based on the rooted triplets in the gene trees. Recently, Lin *et al.* (2009) suggested another approach which maximizes the common rooted triplets in the supertree and the gene trees. They show that their method outperforms modified MinCut and MaxCut on example data sets.

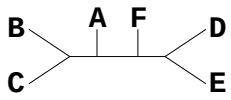


By comparing the accuracies of the reconstructed supertrees with the accuracies of the ML gene trees, we showed the baseline distance to be a reasonable criterion for excluding unsuitable methods. If the baseline distance cannot be improved by a data combination method, it is preferable to use only genes for ML reconstruction that are present in all taxa and to possibly sequence the missing genes in some taxa.

Finally, we observe that superalignment methods usually show the highest accuracy on average. This applies to incomplete data as well as gene-specific substitution parameters. Superalignment also results in the most consistent phylogenetic estimation when each method is not compared to a model tree but to the original result obtained with that very method (Figure 3.7). However, in the presence of high incongruency among true gene trees, if reconstruction error is not the main cause that gene trees differ from the species tree, the implicit weighting by sequence length can have a negative effect on the performance of superalignment leading to outperformance by the supertree method MRP_BR. This bias might be avoided by introducing a normalization, but then, the opposite and still unwanted bias could emerge. Furthermore, it has been discussed (e. g. Gatesy and Springer, 2004) that SA should be preferred over supertree methods since the former does not imply character weighting. On the other hand, Edwards (2009) argued recently that in the presence of gene tree conflict caused by coalescence effects as many genes as possible should be used and they should be weighted equally. This is consistent with our observation that supertree methods outperform superalignment in the presence of strong coalescence effects.

There are also some species tree reconstruction methods that use a coalescent model to account for the differences between true gene trees (e. g. Liu *et al.*, 2009). Kubatko and Degnan (2007) have shown that concatenation of gene alignments may be inappropriate when the gene tree histories differ considerably. The coalescent model can be applied for closely related species (e. g. grasshoppers; Carstens and Knowles, 2007), but severe problems caused by incomplete lineage sorting seem to not play a role among taxa of deep phylogenetic trees (e. g. for Metazoa, see Ewing *et al.*, 2008). Since these methods typically require complete data, we did not include them in our comparison. We rather concentrated on methods that were explicitly designed for missing data and that resolve conflicts of unknown source.

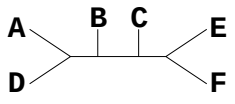
Our results are in general concordance with previously published comparisons. Eulenstein *et al.* (2004) used simulated data and find MRP and MRF to perform similar and better than MinCut and ModMinCut. Dutilh *et al.* (2007) used real data sets and



also find superalignment to perform best. Swenson *et al.* (2010) compare superalignment and weighted and unweighted MRP using biologically motivated simulations and find the highest accuracy for superalignment. We apply, however, a broader range of methods than previous studies.

All conclusions presented here are based on the accuracy measured by the mean *RF* distance. This does not imply that the methods presented as better on average always show superior results and could, thus, be used as a gold standard. Rather, we highly recommend to use several of the superior methods (considering also various levels of data combination) and to compare their results. The source of variation, i. e. why the reconstructed gene trees differ from the species tree, should also be taken into account, since it has an influence on the relative performance of the methods. If a treelike evolutionary history is assumed and true gene tree incongruence is unlikely or rare, superalignment results in the most accurate trees. This also holds in the presence of gene-specific substitution parameters and branch lengths, as has been observed before (Gadagkar *et al.*, 2005). But if the difference of the true gene trees to the species trees is the main source of variation, supertree methods are favorable. Applying a superalignment method to data with different underlying topologies or highly varying parameters has also been shown to be problematic (e. g. Mossel and Vigoda, 2005; Kolaczkowski and Thornton, 2004).

In the case of known gene tree variation, methods that model the assumed causes can also be applied. For incomplete lineage sorting, approaches like BEST (Liu, 2008) or TCM (Ewing *et al.*, 2008) are available. Note that these programs work with complete data, while we are interested in reconstructing species phylogenies from incomplete data. When exploring gene tree effects, like horizontal gene transfer or incomplete lineage sorting, gene trees have to be reconstructed and compared to a species tree. If the intention of an analysis is species tree reconstruction, however, external information may be considered: External information, like the rates of horizontal gene transfer, gene duplication or incomplete lineage sorting helps to judge whether complex evolutionary models are necessary to reconstruct the species tree. If these complex scenarios are not assumed to play a major role, application of superalignment minimizes the stochastic error. On the other hand, if gene-tree conflict is present but the underlying biological model is not known, supertree or medium-level methods can be applied. They account for gene tree variation but make no assumptions on the underlying evolutionary model causing the variation.



3.4 Majority-rule Supertrees

3.4.1 Definitions of Majority-rule Supertrees

In the previous section, we studied the practical performance of several supertree methods. Now, we present and analyze algorithms for two supertree methods, which were only described theoretically before.

Several of the supertree methods presented in the previous section have been criticized for not always displaying the majority of the gene tree splits (Goloboff, 2005). This problem is addressed by Cotton and Wilkinson (2007). They searched for a definition of majority-rule (MR) supertrees, which is applicable to gene trees on overlapping taxon sets. Barthélemy and McMorris (1986) showed that in the consensus setting, the majority-rule tree (Section 3.2.2) is a median tree under the Robinson-Foulds distance (RF, Section 2.1.3). However, the RF distance is only applicable to trees on the same taxon set. Thus, Cotton and Wilkinson (2007) gave two supertree definitions for the gene trees $\mathcal{T}^1$, ..., $\mathcal{T}^k$ on taxon sets X^1, ..., X^k and $X = \bigcup_i X^i$:

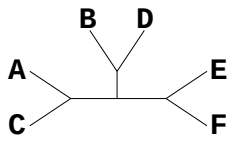
MR(-)supertrees: The objective function is minimizing $\sum_{i=1}^{k} \mathrm{RF}(\mathcal{P}|X^i, \mathcal{T}^i)$ for a supertree $\mathcal{P}$ on taxon set X. Thereby, $\mathcal{P}|X^i$ is the tree obtained by pruning all taxa not in X^i and collapsing nodes of degree 2.

MR(+)supertrees: For each gene tree $\mathcal{T}$, the span $< \mathcal{T} >$ is generated. $< \mathcal{T} >$ is the set of all bifurcating trees on taxon set X, where missing taxa are grafted onto $\mathcal{T}$ and all multifurcations are resolved. The objective function is minimizing $\sum_{i=1}^{k} \mathrm{RF}(\mathcal{P}, \mathcal{R}^i)$ for a supertree $\mathcal{P}$ and each $\mathcal{R}^i \in < \mathcal{T}^i >$.

Dong and Fernández-Baca (2009) define two subvariants for MR(+)supertrees where the spans are defined differently: The graft-only supertree span does not resolve any multifurcations in the gene trees, and the graft/refine supertree span does resolve some but not necessarily all multifurcations. Here we only considere bifurcating trees, then the subvariants are equivalent. Thereby, we use the following abbreviations for the distance functions:

$$d^-(\mathcal{P}, \mathcal{T}^i) = \mathrm{RF}(\mathcal{P}|X^i, \mathcal{T}^i) \quad \text{and} \quad d^+(\mathcal{P}, \mathcal{T}^i) = \min_{\mathcal{R}^i \in < \mathcal{T}^i >} \mathrm{RF}(\mathcal{P}, \mathcal{R}^i)$$

Thus the objective function for MR(-) is $\sum_i d^-(\mathcal{P}, \mathcal{T}^i)$ and the objective function for



d^+ is $\sum_i d^+(\mathcal{P}, \mathcal{T}^i)$. Minimizing these two different objective functions can result in different supertrees for MR(-) and MR(+) (Figure 3.12).

We first present the distance computations between one gene tree and one supertree for d^- and d^+, respectively (Section 3.4.3). The final supertree has to be found by summing up the distances over all gene trees and searching the tree space for the supertree with the minimal distance (Section 3.4.4). Finally, the two methods are also compared in simulations (Section 3.4.5).

3.4.2 Notation for Trees with Overlapping Taxon Sets

In this section, we extend the notations of topologies and splits (Section 1.3) for trees with overlapping taxon sets.

A supertree $\mathcal{P}$ is a tree on taxon set X and a gene tree $\mathcal{T}$ is a tree on taxon set $X_\mathcal{T} \subseteq X$. Both, $\mathcal{P}$ and $\mathcal{T}$, can be represented as sets of splits. The splits in $\mathcal{T}$ are called *partial* if $X_\mathcal{T} \subset X$. In contrast, the splits in $\mathcal{P}$ are called *plenary*. A plenary split $p \in \mathcal{P}$ ($p = Z_1|Z_2$) *extends* a partial split $t \in \mathcal{T}$ ($t = Y_1|Y_2$) if one of the following conditions holds:

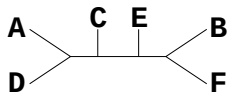
1. $Y_1 \subseteq Z_1$ and $Y_2 \subseteq Z_2$
2. $Y_1 \subseteq Z_2$ and $Y_2 \subseteq Z_1$

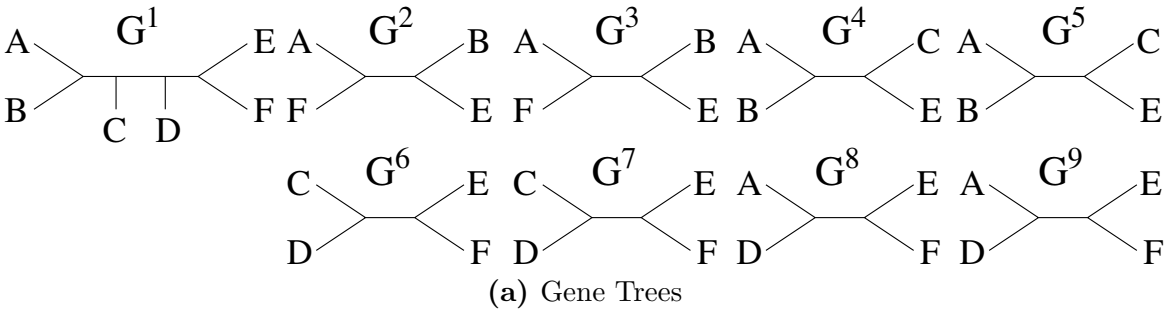
For example, the split $\{A, B, C\}|\{D, E, F\}$ extends the split $\{A, B, C\}|\{F\}$. The *extension* E_t is the set of all plenary splits extending a split t.

For comparing a supertree $\mathcal{P}$ and a gene tree $\mathcal{T}$, we need the following abbreviations:

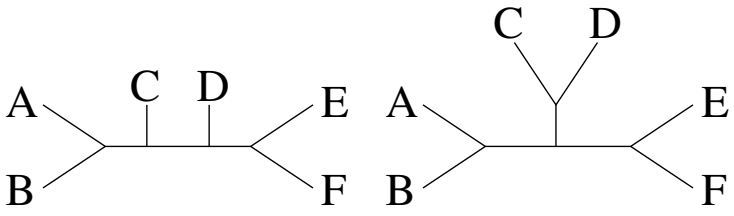
$$\mathcal{T}^* = \{t \in \mathcal{T} : E_t \cap \mathcal{P} \neq \emptyset\}, \quad \overline{\mathcal{T}^*} = \mathcal{T} \backslash \mathcal{T}^*,$$
$$\mathcal{P}^* = \mathcal{P} \cap \bigcup_{t \in \mathcal{T}} E_t, \qquad \overline{\mathcal{P}^*} = \mathcal{P} \backslash \mathcal{P}^*.$$

$\mathcal{P}^*$ are the splits in $\mathcal{P}$ which extend one of the splits in $\mathcal{T}$. $\mathcal{T}^*$ are the splits in $\mathcal{T}$ whose extension contains one of the splits in $\mathcal{P}^*$.



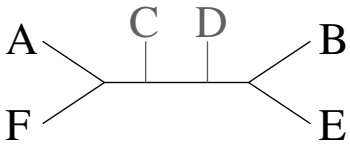


(a) Gene Trees

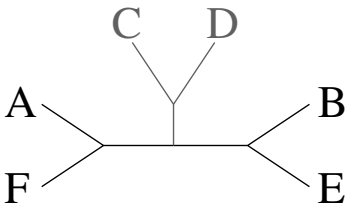


(b) Supertree S^1

(c) Supertree S^2

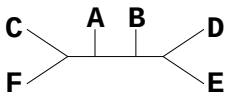


(d) $H^1 \in < G^2 >$



(e) $H^2 \in < G^2 >$

Figure 3.12: Example showing different results for MR(-) and MR(+). The gene tree set consists of the nine trees shown in (a). Note that G^4, ..., G^9, are compatible with both S^1 (b) and S^2 (c), they are included to enforce a unique tree as the result. MR(-) results in a distance of 4 for S^1 $[d^-(S^1, G^1) + d^-(S^1, G^2) + d^-(S^1, G^3) = 0 + 2 + 2]$ and of 6 for S^2 $[d^-(S^2, G^1) + d^-(S^2, G^2) + d^-(S^2, G^3) = 2 + 2 + 2]$. On the other hand, MR(+) results in a distance of 12 for S^1 $[d^+(S^1, G^1) + d^+(S^1, G^2) + d^+(S^1, G^3) = 0 + 6 + 6]$ and of 10 for S^2 $[d^+(S^2, G^1) + d^+(S^2, G^2) + d^+(S^2, G^3) = 2 + 4 + 4]$. S^1 is the unique supertree with MR(-), whereas S^2 is the unique supertree with MR(+). This discrepancy is caused by the "wrong" trees G^2 and G^3. For the supertree S^2, the 2-split $\{C, D\}|\{A, B, E, F\}$ can be placed anywhere causing a d^+ of 4, e. g. as in H^2 (e). But for the supertree S^1, none of the three supertree splits can be found in G^2 and each tree in the span causes a distance of 6, e. g. as in H^1 (d).



(a) S: Supertree **(b)** G: Gene tree **(c)** S_{pr}: Supertree S pruned to taxon set $\{C, D, E, R\}$ **(d)** G_{sp}: Optimal tree from the span of G

Figure 3.13: Example adapted from figure 3(c) in Cotton and Wilkinson (2007).

3.4.3 Distance Computations in the Matrix Representation Framework

The most popular supertree methods are split-based and code the gene tree splits in a matrix representation (MRep, Section 3.2.2). The MR-supertrees are defined via the split-based RF-distance. Thus, we define the MR-supertree methods as MRep-methods as well. Therefore, the first step is to code all splits of the gene tree and the supertree in the matrix representation. We explain the algorithms for d^- and d^+ along with the example in Figure 3.13.

Example: Matrix representations of the trees in Figure 3.13:

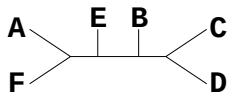
Supertree S (Figure 3.13a):

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
A	1	0	1	1	0	0	0	0	1	0
B	1	0	1	0	1	0	0	0	1	0
C	1	0	0	0	0	1	0	0	1	0
D	0	1	0	0	0	0	1	0	1	0
E	0	1	0	0	0	0	0	1	1	0
R	0	0	0	0	0	0	0	0	0	0

Gene tree G (Figure 3.13b):

	g_1	g_2	g_3	g_4	g_5	g_6
A	–	–	–	–	–	–
B	–	–	–	–	–	–
C	1	1	0	0	1	0
D	1	0	1	0	1	0
E	0	0	0	1	1	0
R	0	0	0	0	0	0

When comparing both trees, we see the following extension relationships: s_1 extends g_2, s_3 extends g_6, s_4 extends g_6, s_5 extends g_6, s_6 extends g_2, s_7 extends g_3, s_8 extends g_4, s_9 extends g_5, and s_{10} extends g_6. Thus, $S^* = \{s_1, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$, $\overline{S^*} = \{s_2\}$, $G^* = \{g_2, g_3, g_4, g_5, g_6\}$, and $\overline{G^*} = \{g_1\}$.



MR(-)supertrees

To compute d^- between a supertree $\mathcal{P}$ and a gene tree $\mathcal{T}$, the MReps are used the following way. First, in the MRep of $\mathcal{P}$, the lines corresponding to the taxa in $X \setminus X_{\mathcal{T}}$ are discarded. Thereby, identical columns emerge which then are merged, thus each column in the resulting MRep is unique. This results in the MRep of $\mathcal{P}|X_{\mathcal{T}}$. The MReps of $\mathcal{P}|X_{\mathcal{T}}$ and of $\mathcal{T}$ are compared, and the number of columns occurring in one but not the other is $\mathrm{RF}(\mathcal{P}|X_{\mathcal{T}}, \mathcal{T}) = d^-(\mathcal{P}, \mathcal{T})$.

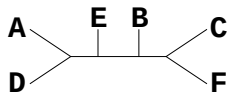
If all trees are fully resolved, we do not need to discard the lines for the missing taxa and merge equal splits explicitly. Instead, for each gene tree split t_i, all supertree splits can be compared, to determine whether they extend t_i. In the matrix representation framework, this is done by comparing each column from $\mathcal{P}$ to t_i. Thereby, an extension of t_i is found if there is one column from $\mathcal{P}$ that equals t_i in all or in none of the positions; in doing so, gaps are ignored. If no extension is found in the supertree splits, t_i adds 2 to d^-. Thus, $d^-(\mathcal{P}, \mathcal{T}) = 2 \times |\overline{\mathcal{T}^*}|$.

Example: Deleting the taxa A and B from the supertree S results in the tree S_{pr} (Figure 3.13c). $\mathrm{RF}(S_{pr}, G) = 2$, thus $d^-(S, G) = 2$. We also see this by comparing the splits in the corresponding MReps. Therefore, the lines for A and B are discarded and identical columns are merged, which results in the MRep of S_{pr}:

	s'_1	s'_2	s'_3	s'_4	s'_5	s'_6
C	1	0	0	0	0	0
D	0	1	1	0	0	0
E	0	1	0	1	0	0
R	0	0	0	0	1	0

Thus, s'_2 is not contained in the gene tree, g_1 is not contained in the supertree and d^- is 2. As described, we can also compute this by comparing the original MReps. Only g_1 is not extended, thus $d^- = 2$.

The objective function of MR(-)supertrees is equivalent to the one of matrix representation with compatibility (MRC, Section 3.2.2). MRC maximizes the number of columns in the MRep which are contained in the supertree. On the other hand, MR(-)supertrees minimize the number of columns not in the supertree. Thus, the two objective functions are the same if the gene trees are fully resolved and if only fully resolved supertrees are considered.



MR(+)supertrees

Defining d^+ in the matrix representation framework is less straightforward. We first define the split extension score and subsequently show that it is equivalent to d^+. The *split extension score* (SES) between $\mathcal{P}$ and $\mathcal{T}$ is the number of splits in $\mathcal{P}$ which occur in no extension of any of the splits in $\mathcal{T}$, thus SES=$|\overline{\mathcal{P}^*}|$.

It is easy to compute whether a supertree split occurs in an extension of a gene tree split (see previous section). Thus, the SES between a supertree and a gene tree can be easily computed in time $\mathcal{O}(|\mathcal{P}| \times |\mathcal{T}|)$. Since the number of splits in a tree grows linear with the number of taxa, the SES can be computed in time $\mathcal{O}(n^2)$.

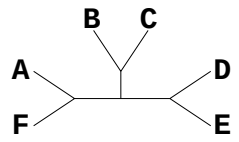
Example: $\overline{S^*} = \{s_2\}$, SES=$|\overline{S^*}| = 1$. The optimal tree from $< G >$ is G_{sp} (Figure 3.13d), thus d^+ is 2.

One can easily show that for fully resolved gene trees, SES is equivalent to d^+, in particular $d^+(\mathcal{P},\mathcal{T}) = 2 \times SES(\mathcal{P},\mathcal{T})$. Therefore, we define $\overline{d^+}(\mathcal{P},\mathcal{T}) = \min_{\mathcal{R}\in<\mathcal{T}>} \overline{\mathrm{RF}}(\mathcal{P},\mathcal{R})$ where $\overline{\mathrm{RF}}(\mathcal{P},\mathcal{R})$ is the number of splits in $\mathcal{P}$ but not in $\mathcal{R}$. For bifurcating trees, $\overline{\mathrm{RF}}(\mathcal{P},\mathcal{R}) = \mathrm{RF}(\mathcal{P},\mathcal{R})/2$. We thus show $\overline{d^+}(\mathcal{P},\mathcal{T}) = SES(\mathcal{P},\mathcal{T})$. We show the inequalities in both directions:

$SES(\mathcal{P},\mathcal{T}) \leq \overline{d^+}(\mathcal{P},\mathcal{T})$: SES counts the splits in $\mathcal{P}$ which occur in no extension of a split in $\mathcal{T}$. Therefore, the splits in $\overline{\mathcal{P}^*}$ cannot be in any $\mathcal{R} \in< \mathcal{T} >$, and $\overline{\mathrm{RF}}(\mathcal{P},\mathcal{R}) \geq |\overline{\mathcal{P}^*}| = SES(\mathcal{P},\mathcal{T})$ for every $\mathcal{R} \in< \mathcal{T} >$.

$\overline{d^+}(\mathcal{P},\mathcal{T}) \leq SES(\mathcal{P},\mathcal{T})$: We show that there is an $\mathcal{R} \in< \mathcal{T} >$ whose $\overline{\mathrm{RF}}(\mathcal{P},\mathcal{R})$ is not larger than $SES(\mathcal{P},\mathcal{T})$.

First, we show that $\mathcal{P}^*$ and $\mathcal{T}$ are compatible: By construction, $\mathcal{P}^* \cup \mathcal{T}$ is pairwise compatible. The pairwise compatibility theorem states that pairwise compatible splits on the same taxon set are compatible (McMorris, 1977; Estabrook and McMorris, 1980). A set of partial splits is compatible if the partition intersection graph is chordal or can be chordalized (Semple and Steel, 2002, see example in Figure 3.14). *Chordal* denotes the property that every cycle with at least four nodes has an edge connecting two non-consecutive nodes. If the graph is *chordalized*, then these edges connecting the non-consecutive nodes are introduced. For partition intersection graphs, the additional property must hold that no nodes of the same split are connected. Here, we have the special case that the set is built from full splits and partial splits which are all pairwise compatible. Then,



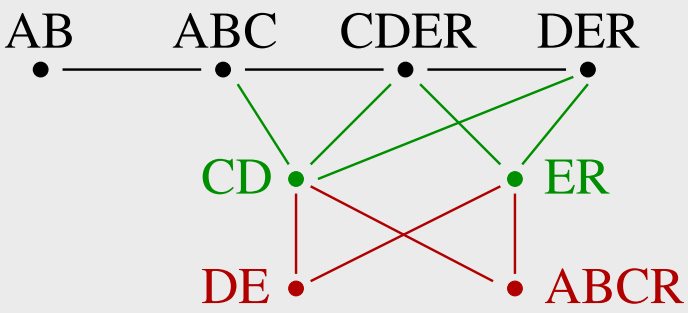


Figure 3.14: Partition intersection graph of the example in Figure 3.13. The node set of the graph corresponds to the taxon sets of each partition. Taxon sets are color-coded as follows: black - S^*, green - G, red - $\overline{S^*}$. Two nodes are connected if their taxon sets overlap. Here, the overlap between S^* and $\overline{S^*}$ is not shown. We see that no cycles of length ≥ 4 are introduced when only nodes in S^* and G are considered. However, the graph of $\overline{S^*}$ and G forms a cycle of length four. This cycle cannot be chordalized since CD|ER $\in G$ and ABCR|DE $\in \overline{S^*}$.

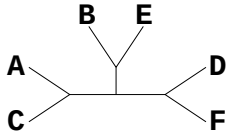
the edges between nodes originating from full and partial splits cannot introduce cycles of lengths greater than 3. Thus, the partition intersection graph is chordal and $\mathcal{P}^* \cup \mathcal{T}$ is compatible.
Second, if $\mathcal{P}^* \cup \mathcal{T}$ is compatible, there is an $\mathcal{R} \in < \mathcal{T} >$ with $\mathcal{R} \supseteq \mathcal{P}^*$. Thus, $\overline{\mathrm{RF}}(\mathcal{P}, \mathcal{R}) \leq |\overline{\mathcal{P}^*}| = SES(\mathcal{P}, \mathcal{T})$.

As for MR(-)supertrees, we only consider fully resolved MR(+)supertrees. Multifurcating supertrees can result, if multiple bifurcating trees have the same distance, then the supertree is the strict consensus of these best trees (see also next section).

3.4.4 Implementation

The two distance functions were implemented in python based on the matrix representation algorithms presented in Section 3.4.3. These implementations are combined with an heuristic tree search. The resulting program is called `PluMiST` (Plus- and Minus Supertrees) and is available upon request. An option of `PluMiST` is the type of algorithm (`minus` or `plus`). According to this option, the corresponding objective function is used. In the following we outline the principle of the approach.

1. **Starting tree** The starting tree is either a given tree, a random tree, or an *iterative minus tree*. The latter builds a starting tree based on minimizing d^- in each step. Therefore, a random order of taxa is generated. The quartet topology for the



first four taxa is computed by choosing the topology most frequent among the gene trees. Then, taxa are added iteratively. In each step, the *informative* gene trees are determined. A gene tree is informative, if it contains the taxon added in this step and at least three of the taxa already inserted. For each possible insertion point, d^- between the proposed tree and the informative gene trees is evaluated. The insertion point with the minimal sum of d^- is chosen as the final insertion point. If no informative gene trees are found, a different taxon is chosen for insertion. Ties are always resolved randomly.
Note that an iterative plus tree cannot be computed that easily. While the starting tree is growing iteratively, computing d^+ would mean to find the optimal positions of taxa both in the starting tree and in the gene tree. The taxon sets of the two trees are overlapping and there may be taxa in one tree but not in the other and vice versa. In this case, the SES is not applicable.

2. **Heuristic search** The starting tree is optimized by *nearest neighbor interchange* (NNI). In each step, all possible interchanges are evaluated, whether they improve the sum of d^- or d^+, respectively, and the best tree is kept for continuing the search. If multiple trees are found, one is chosen randomly for continuing the search but all optimal trees are kept. The search stops if no NNI yields an improvement.

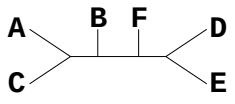
3. **Consensus** The tree with the best score found is returned. If there are multiple best trees, the strict consensus of these is returned as the supertree.

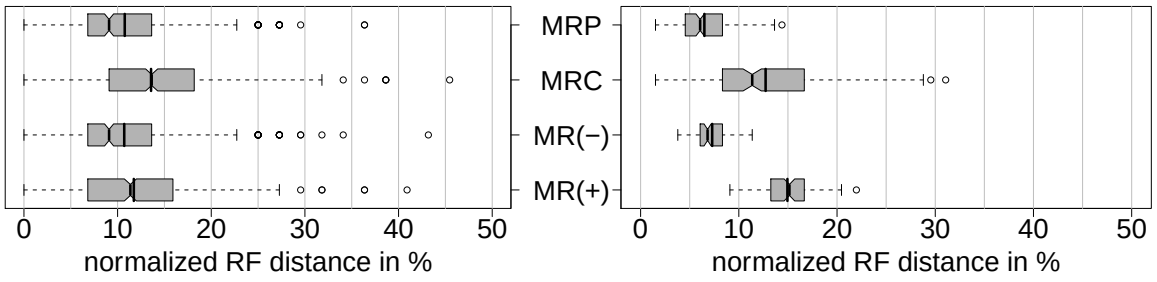
A larger proportion of the tree space can be explored by computing multiple starting trees, each followed by a heuristic search. Thus, steps 1 and 2 can be run multiple times and in step 3, the best trees of all runs are considered for the strict consensus.

3.4.5 Simulation Results

To assess the performance of majority-rule supertrees in simulations, `PluMiST` is used to compute MR(-)- and MR(+)supertrees the following way: Ten replications of starting tree computation and subsequent heuristic search with the respective objective function are performed. Afterwards, the results of these ten replications are merged and the strict consensus of the trees with the smallest distance is computed.

We compare the MR-supertree implementations with two published MRep-methods. First, we use MRC as implemented in Clann. And second, we use MRP_BR as im-





(a) (S,m,E,n) **(b)** (L,m,E,n)

Figure 3.15: Simulation results for MR(-) and MR(+). MRP refers to MRP_BR. For an explanation of the boxes, see Figure 3.4 (page 52).

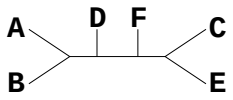
plemented in PAUP*, which is named MRP from now on. In Section 3.2.2, these methods and implementations are described in detail. The simulated alignments from the simulations with missing data and equal simulation parameters of the genes are used (S,m,E,n and L,m,E,n; Section 3.3.2).

For the small data set (Figure 3.15a), we see differences in the performance of the MRep-methods. MR(-) and MRP perform about equally well (mean values of 10.7 % and 10.8 %, respectively). Both are better than MR(+) and MRC (mean values of 11.7 % and 13.6 %, respectively). These differences get more pronounced with the large data set, which also shows a larger amount of missing data (Figure 3.15b). Here, MRP (6.5 %) is best, followed by MR(-) (7.3 %). MR(+) and MRC are clearly outperformed by these methods (mean values of 14.9 % and 12.7 %, respectively).

3.4.6 Summary

We formulated the objective functions for majority-rule supertrees (Cotton and Wilkinson, 2007) in the matrix representation framework. With this tool, we could show that for bifurcating trees the objective function of MR(-) is equivalent to the objective function of MRC. Furthermore, it allows us to formulate a computation of d^+ for fully resolved trees. The complexity of d^+ is unknown (Cotton and Wilkinson, 2007; Dong and Fernández-Baca, 2009) but our computation for bifurcating trees runs in polynomial time.

This allowed us to implement both objective functions in combination with an easy tree-search algorithm (NNI) in the program `PluMiST`. We are aware that other search algorithms can explore the tree space more efficiently, but we only wanted to get a first



insight into the performance of majority-rule supertree methods. The performance of our MR(-)implementation is comparable to MRP for the small data set and slightly worse for the larger data set. Furthermore, MR(-) clearly improves the current MRC implementation.

Finally, the simulation results clearly show a better performance of MR(-) compared to MR(+). This discrepancy traces back to the objective functions, since the same search heuristic was used for the comparison. MR(+) can put missing taxa at the best fitting positions in the gene trees. However, if taxa are missing in many trees, the information for these taxa may be outvoted by the "information" in the extensions. This is consistent with our observation that MR(+) is largely affected by missing data (Figure 3.15b compared to Figure 3.15a). Therefore, we need to be aware of the assumptions underlying the null models of these methods, which are discussed in the next section.

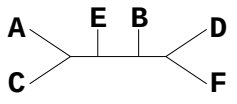
Appendix While writing up this section, a related paper by Dong *et al.* (2010) appeared. The authors show, that the problem of finding a supertree minimizing the sum of d^+ is NP-complete and they provide an exact solution using integer linear programming.

3.5 Effects of Null Models on the Tree Shape Bias of Supertree Methods

3.5.1 Introduction

Although several desiderata for supertree methods exist (Wilkinson *et al.*, 2004, see also Section 3.1), only few of them have been studied in greater detail, examples include shape bias (Wilkinson *et al.*, 2005a) or pareto properties (Wilkinson *et al.*, 2007). Here, we investigate supertree methods in the presence of no or little information. This is modeled by null distributions on the space of topologies (Section 1.3.2). The results in this section were first presented in Kupczok (2009).

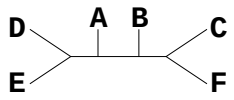
Since we use these theoretical distributions, we use the term *input tree* instead of gene tree for the input of supertree methods. The analyses in this section refer only



to trees with the same taxon set, also known as the consensus setting (e. g. Wilkinson *et al.*, 2007). Furthermore, we will only consider bifurcating topologies. Thus, we restrict the definition of topologies from Section 1.3.2 as follows: Now, a *(tree) topology* for n taxa is an unrooted, leaf-labeled, bifurcating tree with n leaves. Since branch lengths are not relevant in this section, we use the term *tree* simply for topologies. We are only interested in the interior splits, thus we call them *splits* ignoring the trivial splits.

We consider two matrix representation methods, matrix representation with compatibility (MRC) and matrix representation with parsimony (MRP) with Baum-Ragan coding (Section 3.2.2). As in Section 3.4.5 we name them simply MRC and MRP since only one coding is investigated. Several details of the computation of MRC and MRP differ from the previous description in Section 3.2.2. First, the MRep is generated by a python script. The number of columns in an MRep is denoted as length l. This is the total number of splits in the input trees. Furthermore, MRC maximizes the number of compatible columns. Equivalently, the method can also be defined as *minimizing* the number of columns not present in the supertree (as in the definition of MR(-)supertrees, Section 3.4.3). We use this notation from now on and denote the number of columns not present in the supertree as compatibility length (CL). The compatibility length is analoguous to the parsimony length (PL) minimized by MRP. The PL is the sum of the parsimony lengths over all the columns in the MRep (see also Section 1.2.2). Note that for $n = 5$, MRC and MRP are equivalent since only 2-splits are present. A 2-split has a PL of 1 if it occurs in a tree and a PL of 2 otherwise. Thus, for each five-taxa-tree, the PL is l+CL.

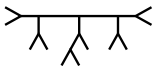
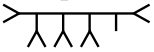
In contrast to the simulations described before, we found all optimal trees in this section by exhaustive search. Note that there can be multiple supertrees with the same minimal CL and PL, respectively. Thus we talk about the "supertree set" and we do not apply a consensus. Optimal MRC trees were found with a python script which evaluates all possible topologies for a particular number of taxa and optimal MRP trees are computed with the branch-and-bound option in PAUP* (Swofford, 2002).



3.5.2 Results

Perfect Distributions

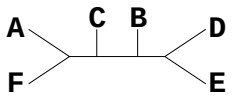
First, we perfectly model the null distributions. Under the perfect PDS model, the MRep contains exactly one split of each kind (e. g. 25 columns for $n = 6$, Table 1.1, page 8). Then the compatibility lengths are equal for all trees. Since each tree contains $n - 3$ inner splits, the CL of each tree is $l - n + 3$. The parsimony lengths of all trees are also equal if each possible split is given as input (Steel, 1993). Thus, for MRC and MRP, the supertree set contains every tree if the perfect PDS model is used as input.

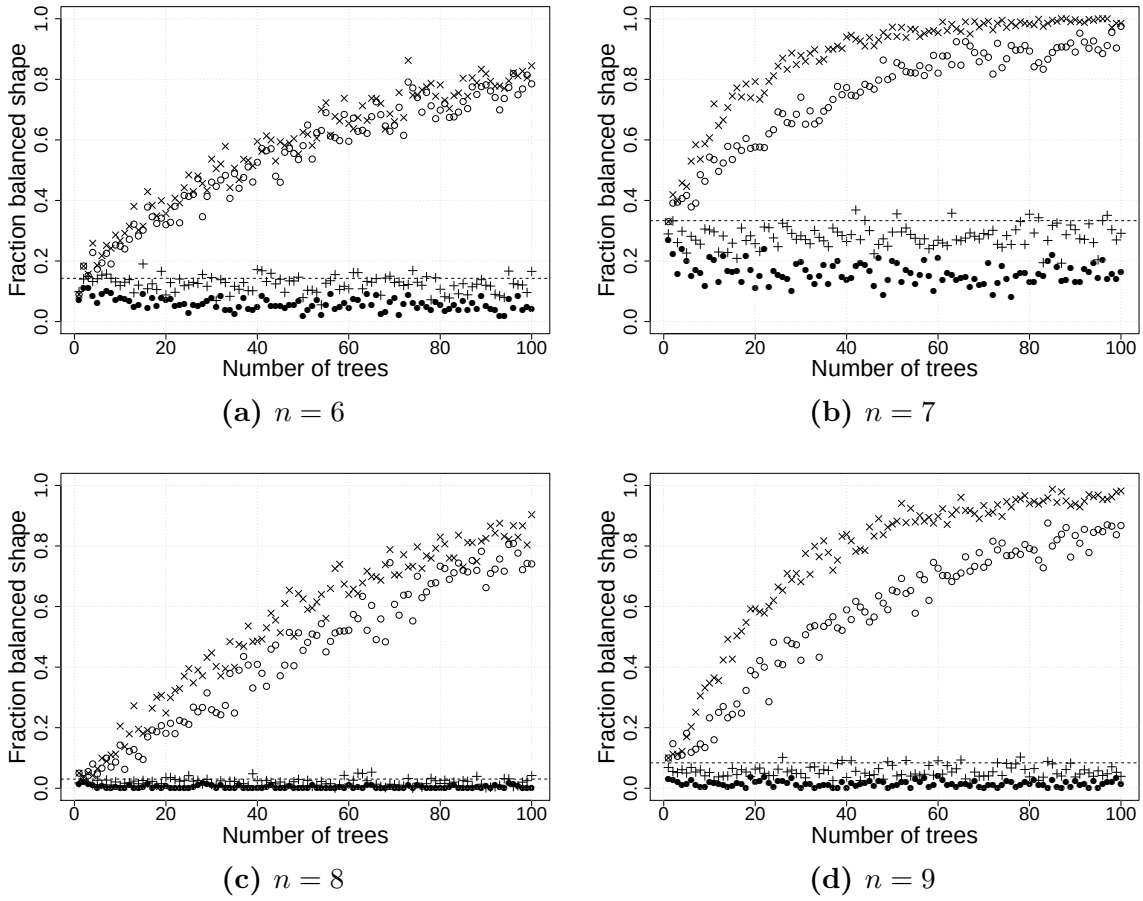
Under the perfect PDA model, the MRep is built by coding each tree once. Then the set of supertrees for MRC contains only trees of the same shape. This shape is called the *optimal shape* for MRC. The analogue holds for MRP. For six to nine taxa, the optimal shape is the same for MRC and MRP (shapes $S_{6,2}$, $S_{7,2}$, $S_{8,3}$ and $S_{9,5}$; shapes marked gray in Table 3.5, page 81). E. g. for $n = 6$, the supertree set contains 15 trees, all being of shape $S_{6,2}$. In the following, we call these four optimal shapes *balanced shapes*. Note that MRC and MRP do not always result in the same optimal shapes. E. g. for $n = 11$, the optimal shape with MRC is  whereas the optimal shape with MRP is  .

Resampling Randomly from the Distributions

Having exactly one split or one tree of each kind as input is a very strong assumption. Thus, we resampled a particular number of trees t from the distribution. For the PDS model, a sample of t trees corresponds to $t \cdot (n-3)$ randomly drawn splits. We evaluate the fraction of balanced shapes among the resulting supertrees. If the supertree set for one data set contains p trees and thereof q exhibit the balanced shape, this data set shows a fraction of q/p of the balanced shape. The results in Figure 3.16 are obtained by averaging this fraction over 100 randomly generated data sets.

Under the PDA model, we would expect a certain fraction of the balanced shape (Table 1.2, page 9). These fractions are marked dashed in Figure 3.16. With the PDS model, the balanced shape occurs a bit less frequently in the supertree sets than expected. In contrast, if the input trees are generated under the PDA model, the supertrees show the balanced shapes more often than expected. The observed fraction





(a) $n = 6$ **(b)** $n = 7$

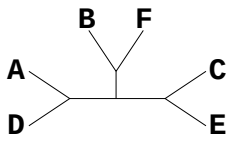
(c) $n = 8$ **(d)** $n = 9$

Figure 3.16: Resulting supertree shapes if random input trees are given (100 repetitions for each number of trees). The dashed line is the expected fraction of the balanced shape under the PDA model (Table 1.2). Labels: × MRC - PDA, ∘ MRP - PDA, + MRC - PDS, • MRP - PDS

of balanced shapes is growing with the number of input trees. Furthermore, the bias is less strong for MRP, in particular for uneven numbers of taxa.

Perfect Distributions with Phylogenetic Information

Next, we ask how the methods behave in the presence of "little" phylogenetic information disturbed by noise. First, noise is modeled by the perfect PDA model. Note that the supertree methods show a bias towards balanced shapes if the perfect PDA model is used as input (shapes marked gray in Table 3.5). The little phylogenetic information is modeled by adding the "true" tree i times to the perfect PDA model. The *critical number* i_c is the smallest i such that the true tree is in the supertree

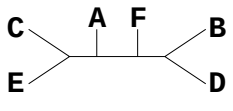


Shape		MRC		MRP	
S_5		1	(6.7 %)	1	(6.7 %)
$S_{6,1}$		6	(5.7 %)	6	(5.7 %)
$S_{6,2}$		1	(0.95 %)	1	(0.95 %)
$S_{7,1}$		60	(6.3 %)	60	(6.3 %)
$S_{7,2}$		1	(0.11 %)	1	(0.11 %)
$S_{8,1}$		720	(6.9 %)	900	(8.7 %)
$S_{8,2}$		630	(6.1 %)	990	(9.5 %)
$S_{8,3}$		1	(0.01 %)	1	(0.01 %)
$S_{8,4}$		270	(2.6 %)	540	(5.2 %)
$S_{9,1}$		8820	(6.5 %)	12600	(9.3 %)
$S_{9,2}$		8820	(6.5 %)	13860	(10.3 %)
$S_{9,3}$		7560	(5.6 %)	13860	(10.3 %)
$S_{9,4}$		3150	(2.3 %)	7560	(5.6 %)
$S_{9,5}$		1	(0.0007 %)	1	(0.0007 %)
$S_{9,6}$		1260	(0.93 %)	1260	(0.93 %)

Table 3.5: Critical numbers (i_c): Numbers of trees necessary to be added to the perfect PDA model until the true tree is in the supertree set. The fraction in parentheses is the number of trees added divided by the number of trees in the perfect PDA model. Shapes marked gray are the optimal shapes with MRC and MRP under the perfect PDA model (the "balanced" shapes).

set. We observe that i_c depends strongly on the shape of the true tree (Table 3.5). It ranges from 1 tree for the balanced shape up to 10 % of the total number of trees. In general, the critical numbers are lower for MRC than for MRP.

If the number of true trees is below i_c, then one or few incorrect supertrees are found. E. g. for $n = 6$ and $S_{6,1}$, the true tree is $\mathcal{T}^1$ (`(A,B,(C,(D,(E,F))))`). If this tree is added once to the supertree set, the supertree for MRC and MRP, respectively, is $\mathcal{T}^2$ (`((A,B),(E,F),(C,D))`). We see that the supertree is balanced and includes the two 2-splits of the true tree (`AB|CDEF` and `EF|ABCD`). For MRC, this result can be explained by the compatibility lengths: The balanced tree $\mathcal{T}^2$ has a length of $315-3\times15+1 = 269$ (315 splits from the PDA model, thereof 3×15 are the 2-splits and one 3-split from the true tree). Analogously, the unbalanced tree $\mathcal{T}^2$ has a length of $315-2\times15-9 = 276$. For this example, i_c is computed by $315 - 3 \times 15 + i_c = 315 - 2 \times 15 - 9$, thus $i_c = 6$ and from $i = 7$ on the true tree is the only supertree. We see that below the critical



number the uneven split distribution, which prefers 2-splits, outvotes the 3-split.

If one "true" tree is added to the perfect PDS model, then the supertree equals the true tree for MRC and MRP and all shapes presented in Table 3.5.

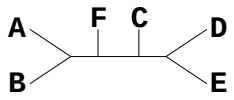
Resampling Randomly from the Distributions and adding Phylogenetic Information

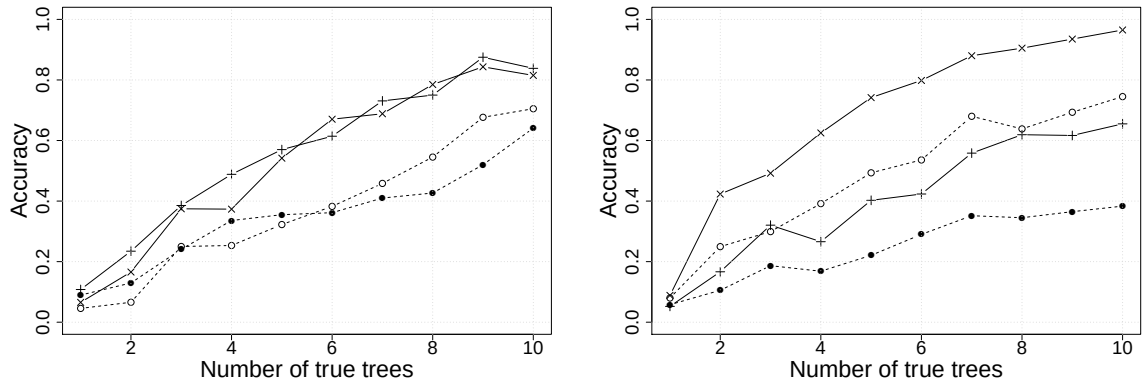
We see that 10 % of phylogenetic information is sufficient such that the supertree equals the true tree in a perfect setting. However, this behavior is disturbed by noise. To show this, we add the "true" tree with a fraction of 10 % to the random trees used for the analysis displayed in Figure 3.16a-c. Therefore, we only take the data sets with 10, 20, ..., 100 trees and add the true tree 1, 2, ..., 10 times. The true tree is either one topology showing the balanced shape or one topology showing the unbalanced shape. The accuracy for one data set is $1/p$, if p supertrees are found and the true tree is among them and 0 otherwise. The accuracies are averaged over 100 data sets.

Under the PDS model (Figure 3.17, left column), there is no observable difference in accuracy between the true trees showing different shapes. For the PDA model (Figure 3.17, right column), however, the tree with the balanced shape is reconstructed correctly with higher probability than the tree with the unbalanced shape. Furthermore, the accuracies are increasing with the number of true trees. Apparently, 10 true trees out of 110 trees provide more information compared to 1 true tree out of 11 trees. The accuracies are higher for MRC than for MRP. This is consistent with the results for the perfect setting that MRP needs more trees until the true tree is in the supertree set (Table 3.5).

3.5.3 Null Models of Majority-rule Supertrees

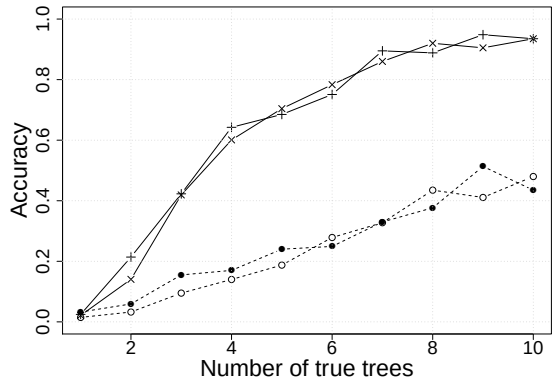
In Section 3.4.5, we found a clearly better performance of MR(-)supertrees compared to MR(+)supertrees. Now, we can explain this with the different implicit null models underlying both specifications of majority-rule supertrees. First, MR(-)supertrees extend the splits independently as is modeled by the PDS model. On the other hand, MR(+)supertrees extend the trees and thus have to keep the constraints introduced by compatible splits. In this section, we showed that these constraints can lead to a shape bias towards more balanced trees. Thus, we expect an unbalanced supertree to be harder to reconstruct by the MR(+)supertree method. Then the conflicting trees



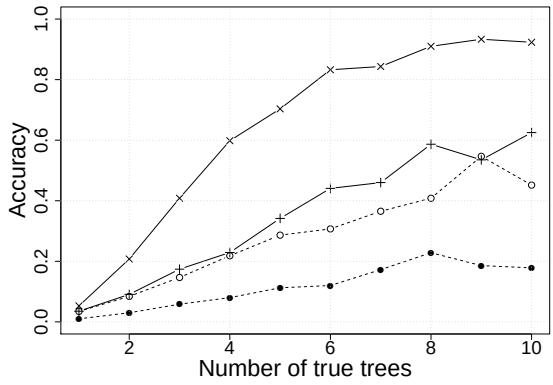


(a) $n = 6$, PDS

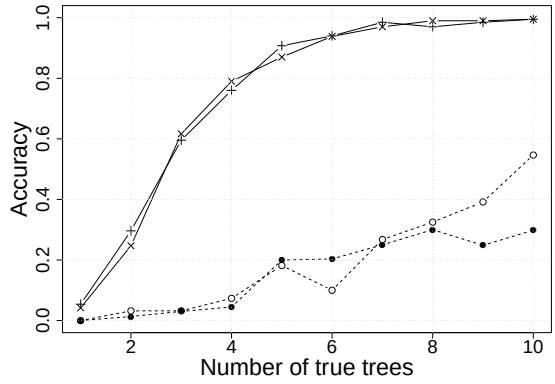
(b) $n = 6$, PDA



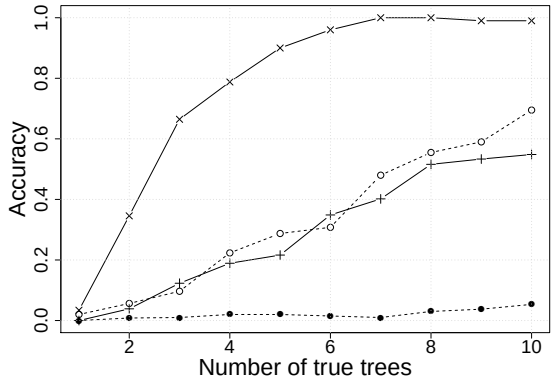
(c) $n = 7$, PDS



(d) $n = 7$, PDA

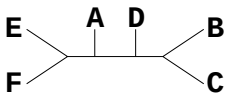


(e) $n = 8$, PDS



(f) $n = 8$, PDA

Figure 3.17: Adding information to random trees. The random trees are taken from Figure 3.16a-c and a fraction of 10 % of the true tree is added. Accuracy is the fraction the true tree was found by the supertree methods (average over 100 repetitions). Labels: $-\times-$ Balanced - MRC, $-+-$ Unbalanced - MRC, ---○---Balanced - MRP, ---●---Unbalanced - MRP. Balanced shapes are $S_{6,2}$, $S_{7,2}$ and $S_{8,3}$. Unbalances shapes are $S_{6,1}$, $S_{7,1}$ and $S_{8,1}$ (see Table 3.5).



	MR(-)		MR(+)	
k	false	missing	false	missing
2	324	233	725	282
3	55	423	350	234
24	0	0	0	190
25	0	0	0	184
26	0	0	0	118
32	0	0	0	51
33	0	0	0	147

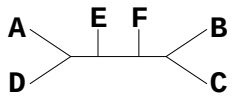
Table 3.6: Distribution of false and missing k-splits with MR(-) and MR(+) among 200 replications with L,m,E,n (Section 3.4.5). Overall, 725 and 350 wrong 2-splits resp. 3-splits are found with MR(+)supertrees and considerably fewer with MR(-)supertrees. This discrepancy is not seen when comparing the absolute numbers of 2- and 3-splits that were missed in the reconstructions. The excess of wrongly reconstructed 2-splits and 3-splits with MR(+) has a negative influence of the accuracy of splits with high k. The model tree contains five splits with $k \geq 24$. All of these are correctly recovered in each MR(-)supertree. However, the respective splits are missing in many MR(+)supertrees.

can cause a very high distance on an unbalanced supertree compared to a balanced supertree (see also example in Figure 3.12). With this relationship we would predict that more splits with smaller k are reconstructed in our simulations compared to splits with higher k. Table 3.6 shows this relationship for the simulation L,m,E,n which has a large amount of missing data.

This data was created with the heuristic method implemented in `PluMiST` (Section 3.4.4). To assess the shape bias of MR(+) in detail, an exact method would be advantageous. Therefore, e. g. the exact solution of Dong *et al.* (2010) could be used.

3.5.4 Conclusions

Wilkinson *et al.* (2005a) report a shape bias towards an unbalanced tree for MRP which at first view contradicts our result that both MRP and MRC prefer balanced shapes. However, their setting is different from ours. They investigate two arbitrarily chosen input trees of different shapes and observe an asymmetry in the parsimony lengths: the PL of a balanced input tree on an unbalanced supertree is shorter than vice versa. This can be explained by an inherent feature of the parsimony score: The PL of a coded k-split does not exceed k on any tree. More balanced trees contain more splits with small k, i. e. where the numbers of taxa are unevenly distributed.

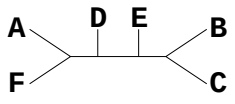


They necessarily have lower maximal parsimony lengths. Splits with higher k do not only have higher maximal PLs, but in addition, the PL distribution on random trees is shifted towards higher values (Maddison and Slatkin, 1991). This explains that coding the balanced tree in the MRep is favorable compared to coding the unbalanced tree.

For more than two random input trees, however, not unbalanced but balanced supertrees are preferred (Figure 3.16). This holds for MRC as well as MRP although compatibility lengths are not asymmetric. Furthermore, it even holds if unbalanced and balanced shapes are not uniformly distributed among the input shapes, but when unbalanced input tree shapes are favored (e. g. for $n = 6$, the input trees are drawn from a distribution which contains the unbalanced shape in 86 %, see also Table 1.2). This shape bias is positively misleading, i. e. it is growing with the number of input trees. However, it grows slower for MRP. An explanation may be that the bias towards unbalanced shapes due to the asymmetric parsimony lengths (Wilkinson *et al.*, 2005a) acts as counterbalance.

This shape bias is only observed for the PDA model (all trees are equally likely), not for the PDS model (all splits are equally likely). The two null models do also behave differently if little phylogenetic information is added. Thereby, one or more "true" trees are added to the perfect distributions. Only under the PDA model, the supertree set may not contain the true tree for some tree shapes. Instead, up to 10 % of the true tree are needed such that it is contained in the supertree set (Table 3.5). When adding one true tree with an unbalanced tree shape, the supertree is not this tree but a balanced tree. This balanced supertree preserves only some splits present in the true tree. The shape bias is also present if not the complete list of trees but random trees are given, and the true tree is added with a fraction of 10 % (Figure 3.17). In this case, a tree with a balanced shape is reconstructed correctly with higher probability than a tree with an unbalanced shape.

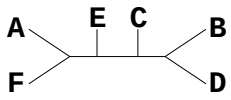
We note that the shape bias presented here is not caused by an unwanted feature of the supertree methods, but is a consequence of the complex space of trees. Under the PDA model, 2-splits are more likely than other splits (Table 1.1) and no distribution on bifurcating trees ensures that all splits are equally likely (Steel and Pickett, 2006). This results in a bias towards balanced trees with any split-based supertree reconstruction method. The same relation causes the problem that uniform priors on tree topologies do imply non-uniform priors on splits in Bayesian analysis, thus posterior probabilities may be biased (Pickett and Randle, 2005).



In general, we do not assume that the shape bias has an impact on usual supertree reconstruction. Here, problems occur if 90 % to 100 % are random data or for small numbers of trees. We do not expect this to be the case in any analysis. However, in sparse data sets, only few input trees may carry the information for some splits. If they are highly conflicting, the shape bias may play a role at least locally.

Furthermore, we cannot draw any conclusions about whether MRC or MRP should be preferred since they show the bias to a different extent depending on the model. If only random trees are given as input, the bias is less strong in MRP (Figure 3.16). On the other hand, if little information is added, MRC can find the true tree better (Figure 3.17).

Our findings allow implications about the design of supertree methods which explicitly model missing data. We conclude that modeling missing data by generating all possible trees may introduce a bias towards more balanced tree shapes when applying split-based supertree methods. To date, supertree methods usually do not model missing data explicitly. However, MR(+)supertrees are an example where tree-shape effects play a role, if an inappropriate null model is chosen. We can now explain the results for MR(+) from Section 3.4.5 with a bias towards balanced shapes. MR(-) resp. MRC treat missing taxa as gaps, which is equivalent to generating all possible binary characters by replacing the gaps with `0`s and `1`s (Rodrigo, 1996). This corresponds to our PDS model which is not affected by the positively misleading tree shape bias.



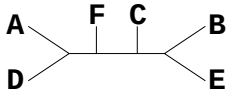
Chapter 4

Conclusions and Outlook

This thesis presents the two aspects of phylogenetic postprocessing. First, distance computations aid in understanding the amount of difference between two phylogenetic trees. In most applications, only topological information is used for the distance. To overcome this, we presented an algorithm to compute a distance measure incorporating the tree topologies and branch lengths, which is motivated by a mathematical representation of the tree space (see Section 1.3 for the tree space and Section 2.2 for the geodesic distance). Furthermore, tree distances can be used to quantify the amount of difference between two trees in statistical terms. Caution is necessary, however, to account for the discrete topology space correctly (Section 2.3).

Distances between trees can also be used to compute the median tree for a set of trees. This definition gave rise to the prevalent majority-rule (MR) consensus method. For trees with overlapping taxon sets, different specifications can be used, namely MR(-)supertrees and MR(+)supertrees. We presented algorithms to compute the respective distances in the matrix representation framework often used for supertree methods (Section 3.4). This allowed for the implementations of these algorithms together with a heuristic tree search strategy. When comparing the two specifications in simulations, we observed a clearly better performance of MR(-) compared to MR(+). This discrepancy is likely to trace back to a tree shape bias in MR(+)supertrees. The null model of tree topologies can insert a tree shape bias in split-based supertree methods (Section 3.5). Only the distribution of equally likely splits behaves in an appropriate way if little information is present. In contrast, the distribution of equally likely trees implies shape-specific effects due to the unequal split distribution.

Additionally to studying those theoretical aspects of supertree methods, we also compared data combination methods using simulations (Section 3.3). There, we in-

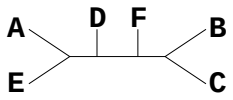


vestigated the performance of these methods in different settings, e. g. with complete or overlapping taxon sets, with equal or different substitution parameters or even with different gene topologies. The results show a good performance of matrix representation methods compared to other supertree and medium-level methods. Furthermore, superalignment is well applicable in the case of differing parameters between genes. We see a negative influence of alignment length on the performance of superalignment if the gene topologies differ considerably.

This thesis provides detailed investigations of methods and programs related to phylogenetic postprocessing. These investigations encompass algorithmic and theoretical work as well as practical considerations. The algorithmic work covers mainly the geodesic distance and MR(+)supertrees. Both were only defined mathematically before, and are now available for computations. Furthermore, we did theoretical work concerning tree distributions. The discreteness of the space of topologies must be considered adequately when trees are tested statistically. Another aspect are the consequences of the non-uniform split distributions on the tree shape bias of split-based supertree methods. Finally, the practical considerations cover a simulation study for data combination methods on different levels. The practical aspect is always taken into account throughout this thesis. The implementation of the geodesic distance is compared to its approximations and the new majority-rule supertree implementations are incorporated into the simulation study.

As every scientific work, this thesis does not only contribute answers to the scientific community, but also points out open questions. The first question is about the elementary information in a tree. Comparing two trees necessarily involves breaking up those trees. Here, mostly splits were considered as the elementary information. However, we see that splits come in different *k*inds and thus do not provide unbiased information. It is to date unknown whether these biases also exist when using other types of elementary information, for example quartets or rooted triplets. Furthermore, it is not clear whether branch lengths should be considered as informative in tree distance computations. Our simulation study raises a related question, namely, what is the elementary evolutionary information. Superalignment weights each alignment character equally and thus considers the base as an element. Supertrees, however, weight each gene equally. The latter approach is more reasonable in the presence of gene-specific evolution.

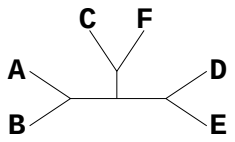
The second important question is how to distinguish between the variation in the

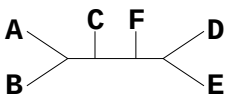


true gene trees and the variation due to model misspecification and stochastic processes. Supertrees only resolve the conflict in the gene trees without considering that distinction. Since "wrong" gene trees can bias the result, the filtering of these gene trees is desirable. However, the determination of a "wrong" gene tree without knowing the species tree is another complex, and yet largely uncharted, task.

Finally, the chapter on tree distances only considers trees on the same taxon set, whereas supertrees are designed for trees on overlapping taxon sets. The case of the majority-rule supertrees highlights that distances between a supertree and a gene tree can be defined in various ways. Our results show that the most intuitive way, namely completing the gene trees, can be misleading. So far, pruning the supertree to the taxa in the gene tree is the only solution for a distance computation between two trees of different sizes. However, the scales of most distances depend on the sizes of the taxon sets. An interesting topic is to design reasonable distances applicable to trees on overlapping taxon sets. Such a distance would help to evaluate the performance of a supertree method, that is, which genes are most similar to the final supertree. Supertree methods could be compared based on this evaluation. In simulations, we applied the baseline distance as an indicator for method performances, that is the average distance of a gene tree to the model tree. In real analyses, however, the model tree is unknown. Then only distances among gene trees or between a gene tree and a computed supertree are applicable. Furthermore, reconstructing the median tree based on this distance could serve as a supertree method.

To conclude, the presented algorithms and results contribute to our understanding of phylogenetic trees, how to extract information from trees, compute distances between trees and combine trees.




A
B
C
F
D
E

Bibliography

Aho, Alfred V., Yehoshua Sagiv, Thomas G. Szymanski, and Jeffrey D. Ullman (1981) Inferring a Tree from Lowest Common Ancestors with an Application to the Optimization of Relational Expressions. *SIAM J. Comput.*, **10**(3):405–421. (Cited on pages 39 and 45)

Amenta, Nina, Matthew Godwin, Nicolay Postarnakevich, and Katherine St. John (2007) Approximating Geodesic Tree Distance. *Inf. Process. Lett.*, **103**(2):61–65. (Cited on pages 24 and 25)

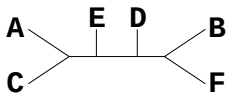
Baker, William J., Vincent Savolainen, Conny B. Asmussen-Lange, Mark W. Chase, John Dransfield, Felix Forest, Madeline M. Harley, Natalie W. Uhl, and Mark Wilkinson (2009) Complete Generic-Level Phylogenetic Analyses of Palms (Arecaceae) with Comparisons of Supertree and Supermatrix Approaches. *Syst. Biol.*, **58**(2):240–256. (Cited on page 40)

Barrett, Martin, Michael J. Donoghue, and Elliott Sober (1991) Against Consensus. *Syst. Zool.*, **40**(4):486–493. (Cited on page 38)

Barthélemy, Jean-Pierre and F. R. McMorris (1986) The Median Procedure for n-Trees. *J. Classif.*, **3**:329–334. (Cited on pages 29 and 68)

Baum, Bernard R. (1992) Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees. *Taxon*, **41**(1):3–10. (Cited on pages 38, 42, 43, 44, and 48)

Baum, Bernhard R. and Mark A. Ragan (2004) The MRP method. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 1, pp. 17–34. Kluwer Academic, Dordrecht, The Netherlands. (Cited on pages 43 and 44)



Billera, Louis J., Susan P. Holmes, and Karen Vogtmann (2001) Geometry of the Space of Phylogenetic Trees. *Adv. Appl. Math.*, **27**:733–767. (Cited on pages 9, 12, 13, 15, 16, and 29)

Bininda-Emonds, Olaf R. P. (2003) Novel Versus Unsupported Clades: Assessing the Qualitative Support for Clades in MRP Supertrees. *Syst. Biol.*, **52**(6):839–848. (Cited on page 40)

Bininda-Emonds, Olaf R. P., ed. (2004) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic, Dordrecht. (Cited on page 37)

Bininda-Emonds, Olaf R. P., Marcel Cardillo, Kate E. Jones, Ross D. E. MacPhee, Robin M. D. Beck, Richard Grenyer, Samantha A. Price, Rutger A. Vos, John L. Gittleman, and Andy Purvis (2007) The delayed rise of present-day mammals. *Nature*, **446**:507–512. (Cited on page 38)

Bininda-Emonds, Olaf R. P. and Michael J. Sanderson (2001) Assessment of the Accuracy of Matrix Representation with Parsimony Analysis Supertree Construction. *Syst. Biol.*, **50**(4):565–579. (Cited on pages 40 and 65)

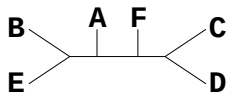
Blum, Michael G B and Olivier François (2006) Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance. *Syst. Biol.*, **55**(4):685–691. (Cited on page 35)

Bridson, Martin R. and André Haefliger (1999) *Metric Spaces of Non-Positive Curvature*. Springer, Berlin Heidelberg New York. (Cited on pages 15 and 19)

Bryant, David (2003) A Classification of Consensus Methods for Phylogenetics. In *Bioconsensus: Proceedings of Tutorial and Workshop on Bioconsensus II*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pp. 55–66. DIMACS-AMS. (Cited on pages 29 and 43)

Bryant, David, Andy McKenzie, and Mike Steel (2003) The Size of a Maximum Agreement Subtree for Random Binary Trees. *Dimacs Series in discrete mathematics and theoretical computer science*, **61**:55–65. (Cited on page 30)

Bull, J. J., John P. Huelsenbeck, Clifford W. Cunningham, David L. Swofford, and Peter J. Waddell (1993) Partitioning and Combining Data in Phylogenetic Analysis. *Syst. Biol.*, **42**(3):384–387. (Cited on page 38)



Burleigh, J. Gordon, Oliver Eulenstein, David Fernandez-Baca, and Michael J. Sanderson (2004) MRF supertrees. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 3, pp. 65–86. Kluwer Academic, Dordrecht, The Netherlands. (Cited on pages 42 and 44)

Camin, J. H. and R. R. Sokal (1965) A Method for Deducing Branching Sequences in Phylogeny. *Evolution*, **19**(3):311–326. (Cited on pages 3, 42, and 44)

Carstens, Bryan C and L. Lacey Knowles (2007) Estimating Species Phylogeny from Gene-Tree Probabilities Despite Incomplete Lineage Sorting: An Example from Melanoplus Grasshoppers. *Syst. Biol.*, **56**(3):400–411. (Cited on page 66)

Chen, Duhong, Lixia Diao, Oliver Eulenstein, David Fernández-Baca, and Michael J. Sanderson (2003) Flipping: A Supertree Construction Method. In M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds., *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 61, pp. 135–160. American Mathematical Society, Providence, Rhode Island. (Cited on pages 42 and 44)

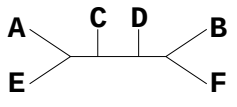
Chen, Duhong, Oliver Eulenstein, and David Fernández-Baca (2004) Rainbow: a toolbox for phylogenetic supertree construction and analysis. *Bioinformatics*, **20**(16):2872–2873. (Cited on page 45)

Chen, Duhong, Oliver Eulenstein, David Fernandez-Baca, and Michael Sanderson (2006) Minimum-Flip Supertrees: Complexity and Algorithms. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **3**(2):165–173. (Cited on page 44)

Chen, Feng, Aaron J Mackey, Jeroen K Vermunt, and David S Roos (2007) Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. *PLoS ONE*, **2**(4):e383. (Cited on page 3)

Ciccarelli, Francesca D., Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork (2006) Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, **311**:1283–1287. (Cited on pages 2 and 37)

Cotton, James A. and Mark Wilkinson (2007) Majority-Rule Supertrees. *Syst. Biol.*, **56**(3):445–452. (Cited on pages 65, 68, 71, and 76)



Creevey, Christopher J. and James O. McInerney (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*, **21**(3):390–392. (Cited on page 44)

Criscuolo, Alexis, Vincent Berry, Emmanuel J. P. Douzery, and Olivier Gascuel (2006) SDM: A Fast Distance-Based Approach for (Super)Tree Building in Phylogenomics. *Syst. Biol.*, **55**(5):740–755. (Cited on pages 37, 42, 46, and 65)

Daubin, Vincent, Manolo Gouy, and Guy Perrière (2002) A Phylogenomic Approach to Bacterial Phylogeny: Evidence of a Core of Genes Sharing a Common History. *Genome Res.*, **12**:1080–1090. (Cited on page 38)

Dong, Jianrong and David Fernández-Baca (2009) Properties of Majority-Rule Supertrees. *Syst. Biol.*, **58**(3):360–367. (Cited on pages 65, 68, and 76)

Dong, Jianrong, David Fernandez-Baca, and F. R. McMorris (2010) Constructing majority-rule supertrees. *Algorithms Mol Biol*, **5**(1):2. (Cited on pages 77 and 84)

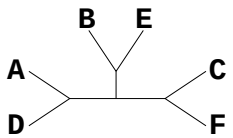
Driskell, Amy C., Cécile Ané, J. Gordon Burleigh, Michelle M. McMahon, Brian C. O'Meara, and Michael J. Sanderson (2004) Prospects for Building the Tree of Life from Large Sequence Databases. *Science*, **306**:1172–1174. (Cited on pages 2, 37, and 49)

Dutilh, B. E., V. van Noort, R. T. J. M. van der Heijden, T. Boekhout, B. Snel, and M. A. Huynen (2007) Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*, **23**(7):815–824. (Cited on pages 3, 13, 40, and 66)

Ebersberger, Ingo (2007) Human genetic ancestry. `http://www.newton.cam.ac.uk/webseminars/pg+ws/2007/plg/plgw01/0907/ebersberger/`. (Cited on pages 23 and 29)

Ebersberger, Ingo, Arndt von Haeseler, and Heiko A. Schmidt (2006) Phylogenetic Reconstruction. In Thomas Lengauer, ed., *Bioinformatics – From Genomes to Therapies*, vol. 1, pp. 83–128. Wiley-VCH Verlag, Weinheim, Germany, 2 edn. (Cited on page 37)

Edgar, Robert C and Serafim Batzoglou (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**(3):368–373. (Cited on page 3)



Edwards, Scott V (2009) Is a new and general theory of molecular systematics emerging? *Evolution*, **63**(1):1–19. (Cited on page 66)

Estabrook, G. F. and F. R. McMorris (1980) When is One Estimate of Evolutionary Relationships a Refinement of Another? *J. Math. Biol.*, **10**:367–373. (Cited on page 73)

Estabrook, George F., Fred R. McMorris, and Christopher A. Meacham (1985) Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Syst. Zool.*, **34**(2):193–200. (Cited on page 13)

Eulenstein, Oliver, Duhong Chen, J. Gordon Burleigh, David Fernández-Baca, and Michael J. Sanderson (2004) Performance of Flip Supertree Construction with a Heuristic Algorithm. *Syst. Biol.*, **53**(2):299–308. (Cited on pages 40 and 66)

Ewing, Gregory B, Ingo Ebersberger, Heiko A. Schmidt, and Arndt von Haeseler (2008) Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.*, **8**:118. (Cited on pages 23, 29, 57, 66, and 67)

Felsenstein, Joe (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. (Cited on page 7)

Felsenstein, Joseph (1981) Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *J. Mol. Evol.*, **17**:368–376. (Cited on page 3)

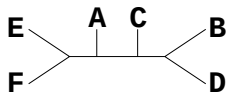
Felsenstein, Joseph (2005) *PHYLIP (Phylogeny Inference Package) version 3.6*. Department of Genome Sciences, University of Washington, Seattle. Distributed by the author. (Cited on pages 15 and 43)

Finden, C. R. and A. D. Gordon (1985) Obtaining Common Pruned Trees. *J. Classif.*, **2**:255–276. (Cited on page 14)

Fitch, Walter M. (1971) Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool.*, **20**(4):406–416. (Cited on page 44)

Fitch, Walter M. and Emanuel Margoliash (1967) Construction of Phylogenetic Trees. *Science*, **155**:279–284. (Cited on pages 4 and 46)

Fitzpatrick, David, Mary Logue, Jason Stajich, and Geraldine Butler (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.*, **6**(1):99. (Cited on page 40)



Gadagkar, Sudhindra R, Michael S Rosenberg, and Sudhir Kumar (2005) Inferring Species Phylogenies From Multiple Genes: Concatenated Sequence Tree versus Consensus Gene Tree. *J. Exp. Zool. B Mol. Dev. Evol.*, **304B**:64–74. (Cited on pages 13, 38, 47, and 67)

Gatesy, John, Richard H. Baker, and Cheryl Hayashi (2004) Inconsistencies in Arguments for the Supertree Approach: Supermatrices versus Supertrees of Crocodylia. *Syst. Biol.*, **53**(2):342–355. (Cited on pages 40, 47, and 48)

Gatesy, John and Mark S. Springer (2004) A Critique of Matrix Representation with Parsimony Supertrees. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 17, pp. 369–388. Kluwer Academic, Dordrecht, The Netherlands. (Cited on page 66)

Goddard, Wayne, Ewa Kubicka, Grzegorz Kubicki, and Fred R. McMorris (1994) The Agreement Metric for Labeled Binary Trees. *Math. Biosci.*, **123**:215–226. (Cited on page 33)

Goldman, Nick, Jon P. Anderson, and Allen G. Rodrigo (2000) Likelihood-Based Tests of Topologies in Phylogenetics. *Syst. Biol.*, **49**(4):652–670. (Cited on page 14)

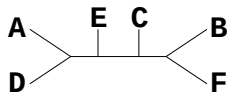
Goloboff, Pablo A. (2005) Minority rule supertrees? MRP, Compatibility, and Minimum Flip may display the least frequent groups. *Cladistics*, **21**(3):282–294. (Cited on pages 65 and 68)

Goloboff, Pablo A. and Diego Pol (2002) Semi-strict supertrees. *Cladistics*, **18**:514–525. (Cited on page 39)

Gordon, Allan D. (1986) Consensus Supertrees: The Synthesis of Rooted Trees Containing Overlapping Sets of Labelled Leaves. *J. Classif.*, **3**:335–348. (Cited on page 39)

Guindon, Stéphane and Olivier Gascuel (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.*, **52**(5):696–704. (Cited on page 23)

Hasegawa, Masami, Hirohisa Kishino, and Taka-Aki Yano (1985) Dating of the Human–Ape Splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol. Evol.*, **22**:160–174. (Cited on pages 3 and 41)



Hein, Jotun (1990) Reconstructing Evolution of Sequences Subject to Recombination Using Parsimony. *Math. Biosci.*, **98**:185–200. (Cited on page 13)

Hillis, David M., Tracy A. Heath, and Katherine St. John (2005) Analysis and Visualization of Tree Space. *Syst. Biol.*, **54**(3):471–482. (Cited on pages 13 and 29)

Ho, Simon Y. W. and Lars Jermiin (2004) Tracing the Decay of the Historical Signal in Biological Sequence Data. *Syst. Biol.*, **53**(4):623–637. (Cited on page 2)

Jones, David T., William R. Taylor, and Janet M. Thornton (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**(3):275–282. (Cited on pages 3 and 41)

Jukes, Thomas H. and Charles R. Cantor (1969) Evolution of protein molecules. In H. N. Munro, ed., *Mammalian protein metabolism*, vol. 3, pp. 21–132. Academic Press, New York. (Cited on page 3)

Kluge, Arnold G. (1989) A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates (Boidae, Serpentes). *Syst. Zool.*, **38**(1):7–25. (Cited on pages 37 and 42)

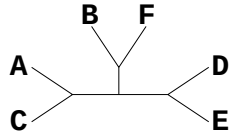
Kolaczkowski, Bryan and Joseph W. Thornton (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **432**:980–984. (Cited on page 67)

Kubatko, Laura Salter and James H. Degnan (2007) Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Syst. Biol.*, **56**(1):17–24. (Cited on page 66)

Kuhner, Mary K. and Joseph Felsenstein (1994) A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Mol. Biol. Evol.*, **11**(3):459–468. (Cited on pages 13 and 15)

Kupczok, Anne (2009) Consequences of different null models on the tree shape bias of supertree methods. Submitted to *Syst. Biol.* (Cited on page 77)

Kupczok, Anne and Arndt von Haeseler (2009) Comment on 'A congruence index for testing topological similarity between trees'. *Bioinformatics*, **25**(1):147–149. (Cited on page 30)



Kupczok, Anne, Arndt von Haeseler, and Steffen Klaere (2008) An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees. *J. Comput. Biol.*, **15**(6):577–591. (Cited on pages 4 and 16)

Kupczok, Anne, Heiko A. Schmidt, and Arndt von Haeseler (2009) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Submitted to *BMC Evol. Biol.* (Cited on page 47)

Lanave, Cecilia, Giuliano Preparata, Cecilia Saccone, and Gabriella Serio (1984) A New Method for Calculating Evolutionary Substitution Rates. *J. Mol. Evol.*, **20**:86–93. (Cited on page 3)

Landan, Giddy and Dan Graur (2007) Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. *Mol. Biol. Evol.*, **24**(6):1380–1383. (Cited on page 3)

Lapointe, François-Joseph and Guy Cucumel (1997) The Average Consensus Procedure: Combining of Weighted Trees Containing Identical or Overlapping Sets of Taxa. *Syst. Biol.*, **46**(2):306–312. (Cited on pages 37, 42, 46, and 65)

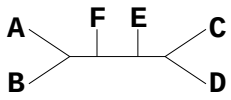
Lapointe, François-Joseph and Claudine Levasseur (2004) Everything you always wanted to know about average consensus and more. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 4, pp. 87–106. Kluwer Academic, Dordrecht, The Netherlands. (Cited on pages 42 and 46)

Lemey, Philippe, Marco Salemi, and Anne-Mieke Vandamme, eds. (2009) *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press. (Cited on page 4)

Levasseur, Claudine and François-Joseph Lapointe (2006) Total Evidence, Average Consensus and Matrix Representation with Parsimony: What a Difference Distances Make. *Evol. Bioinform. Online*, **2**:249–253. (Cited on page 40)

Lin, Harris T, J. Gordon Burleigh, and Oliver Eulenstein (2009) Triplet supertree heuristics for the tree of life. *BMC Bioinformatics*, **10**(Suppl 1):S8. (Cited on page 65)

Liu, Liang (2008) BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, **24**(21):2542–2543. (Cited on page 67)



Liu, Liang, Lili Yu, Laura Kubatko, Dennis K. Pearl, and Scott V. Edwards (2009) Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.*, **53**:320–328. (Cited on page 66)

Maddison, Wayne P. (1997) Gene Trees in Species Trees. *Syst. Biol.*, **46**(3):523–536. (Cited on page 2)

Maddison, Wayne P. and Montgomery Slatkin (1991) Null Models for the Number of Evolutionary Steps in a Character on a Phylogenetic Tree. *Evolution*, **45**(5):1184–1197. (Cited on page 85)

Margush, Tim and Fred R. McMorris (1981) Consensus n-trees. *Bull. Math. Biol.*, **43**:239–244. (Cited on page 42)

McMahon, M.M. and M.J. Sanderson (2006) Phylogenetic Supermatrix Analysis of GenBank Sequences from 2228 Papilionoid Legumes. *Syst. Biol.*, **55**(5):818–36. (Cited on pages 2 and 37)

McMorris, F. R. (1977) On the compatibility of binary qualitative taxonomic characters. *Bull. Math. Biol.*, **39**(2):133–138. (Cited on page 73)

Mossel, Elchanan and Eric Vigoda (2005) Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees. *Science*, **309**:2207–2209. (Cited on page 67)

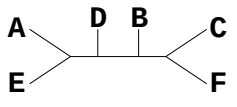
Notredame, C, D G Higgins, and J Heringa (2000) T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *J. Mol. Biol.*, **302**:205–17. (Cited on page 23)

O'Brien, Kevin P., Maido Remm, and Erik L. L. Sonnhammer (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucl. Acids Res.*, **33**(suppl_1):D476–480. (Cited on page 23)

Owen, Megan (2008) *Distance Computation in the Space of Phylogenetic Trees*. Ph.D. thesis, Cornell University. (Cited on page 27)

Owen, Megan and J. Scott Provan (2010) A Fast Algorithm for Computing Geodesic Distances in Tree Space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14 Jan. 2010. (Cited on page 27)

Page, Roderic D. M. (1996) On consensus, confidence, and "total evidence". *Cladistics*, **12**:83–92. (Cited on page 38)



Page, Roderic D. M. (2002) Modified Mincut Supertrees. In *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI 2002)*, vol. 2452 of *Lecture Notes in Computer Science*, pp. 537–551. Springer, New York. (Cited on pages 42 and 45)

Pattengale, Nicholas Dylan (2005) *Tools for Phylogenetic Postprocessing*. Master's thesis, University of New Mexico. (Cited on page 29)

Philippe, Hervé, Elizabeth A. Snell, Eric Bapteste, Philippe Lopez, Peter W. H. Holland, and Didier Casane (2004) Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments. *Mol. Biol. Evol.*, **21**(9):1740–1752. (Cited on page 37)

Philippe, Hervé and Maximilian J Telford (2006) Large-scale sequencing and the new animal phylogeny. *Trends Ecol. Evol. (Amst.)*, **21**(11):614–620. (Cited on page 3)

Phillips, Matthew J., Frédéric Delsuc, and David Penny (2004) Genome-Scale Phylogeny and the Detection of Systematic Biases. *Mol. Biol. Evol.*, **21**(7):1455–1458. (Cited on page 2)

Piaggio-Talice, Raul, Gordon Burleigh, and Oliver Eulenstein (2004) Quartet Supertrees. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 8, pp. 173–191. Kluwer Academic, Dordrecht. (Cited on pages 42, 45, and 65)

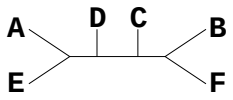
Pickett, Kurt M. and Christopher P. Randle (2005) Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. *Mol. Phylogenet. Evol.*, **34**:203–211. (Cited on page 85)

Purvis, Andy (1995) A Modification to Baum and Ragan's Method for Combining Phylogenetic Trees. *Syst. Biol.*, **44**:251–255. (Cited on pages 42 and 43)

de Queiroz, Alan, Michael J. Donoghue, and Junhyong Kim (1995) Separate Versus Combined Analysis of Phylogenetic Evidence. *Annu. Rev. Ecol. Syst.*, **26**:657–681. (Cited on page 38)

de Queiroz, Alan and John Gatesy (2007) The supermatrix approach to systematics. *Trends Ecol. Evol. (Amst.)*, **22**(1):34–41. (Cited on page 37)

Ragan, Mark A. (1992) Phylogenetic Inference Based on Matrix Representation of Trees. *Mol. Phylogenet. Evol.*, **1**(1):53–58. (Cited on pages 38, 42, 43, and 44)



Rambaut, Andrew and Nick C. Grassly (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**(3):235–238. (Cited on page 50)

Ranwez, Vincent, Vincent Berry, Alexis Criscuolo, Pierre-Henri Fabre, Sylvain Guillemot, Celine Scornavacca, and Emmanuel J P Douzery (2007) PhySIC: A Veto Supertree Method with Desirable Properties. *Syst. Biol.*, **56**(5):798–817. (Cited on page 39)

Robinson, David F. and Leslie R. Foulds (1978) Comparison of weighted labelled trees. In A.F. Horadam and W.D. Wallis, eds., *Combinatorial Mathematics VI. Proceedings of the Sixth Australian Conference on Combinatorial Mathematics, Armidale, Australia*, vol. 748 of *Lecture notes in mathematics*, pp. 119–126. Springer Verlag, Berlin. (Cited on pages 13 and 14)

Robinson, David F. and Leslie R. Foulds (1981) Comparison of Phylogenetic Trees. *Math. Biosci.*, **53**:131–147. (Cited on pages 13 and 14)

Rodrigo, Allen G. (1996) On Combining Cladograms. *Taxon*, **45**:267–274. (Cited on pages 42, 44, and 86)

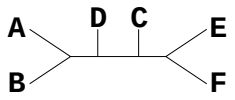
Rodríguez-Ezpeleta, Naiara, Henner Brinkmann, Batrice Roure, Nicolas Lartillot, B. Franz Lang, and Herv Philippe (2007) Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies. *Syst. Biol.*, **56**(3):389–399. (Cited on page 2)

Rokas, Antonis (2006) Genomics and the Tree of Life. *Science*, **313**:1897–1899. (Cited on page 1)

Rokas, Antonis and Sean B. Carroll (2005) More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. *Mol. Biol. Evol.*, **22**(5):1337–1344. (Cited on page 13)

Rokas, Antonis, Barry L Williams, Nicole King, and Sean B Carroll (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**:798–804. (Cited on page 2)

Ross, Howard A. and Allen G. Rodrigo (2004) An assessment of matrix representation with compatibility in supertree reconstruction. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 2,



pp. 35–63. Kluwer Academic, Dordrecht, The Netherlands. (Cited on pages 42 and 44)

Saitou, Naruya and Masatoshi Nei (1987) The Neighbor–joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.*, **4**(4):406–425. (Cited on page 4)

Salamin, Nicolas, Trevor R. Hodkinson, and Vincent Savolainen (2002) Building Supertrees: An Empirical Assessment Using the Grass Family (Poaceae). *Syst. Biol.*, **51**(1):136–150. (Cited on pages 39 and 65)

Salamin, Nicolas, Trevor R. Hodkinson, and Vincent Savolainen Coates (2005) Towards Building the Tree of Life: A Simulation Study for All Angiosperm Genera. *Syst. Biol.*, **54**(2):183–196. (Cited on page 47)

Sanderson, Michael J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**(2):301–302. (Cited on page 43)

Schmidt, Heiko A. (2003) *Phylogenetic Trees from Large Datasets*. Ph.D. thesis, Universität Düsseldorf. (Cited on pages 2, 37, 42, and 46)

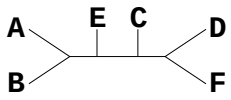
Schmidt, Heiko A. (2009) Testing Tree Topologies. In Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme, eds., *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, chap. 12, pp. 377–400. Cambridge University Press. (Cited on page 14)

Schmidt, Heiko A., Korbinian Strimmer, Martin Vingron, and Arndt von Haeseler (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**(3):502–504. (Cited on page 46)

Semple, Charles and Mike Steel (2000) A supertree method for rooted trees. *Discr. Appl. Math.*, **105**:147–158. (Cited on pages 42, 45, and 65)

Semple, Charles and Mike Steel (2002) A characterization for a set of partial partitions to define an X-tree. *Discrete Math.*, **247**:169–186. (Cited on page 73)

Semple, Charles and Mike Steel (2003) *Phylogenetics*, vol. 24 of *Oxford Lecture Series in Mathematics and Its Applications*. Oxford University Press, Oxford, UK. (Cited on page 7)



Smythe, Ashleigh B., Michael J. Sanderson, and Steven A. Nadler (2006) Nematode Small Subunit Phylogeny Correlates with Alignment Parameters. *Syst. Biol.*, **55**(6):972–992. (Cited on page 13)

Snir, Sagi and Satish Rao (2006) Using Max Cut to Enhance Rooted Trees Consistency. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**(4):323–333. (Cited on pages 42 and 45)

Steel, Mike (1992) The Complexity of Reconstructing Trees from Qualitative Characters and Subtrees. *J. Classif.*, **9**:91–116. (Cited on page 39)

Steel, Mike (1993) Distributions on bicoloured binary trees arising from the principle of parsimony. *Discr. Appl. Math.*, **41**:245–261. (Cited on page 79)

Steel, Mike, Andreas W. M. Dress, and Sebastian Böcker (2000) Simple but Fundamental Limitations on Supertree and Consensus Tree Methods. *Syst. Biol.*, **42**:363–368. (Cited on page 39)

Steel, Mike and Kurt M. Pickett (2006) On the impossibility of uniform priors on clades. *Mol. Phylogenet. Evol.*, **39**:585–586. (Cited on pages 8 and 85)

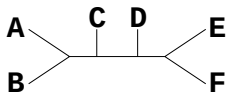
Stockham, Cara, Li-San Wang, and Tandy Warnow (2002) Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics*, **18**(suppl_1):S285–293. (Cited on page 13)

Strimmer, Korbinian and Arndt von Haeseler (1996) Quartet Puzzling: A Quartet Maximum–Likelihood Method for Reconstructing Tree Topologies. *Mol. Biol. Evol.*, **13**(7):964–969. (Cited on page 46)

Swenson, M. Shel, Francois Barbancon, Tandy Warnow, and C. Randal Linder (2010) A simulation study comparing supertree and combined analysis methods using SMIDGen. *Algorithms Mol Biol*, **5**(1):8. (Cited on page 67)

Swofford, David L. (2002) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts. (Cited on pages 33, 41, and 78)

de Vienne, Damien M, Tatiana Giraud, and Olivier C Martin (2007) A congruence index for testing topological similarity between trees. *Bioinformatics*, **23**(23):3119–3124. (Cited on pages 30, 31, and 35)



de Vienne, Damien M., Tatiana Giraud, and Olivier C. Martin (2009) In response to comment on 'A congruence index for testing topological similarity between trees'. *Bioinformatics*, **25**(1):150–151. (Cited on page 30)

Vinh, Le Sy and Arndt von Haeseler (2004) IQPNNI: Moving Fast Through Tree Space and Stopping in Time. *Mol. Biol. Evol.*, **21**(8):1565–1571. (Cited on page 41)

Vogtmann, Karen (2003) Geodesics in the Space of Trees. *Tech. rep.*, Cornell University. (Cited on pages 16, 17, and 19)

Waterman, M. S. and T. F. Smith (1978) On the Similarity of Dendrograms. *J. Theor. Biol.*, **73**:789–800. (Cited on page 13)

Wilkinson, Mark, James A. Cotton, Chris Creevey, Oliver Eulenstein, Simon R. Harris, François-Joseph Lapointe, Claudine Levasseur, James O. McInerney, Davide Pisani, and Joseph L. Thorley (2005a) The Shape of Supertrees to Come: Tree Shape Related Properties of Fourteen Supertree Methods. *Syst. Biol.*, **54**(3):419–431. (Cited on pages 77, 84, and 85)

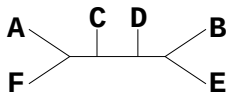
Wilkinson, Mark, James A. Cotton, François-Joseph Lapointe, and Davide Pisani (2007) Properties of Supertree Methods in the Consensus Setting. *Syst. Biol.*, **56**(2):330–337. (Cited on pages 39, 77, and 78)

Wilkinson, Mark, Davide Pisani, James A. Cotton, and Ian Corfe (2005b) Measuring Support and Finding Unsupported Relationships in Supertrees. *Syst. Biol.*, **54**(5):823–831. (Cited on page 40)

Wilkinson, Mark, Joseph L. Thorley, Davide Pisani, François-Joseph Lapointe, and James O. McInerney (2004) Some desiderata for liberal supertrees. In Olaf R. P. Bininda-Emonds, ed., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, chap. 10, pp. 227–246. Kluwer Academic, Dordrecht, The Netherlands. (Cited on pages 39 and 77)

Willson, Stephen J. (1999) Building Phylogenetic Trees from Quartets by Using Local Inconsistency Measures. *Mol. Biol. Evol.*, **16**(5):685–693. (Cited on page 45)

Wortley, Alexandra H and Robert W Scotland (2006) The Effect of Combining Molecular and Morphological Data in Published Phylogenetic Analyses. *Syst. Biol.*, **55**(4):677–685. (Cited on page 2)



Appendix A

Simulation Results

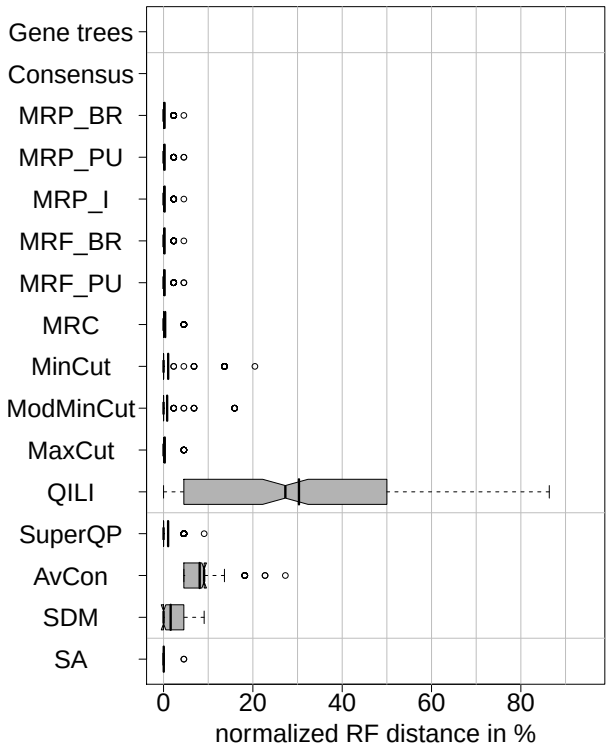
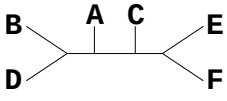


Figure A.1: Distribution of normalized RF distances (200 simulations) for the simulation setting with long sequences (S,m,E,l). Only missing data is studied, thus the baseline distance is not applicable.



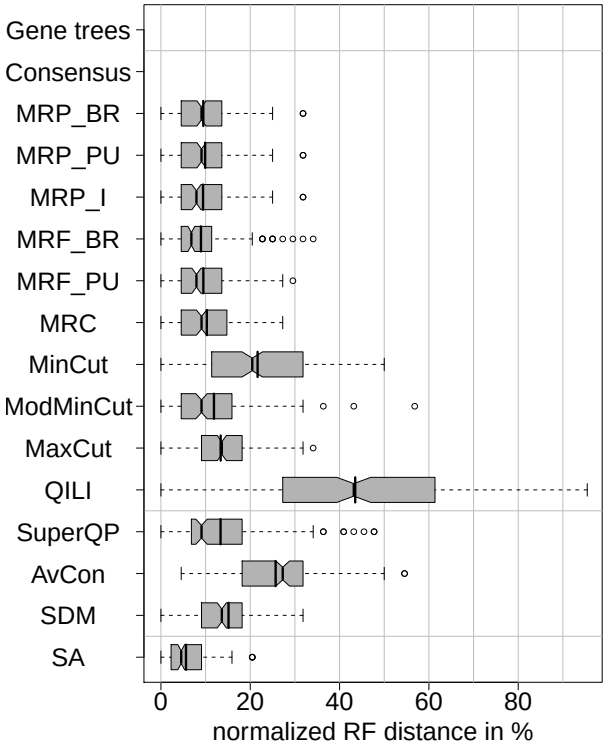


Figure A.2: Distribution of normalized *RF* distances for the simulation with bootstrapping (200 simulations). Alignments were taken from the data sets that were the basis of the results shown in Figure 3.4b (S,m,E,n). Only missing data is studied, thus the baseline distance is not applicable.

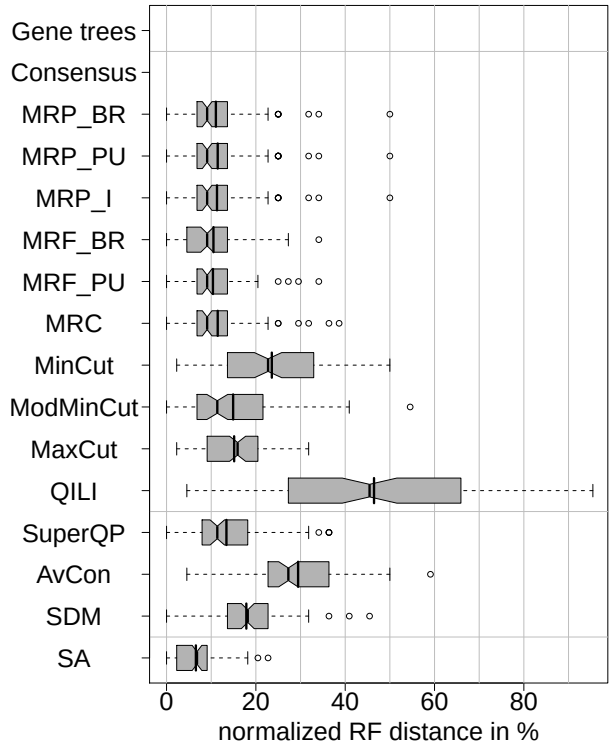


Figure A.3: Distribution of normalized *RF* distances (100 simulations) for the simulation setting S,m,R_3,n.

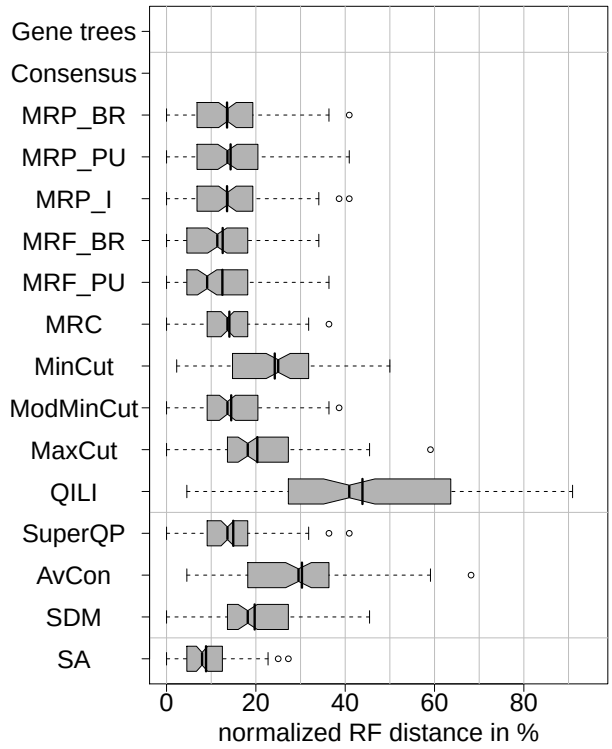
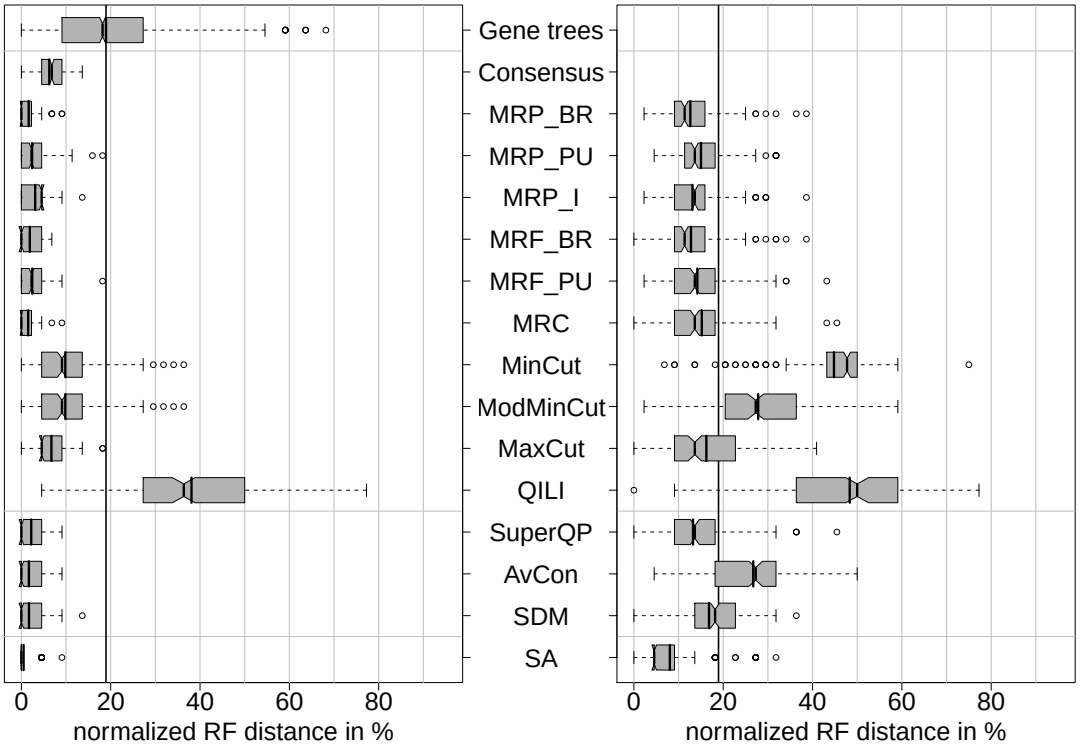
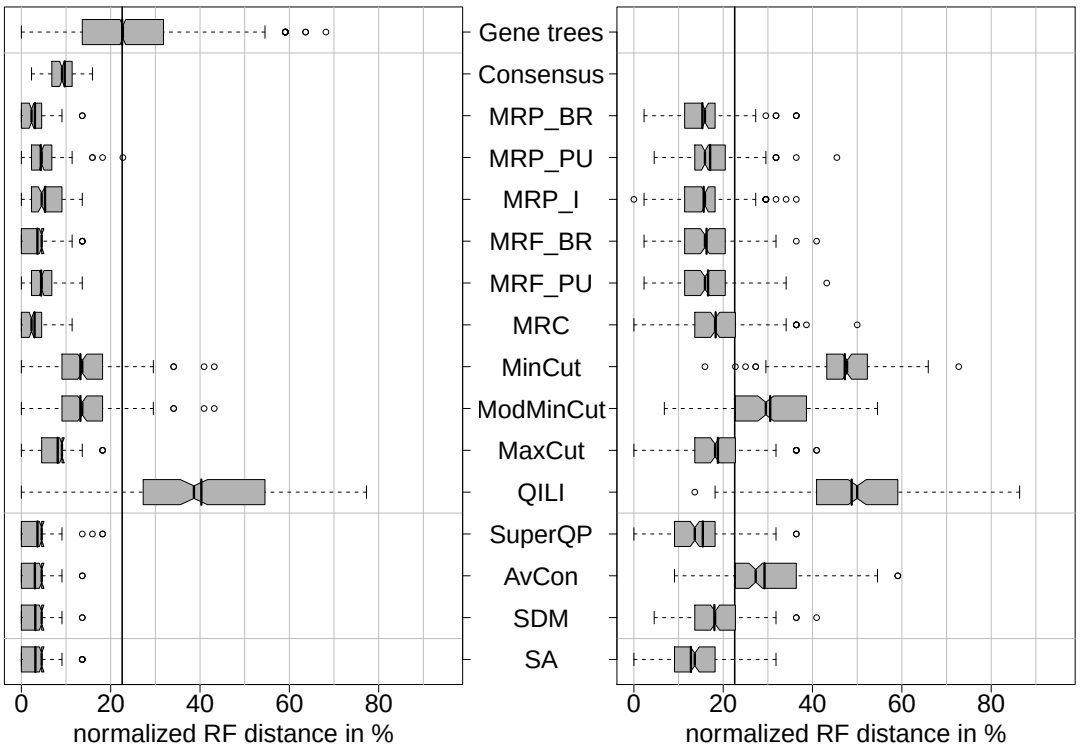


Figure A.4: Distribution of normalized *RF* distances (100 simulations) for the simulation setting S,m,$R_{1.67}$,n.



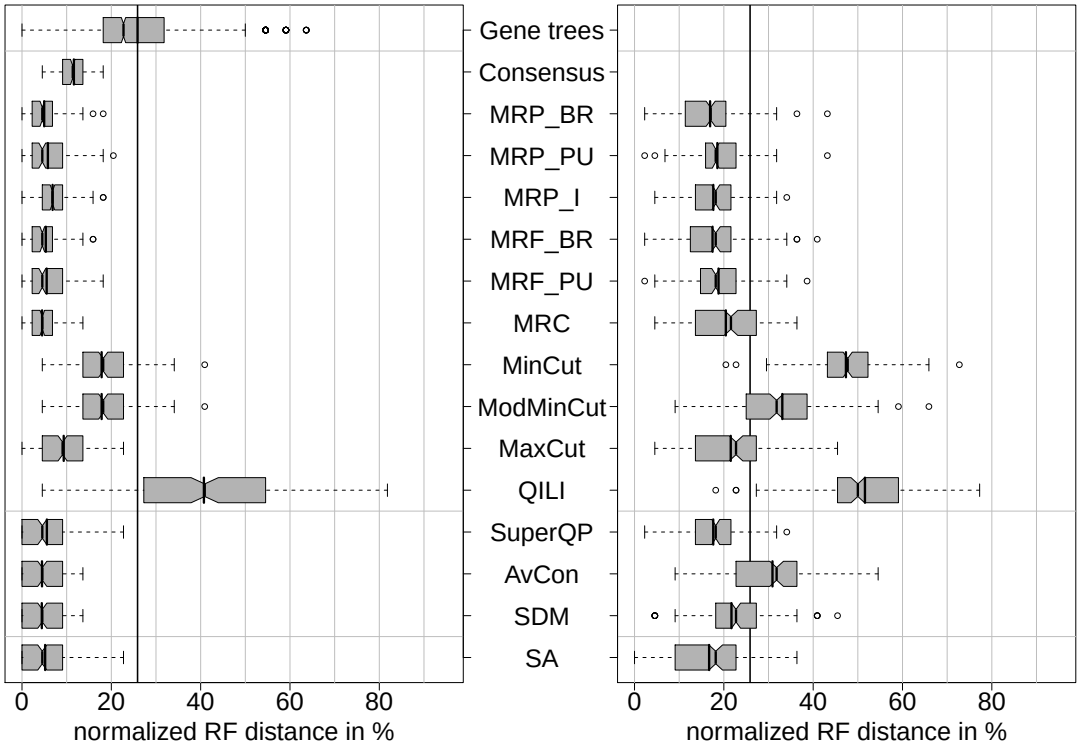
(a) S,c,$T_{0.001}$,n **(b)** S,m,$T_{0.001}$,n

Figure A.5: Distribution of normalized *RF* distances (200 simulations).



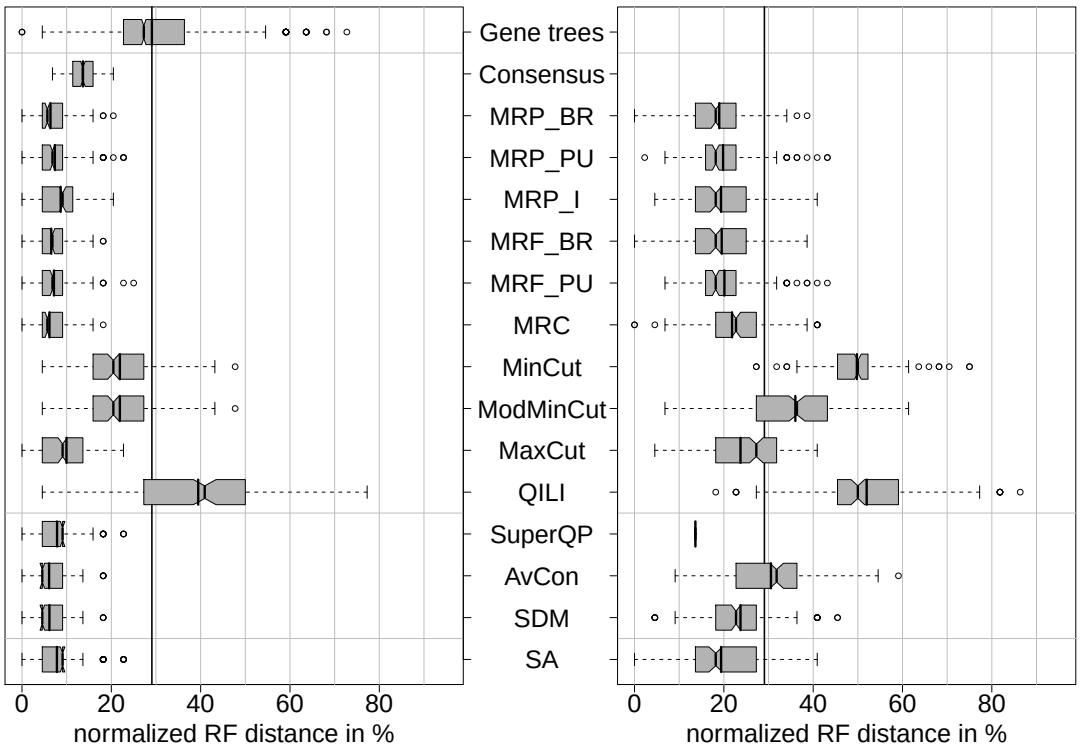
(a) S,c,$T_{0.002}$,n **(b)** S,m,$T_{0.002}$,n

Figure A.6: Distribution of normalized *RF* distances (200 simulations).



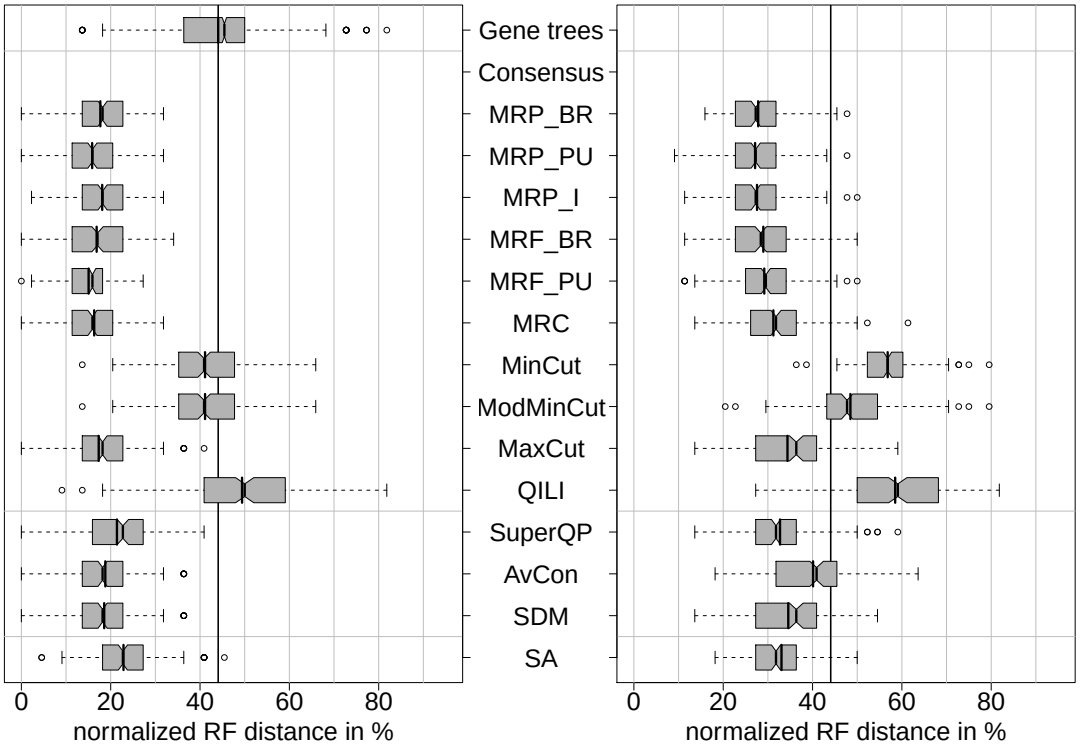
(a) S,c,$T_{0.003}$,n **(b)** S,m,$T_{0.003}$,n

Figure A.7: Distribution of normalized *RF* distances (200 simulations).



(a) S,c,$T_{0.004}$,n **(b)** S,m,$T_{0.004}$,n

Figure A.8: Distribution of normalized *RF* distances (200 simulations).



(a) S,c,$T_{0.01}$,n **(b)** S,m,$T_{0.01}$,n

Figure A.9: Distribution of normalized *RF* distances (200 simulations).

Appendix B

List of Abbreviations and Symbols

$\mathcal{A}$	A legal topology
$\mathbb{A}$	A set of legal topologies
α	Size of a statistical test
AvCon	Average Consensus
BR	Baum-Ragan (coding)
BS	Branch-score distance
CL	Compatibility length
d	Dimension: number of splits unique to one topology compared to another
d^+	Objective function for MR(+)supertrees
d^-	Objective function for MR(-)supertrees
DAG	Directed acyclic graph
$\mathbf{e}^{\mathcal{S}}$	Orthogonal unit vector corresponding to split $\mathcal{S}$
E_t	The extension of a partial split t
g	A piecewise linear function
`GeoMeTree`	Geodesic Metric on Trees
GTR	General time-reversible
HKY	Hasegawa, Kishino and Yano
I	A transition
JC	Jukes-Cantor
JTT	Jones, Taylor and Thornton
k-split	A split into k and $n-k$ taxa
l	Number of columns in an MRep
$\lambda_{\mathbf{p}}$	Weight function of tree $\mathbf{p}$
m	Number of possible splits for n taxa

MAST	Maximum agreement subtree
MD	MAST-based distance
MinCut	Minimal cut
ML	Maximum likelihood
ModMinCut	Modified minimal cut
MP	Maximum parsimony
MR	Majority-rule
MRC	Matrix representation with compatibility
MRep	Matrix representation
MRF	Matrix representation with flipping
MRP	Matrix representation with parsimony
n	Number of taxa
NNI	Nearest-neighbor interchange
$\mathbf{p}$	A weighted tree
$\mathcal{P}$	A supertree toplogy
PL	Parsimony length
`PluMiST`	Plus- and Minus Supertrees
PDA	Proportional to distinguishable arrangements
PDS	Proportional to distinguishable splits
PU	Purvis (coding)
QILI	Quartet inference and local inconsistency
QP	Quartet puzzling
RF	Robinson-Foulds distance
RF_w	Weighted Robinson-Foulds distance
$\mathcal{S}$	A split
$\mathbb{S}'$	Splits in the symmetric difference of two trees
SA	Superalignment
SDM	Super distance matrix
SES	Split extension score
$\mathbb{S}_n$	Set of splits for n taxa
SuperQP	SuperQuartetPuzzling
$\mathcal{T}$	A topology
$< \mathcal{T} >$	The span of a topology
θ	Coalescent parameter
TBR	Tree-bisection-reconnection

T_n	Number of topologies for n taxa
$\mathbb{T}_n$	Space of weighted trees for n taxa
X	A taxon set
$X_1 \vert X_2$	A split

Curriculum Vitae

Contact Information

Anne Kupczok
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories
Dr. Bohr Gasse 9
1030 Vienna, Austria
Phone: +43 1 / 4277 24027
Fax: +43 1 / 4277 24098
Email: anne.kupczok@univie.ac.at

Date of birth October 12, 1982

Place of birth Leipzig, Germany

Nationality German

Education

July 2006 Diploma in Bioinformatics

October 2001 - July 2006 Student of Bioinformatics at the Friedrich Schiller University, Jena

June 2001 "Abitur" (German A-Levels) focused on biology and mathematics

1993 - 2001 Student at the secondary school "Gymnasium Erasmus Reinhold", Saalfeld

Research Experience

November 2006 - January 2010 PhD Student at the CIBIV (Center for Integrative Bioinformatics Vienna)
Supervisor: Prof. Arndt von Haeseler
Topic: Postprocessing Phylogenies: Tree Distances and Supertrees

July 2006 - September 2006 Scholarship at the Max-Planck-Institute for Chemical Ecology, Jena, to continue work of diploma thesis

January 2006 - July 2006 Diploma student at the Max-Planck-Institute for Chemical Ecology, Jena
Supervisors: Dr. Karl Schmid, Dr. Jürgen Sühnel
Topic: Evolutionary Genomics of the Plant-pathogenic Bacteria Xanthomonas

March 2003 - April 2005 Research assistant in the Biosystems Analysis Group, Jena
Supervisor: Dr. Peter Dittrich
Topic: Dynamics of RNA Evolution

Publications

1. Anne Kupczok and Arndt von Haeseler (2009) Comment on 'A congruence index for testing topological similarity between trees'. *Bioinformatics*, **25**(1):147–149.
2. Anne Kupczok, Arndt von Haeseler, and Steffen Klaere (2008) An Exact Algorithm for the Geodesic Distance between Phylogenetic Trees. *J. Comput. Biol.*, **15**(6):577–591.
3. Anne Kupczok and Peter Dittrich (2006) Determinants of Simulated RNA Evolution. *J. Theor. Biol.*, **238**(3):726-735.