A phylogenetic potpourri

Computational methods for analysing genome-scale data

Dissertation

der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. Alexander F. Auch
aus Stuttgart

Tübingen
2009

Tag der mündlichen Qualifikation:	13.01.2010
Dekan:	Prof. Dr. Oliver Kohlbacher
1. Berichterstatter:	Prof. Dr. Daniel H. Huson
2. Berichterstatter:	Dr. Alexandros Stamatakis The Exelixis Lab Technische Universität München

Erklärung

Hiermit erkläre ich, daß ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und daß alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind. Eine detaillierte Abgrenzung meiner eigenen Leistungen von den Beiträgen meiner Kooperationspartner und von Implementierungsleistungen, die im Rahmen von mir betreuter Studien- und Diplomarbeiten erbracht worden sind, habe ich explizit in Anhang B vorgenommen.

Tübingen, September 2009 *Alexander Auch*

Zusammenfassung

Seit den Anfängen der Erforschung evolutionärer Prozesse gilt das Bestreben dieser Disziplin der Rekonstruktion eines möglichst wirklichkeitsgetreuen Stammbaums des Lebens. Dieser Zweig der Wissenschaft wird nach Ernst Haeckel als “Phylogenetik” bezeichnet – die Entwicklungsgeschichte der Stämme. Die ersten phylogenetischen Methoden benutzten morphologische Merkmale zur Unterscheidung von Arten, um daraus einen Stammbaum des Lebens zu erstellen. Allerdings ist diese Methodik nur beschränkt auf Mikroorganismen anwendbar, da diese nur wenige gut zu unterscheidende morphologische Merkmale besitzen. Erst die Entschlüsselung der DNA-Struktur durch Francis Crick und James Watson, sowie die Entwicklung der Sanger-Sequenziertechnologie ermöglichten es, genetische Informationen zur phylogenetischen Rekonstruktion heranzuziehen.

Noch unbeantwortet ist hingegen die Frage nach der tatsächlichen Existenz eines prokaryotischen Baums des Lebens. Prokaryoten (Bakterien und Archaea) besitzen Mechanismen für den direkten Austausch von genetischem Material zwischen Zellen, die zu verschiedenen Arten gehören können (horizontaler Gentransfer). Dies bedeutet, daß ein Gen auch durch andere Wege als die klonale Vermehrung erhalten werden kann, die eben nicht durch eine Baumstruktur repräsentiert werden können. In dieser Dissertation stellen wir die GBDP-Methodik (”Genome BLAST distance phylogeny”) vor, mit der Phylogenien aus ganzen Genomen berechnet werden können. Die Ergebnisse der GBDP-Methodik werden mit einer Taxonomie verglichen, die auf der Phylogenie von Einzelgenen basiert. Des weiteren untersuchen wir den Anteil von horizontalem Gentransfer in einer Gruppe von Genen, die in allen von uns untersuchten prokaryotischen Genomen vorkommen. Für diese Untersuchung benutzen wir sowohl eine aktuelle Methode, wie zwei von uns neu vorgestellte Ansätze. Zusätzlich schlagen wir hier eine neue Methode zur Spezies-Bestimmung bei Prokaryoten vor, die auf der GBDP-Methodik basiert.

Im letzten Teil der Dissertation werden mehrere Software-Pakete vorgestellt. Zusammen mit AxParafit und AxPcoords stellt CopyCat das erste Grid-fähige Software-Paket dar, das speziell im Hinblick auf großangelegte kophylogenetische Analysen entwickelt wurde. Mit diesen Programmen können große Wirts- und Parasitenphylogenien miteinander auf Über-

einstimmungen hin untersucht werden. Des weiteren wird MEGAN vorgestellt, eine benutzerfreundliche Software-Applikation für die Analyse von Metagenomik-Datensätzen, sowie MetaSim, ein Simulationsprogramm für Metagenomik-Datensätze, das zur Unterstützung der Entwicklung und Verifikation von Metagenomik-Software entwickelt wurde.

Abstract

Since the dawn of evolutionary biology, it was the dream of scientists to obtain a meaningful genealogy of species, a "tree of life". The term "phylogenetics" was coined by Ernst Haeckel for that area of research, meaning the history of the evolutionary relationships between species. First phylogenetic approaches focused on morphological differences between species. However, the analysis of the phylogeny of microbial organisms is hindered due to the limited number of observable morphological differences. With the discovery of the structure of DNA by Francis Crick and James Watson, and the development of the Sanger sequencing technology, it became feasible to use genetic information for phylogenetic inference.

Regarding the prokaryotic universe (Bacteria and Archaea), a main question of phylogenetics is whether there exists a prokaryotic "tree of life" actually. Those organisms exhibit mechanisms for the direct exchange of genetic material between cells that can belong to different species (called horizontal gene transfer). Accordingly, genes can be derived from different organisms rather than via clonal reproduction, as expressed by a phylogenetic tree. In this thesis, we introduce the GBDP ("Genome BLAST distance phylogeny") framework for inferring phylogenies based on whole genomes, and we compare the results with a current taxonomic tree based on single genes. Furthermore, we investigate the amount of horizontal gene transfer in a common set of prokaryotic genes by using a state-of-the-art method, as well as two newly developed approaches. Additionally, a new method for species delineation is proposed that is based on the GBDP method for deriving whole genome phylogenies.

In the last part of the thesis, several software packages are presented. CopyCat, together with AxParafit and AxPcoords, represents the first Grid-enabled software package that is optimized for large-scale cophylogenetic studies. With these tools, large host and parasite phylogenies can be screened for correlations. Furthermore, MEGAN, a user-friendly software application for the analysis of metagenomic datasets is presented. Metagenomics is the study of microorganismal communities by direct extraction of DNA from environmental samples. To aid the development and testing of metagenomic software, we developed MetaSim, a tool to generate simulated metagenomic datasets.

Acknowledgements

I have to thank my advisor Daniel Huson for providing me the opportunity to work in his department during my PhD. Also, I want to thank my co-advisor Alexandros Stamatakis for his kind support and many constructive suggestions which helped to improve my work.

Amongst others, I am deeply grateful to Markus Göker for many helpful comments, ideas, and for sharing his profound biological knowledge with me. Whenever I had to struggle with a scientific problem, he took the time to listen to me – and in most cases :-), he also had the right idea. Tremendous thanks go to my cooperation partners, Stefan Henz, Janko Dietzsch, Stephan Steigele, Guido Grimm, Matthias Schlee, Heinz Stockinger, Barbara Holland, Esther Rheinbay, Jan P. Meier-Kolthoff, Felix Ott, Friedrich Götz, and Tilmann Weber. It was a great pleasure and honor for me to work with you, guys!

I also want to thank my colleagues, namely Tobias Dezulian, Christian Rausch, Sandra Gesing, Holger Gast, Julia Trieflinger, Marine Gaudefroy-Bergmann, Jan Schulze, Kay Nieselt, Suparna Mitra, and Regula Rupp. Special thanks goes to Daniel Richter, who had to endure me as office roommate, and during many semesters where we jointly organized teaching courses. It always was a pleasure to work with you! Moreover, I remember many funny conversations I had with Daniel Richter during my PhD. It was a great time!

Last but not least, I want to express my gratitude towards Sabine Auch, Manfred Gerblinger, and Jochen Schumacher. They always listened to me when I stroke a bad patch, and their optimism helped me a lot. In this context, I want to dedicate my work to Epicurus, who showed me what (and who!) is really important in life.

In accordance with the standard scientific protocol, I will use the personal pronoun “we” to indicate the reader and the writer or (as explained in Appendix B) my scientific collaborators and myself.

Contents

I mean, after all, you have to consider we're only made out of dust. That's admittedly not much to go on and we shouldn't forget that. But even considering, I mean it's a sort of bad beginning, we're not doing too bad. So I personally have faith that even in this lousy situation we're faced with we can make it.

You get me?

From "The Three Stigmata of Palmer Eldritch" by Philip K. Dick

List of Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
COG	Clusters of Orthologous Groups
GBDP	Genome BLAST distance phylogeny
HGT	Horizontal Gene Transfer
HSP	High-scoring segment pair
LBA	Long Branch Attraction
LnL	Log-Likelihood
MAST	Maximum Agreement Subtree metric
ML	Maximum Likelihood
MP	Maximum Parsimony
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information, USA
NJ	Neighbour-joining
ORF	Open Reading Frame
WGP	Whole Genome Phylogeny
XGD	Xenologous Gene Displacement

Chapter 1

Introduction

Naturally, when writing a thesis dealing with evolutionary processes in 2009, one has to refer to the 150th anniversary of Charles Darwin's groundbreaking work "On the Origin of Species" (Darwin 1859), which, together with the work of Alfred Russel Wallace (Wallace 1858), provided the basis for the theory of evolution. To express it using Bernard of Chartres famous metaphor, I am like a dwarf on the shoulder of giants, standing there with deep gratitude towards all scientists who supplied the prerequisites for my humble contribution.

1.1 Background

This thesis mainly deals with phylogenetics, the study of evolutionary relationships between organisms, which is related to modern taxonomy providing a hierarchical classification of species. The idea of a hierarchical classification of organisms can be traced back to Aristotle ("scala naturae", "ladder of life"), who already divided organisms into groups that can be compared to modern classification (vertebrates vs. invertebrates, mammals vs. egg-bearing species, etc.). Aristotle's hierarchical system can be seen as the progenitor of modern taxonomy (Greene and Mayr 1992), invented by Carl von Linné (Linnaeus 1758). His work even had great influence on Darwin (Gotthelf 1999). But by the work of Darwin and Ernst Haeckel (Haeckel 1894), the aristotelian idea of a "ladder of life" was transformed into the concept of a "tree of life" that explicitly depicts speciation events (resulting in a phylogenetic tree). This modern concept of hierarchical classification inherently embraces the idea of evolution.

The beginning of computational phylogenetics is closely linked with the dawn of the age of molecular biology. Consequently, computational methods are not older than approximately 40 years (a comprehensive overview is given in Felsenstein 2004). While being a young discipline, methodology was greatly improved with the continuing evolution of wet-lab sequencing

technology. Early work mainly focused on single genes (e.g., Woese 1987), whereas the multitude of sequencing projects and completely sequenced genomes nowadays allows to compare the evolutionary history of a plenitude of single genes, as well as that of whole genomes. Thus, the question whether a single gene can reflect the phylogeny of the associated species, or whether a species phylogeny of prokaryotes actually exists, can now be approached (Rokas et al. 2003; Dagan and Martin 2006; McInerney et al. 2008; Dagan and Martin 2009).

However, some scientists even challenge the notion of a hierarchical classification or the existence of a meaningful species concept for prokaryotes (Bapteste and Boucher 2009; Boucher and Bapteste 2009). They argue that a prokaryotic lineage comprises different genetic elements that may or may not share a common evolutionary history, due to mechanisms like horizontal gene transfer (HGT). Some prokaryotic genomes may resemble a genetic chimera, even to an extent where a tree-like representation of the species phylogeny may become meaningless. In that case, it may be necessary to subdivide the species concept into evolutionary units that consist of genetic elements replicated together (i.e., a chromosome, plasmid, operon, or a transposon).

1.2 Outline of this thesis

During my PhD, I focused on several topics, which all are related to computational phylogenetics. In Chapter 2, a method for inferring whole-genome phylogenies is outlined. The basic idea consists of using completely sequenced prokaryotic genomes instead of single genes, thus leading to an "averaged" genomic phylogeny. This allows one to compare the obtained whole genome phylogenies with phylogenies based on single genes (or taxonomic trees) to attempt to illuminate the amount of congruence between both phylogenies.

Differences between a gene phylogeny and a "species" (or whole-genome) phylogeny can occur due to horizontal transfer of genes between prokaryotes. Consequently, we tried to assess the amount of horizontal gene transfer within a common set of prokaryotic genes in Chapter 3. The concept of a "true" species phylogeny depends on the existence of a common set of genes that resist intergenomic exchange. In our study, we tried to determine how much the phylogenies of these genes differ from a possible species phylogeny.

In Chapter 4, we focused on the study of cophylogenetic relationships. Here, the question is addressed whether the phylogenies of two groups of organisms (e.g., hosts and parasites) resemble each other. We developed a software toolkit that allows to handle large-scale cophylogenetic datasets, thus providing the instrument for studying deep cophylogenetic relationships.

A software package for the taxonomic classification of microbial communities is presented in Chapter 5. Metagenomics, the genomic analysis of entire microbial communities provides an insight into the biodiversity of different habitats like soil, marine water, acid mines, human and mouse gut.

Finally, the results of this thesis are summarized in Chapter 6.

Chapter 2

Whole-Genome Phylogeny

2.1 Introduction

At the beginning of the era of molecular systematics and phylogenetics, sequencing technology was limited to sequencing single genes or even single loci. Hence, taxonomists looked for suitable marker genes that could be easily detected and amplified, and also had to carry sufficient information for reconstruction of deep phylogenetic relationships.

The most widely used marker, at least for prokaryotic phylogenies, is the 16S rRNA gene (Woese 1987). A phylogenetic reconstruction based on this marker gene led to the proposal of a tripartite natural system of organisms by Woese and coworkers (Woese et al. 1990). They observed a deep dissimilarity between two groups of prokaryotic organisms that previously were thought to be more similar to each other than to the Eukaryota. Namely, the new taxonomic system comprised the domains Eukaryota, Archaea, and Bacteria, whereas the latter two are distinct prokaryotic taxa. Since this time, a discussion about the ancestry of all eukaryotic organisms began (see also Zimmer 2009), and to this day, no clear answer has been found whether the eukaryotic ancestor was among archaeal prokaryotes (Cavalier-Smith 2002; Ciccarelli et al. 2006; Saruhashi et al. 2008), or rather a chimeric organism (Rivera and Lake 2004; Simonson et al. 2005; Rivera 2007; Pisani et al. 2007). Furthermore, some groups even argue that the last universal common ancestor may have been an eukaryotic-like cell (Forterre and Philippe 1999; Kurland et al. 2006; Glansdorff et al. 2008).

Like the mythological character Chimera, composed of a mixture between different animals, genomes may contain segments of different origin. A popular example may be the eukaryotic cells, which also contain chromosomes of organelles like Plastids and Mitochondria, which were thought to have been derived via endosymbiosis (Gray 1989). Moreover, the so-called "B chromosomes" use an organism to travel as hitchhikers (Gregory 2005, p. 225), while they do not contribute any indispensable function for the

captured organism. But genomes are also shaped by smaller factors like recombination, inclusion of transposable elements, or other mechanisms.

Such events can have a devastating effect on phylogenetic reconstruction. Using sequences from different loci may lead to contradictory evolutionary scenarios with respect to the inferred tree topologies (Rokas et al. 2003). Here, the question arises, which of these genes may harbour the "true" phylogenetic signal, provided that one can assume that a true tree-like representation of evolutionary history even exists (Doolittle 1999a; 2000; McInerney et al. 2008).

Additional factors like saturation may further diminish the quality of a phylogenetic signal based on single locus data (Forterre and Philippe 1999; Gribaldo and Philippe 2004). Thus, the information content of such a small portion of the whole genome may be rather narrow, which eventually leads to the reconstruction of inaccurate trees (Aguileta et al. 2008).

An enhanced approach is the supermatrix method, which tries to combine as many as possible multiple sequence alignments (MSA) of orthologous genes into one single MSA, which can then be analyzed together. Datasets comprising more than 100,000 base pairs have already been assembled using this method (Rokas et al. 2003; Goremykin and Hellwig 2005; Hejnol et al. 2009). But establishing orthology of genes is error-prone and considered to be a hard task (see Section 3.1.2, page 57). Even if orthology of a distinct set of genes can be established, the question arises if the derived sequence length is sufficient to reliably infer phylogenies (Forterre and Philippe 1999). Furthermore, there exist genes that are not entirely homologous across domains, leading to a modularized view of homology that cannot be established for the entire gene (Fitch 2000; Di Giulio 2006; Glansdorff et al. 2008). MSA-based methods may only be applied when homologous fragments (e.g., protein domains) are collinear, i.e., the fragments have to be in the same consecutive order. This is a prerequisite that is not always met when dealing with genes that consist of a complex domain structure (Leipe et al. 1999; Apic et al. 2001a;b; Vishwanath et al. 2004). Yet another problem of MSA-based approaches is, that the exclusion of ambiguously aligned positions leads to a certain loss of information (Lee 2001). Particularly, fast evolving sites may harbour a strong phylogenetic signal regarding relationships between closely related taxa. But if these regions are discarded, resolution of the derived phylogenies may thus be greatly diminished.

In the age of genomics, an increasing number of fully sequenced genomes is available, and the speed of publication of new genomes accelerates. Thus, it may be quite natural to open the door to an era that is dominated by whole genome based phylogenies. The previously discussed pitfalls of single locus and MSA-based phylogenies do not hamper reconstructions that are based on entire genomic content. Saturation, genetic transfer, length restrictions, and paralogy thus can be handled by inclusion of the full range of data a whole genome perspective can provide. Even if a tree-like phyloge-

netic representation of the evolutionary history of life cannot be assumed to exist, whole genome based methods are better suited than other methods to reveal the basic vertical inheritance scheme underlying prokaryotic evolution (McInerney et al. 2008).

Several methods now exist to infer phylogenies based on whole genomic data. Some of these approaches are based on vectors of word-count frequencies, which can then be used for distance calculation (Qi et al. 2004), or the average length of maximum common substrings (Ulitsky et al. 2006). Other methods utilize gene presence/absence (Snel et al. 1999), gene order data (Tang and Moret 2003), or even complexity-based metrics (Otu and Sayood 2003). In the following, we present our own whole genome phylogeny approach, which is based on pairwise local alignments (Henz et al. 2003; 2005; Auch et al. 2006b;a; 2009a;b).

All of these methods have in common that they do not utilize multiple sequence alignments, and that they are based on calculating distances between taxa prior to phylogenetic reconstruction. Afterwards, phylogenies can be inferred using well-known tree-based methods like `NJ` (Saitou and Nei 1987; Studier and Keppler 1988) and `UPGMA` (Sokal and Michener 1958), or even network-based approaches (Huson 1998; 2003).

2.2 Methods

2.2.1 The "Genome BLAST distance phylogeny" approach

Basic steps of the `GBDP` strategy

Because of the huge amount of sequence data, which has to be considered in whole-genome phylogenies, and due to methodological restrictions, most WGP methods utilize distance-based instead of character-based tree reconstruction methods. In those cases, first a distance calculation is conducted, then one or several reconstruction methods are applied.

Basically, `GBDP` (Genome BLAST distance phylogeny) consists of the following steps:

1. All-against-all `BLAST` comparison of every genome against every other genome

2. Filtering of HSP (High-scoring segment pairs) data according to their e-Value and overlapping regions

3. Application of different distance methods

4. Tree or network reconstruction

After phylogenetic reconstruction, an evaluation of the obtained distance matrices and trees can be accomplished by calculating δ values (Holland et al. 2002), which provide a measure for the additivity of the distance matrices, and by computing support values for tree edges based on estimated distance variances (see Section 2.2.5, p. 24). Additivity as measured by the δ value of a distance matrix is a valuable criterion for the accuracy of the reconstructed phylogenetic trees (Auch et al. 2006b).

In the following, the HSP filtering step will be discussed. Afterwards, the `GBDP` distance functions are introduced, which all share a common structure. A `GBDP` distance function consists of a similarity function, which is normalized by selecting a corresponding denominator, and eventually, a dissimilarity conversion formula is applied to derive a distance (or "dissimilarity") value. Thereafter, a tree (or network) reconstruction algorithm is applied to the distance matrix.

HSP filtering

Basic filtering is done by defining maximal thresholds for the expectation value of HSPs. Since the e-Value decreases with growing alignment length, later filtering steps compensate for a high e-Value threshold setting. Thus, we used a high setting of 10^{-3} (empirically determined) for inferring most whole-genome phylogenies.

Additionally, overlapping segments are filtered by applying two different strategies based on the idea of placing all HSPs in a priority queue and inserting them in a selection list in the order determined by the priority queue (Figure 2.1). This approach is outlined in more detail in Henz et al. (2005) and is based on the greedy strategy described in Halpern et al. (2002, p. 136). These two strategies are the "Greedy" as well as the "Greedy with Trimming" filtering approaches, which are discussed in the following paragraphs.

"Greedy" filtering: HSPs in the priority queue are ranked according to their length. Each HSP is checked against overlap with all HSPs already contained in the initially empty selection list (see Figure 2.1). In case of an overlap, the current HSP will not be included in the selection list, and will thus be discarded.

"Greedy with Trimming" filtering: This is a slightly refined variant of the previously described approach. Instead of discarding the whole HSP, the overlapping part of the corresponding HSP is cut off and the HSP is afterwards re-inserted into the priority queue according to its new length (see Figure 2.1). When a HSP is trimmed, interval borders and its length are adjusted. Percentage identity as well as the normalized score are not

adjusted (i.e., recalculated from trimmed alignment data), since we assume a simplified model with equally distributed patterns. This simplification leads to a significant decrease of disk storage and memory (RAM) requirements in the current implementation of `GBDP` (see Section 2.2.3 for details).

Distance functions

After filtering of HSPs, several distance functions are applied to the remaining set of HSPs to infer inter-genomic distances between organisms. Initially, a similarity is calculated, which is afterwards converted to a dissimilarity for phylogenetic inferrence using two different formulas. In the following, the term "distance" is also used for dissimilarities, because these dissimilarity values can be seen as a distance in a biological sense Felsenstein (2004, p. 158).

Similarity formulas: For similarity calculation, we used different denominators for normalizing the values to a range between 0 and 1. For the definition of the following similarity formulas, we simply use the term g to represent the denominator. It can be substituted by any of the denominators defined below. We define H as a set of tuples (q, s), each corresponding to the intervals of an HSP (query and subject). $|X|$ and $|Y|$ denote the length of genomes X and Y respectively, whereas $|X_{cov}|$ and $|Y_{cov}|$ are defined as the number of characters that are covered by at least one HSP in genome X and Y, respectively.

Since similarity formula 2.1 is independent of the level of coverage, i.e., whether more than one HSP interval maps to a specific region, no prefiltering using the *Greedy* or *Greedy with trimming* strategy is performed in this specific case.

$$s_{cov}(X, Y) \quad := \quad \frac{|X_{cov}| + |Y_{cov}|}{g} \tag{2.1}$$

Some genomes consist of large regions of repeated sequences, leading to an underestimation of distances when applying this formula. This has been observed (see Henz et al. 2005) for distances between *Neisseria meningitidis* strains, which contain a large fraction of repetitive elements (Parkhill et al. 2000; Liu et al. 2002). To tackle this problem, we designed similarity function 2.2, which uses the overlap filtering methods mentioned above. Here, $|X_{match}|$ and $|Y_{match}|$ describe the sum of all interval lengths in genome X and Y respectively, obtained from the remaining HSPs only.

$$s_{match}(X, Y) \quad := \quad \frac{|X_{match}| + |Y_{match}|}{g} \tag{2.2}$$

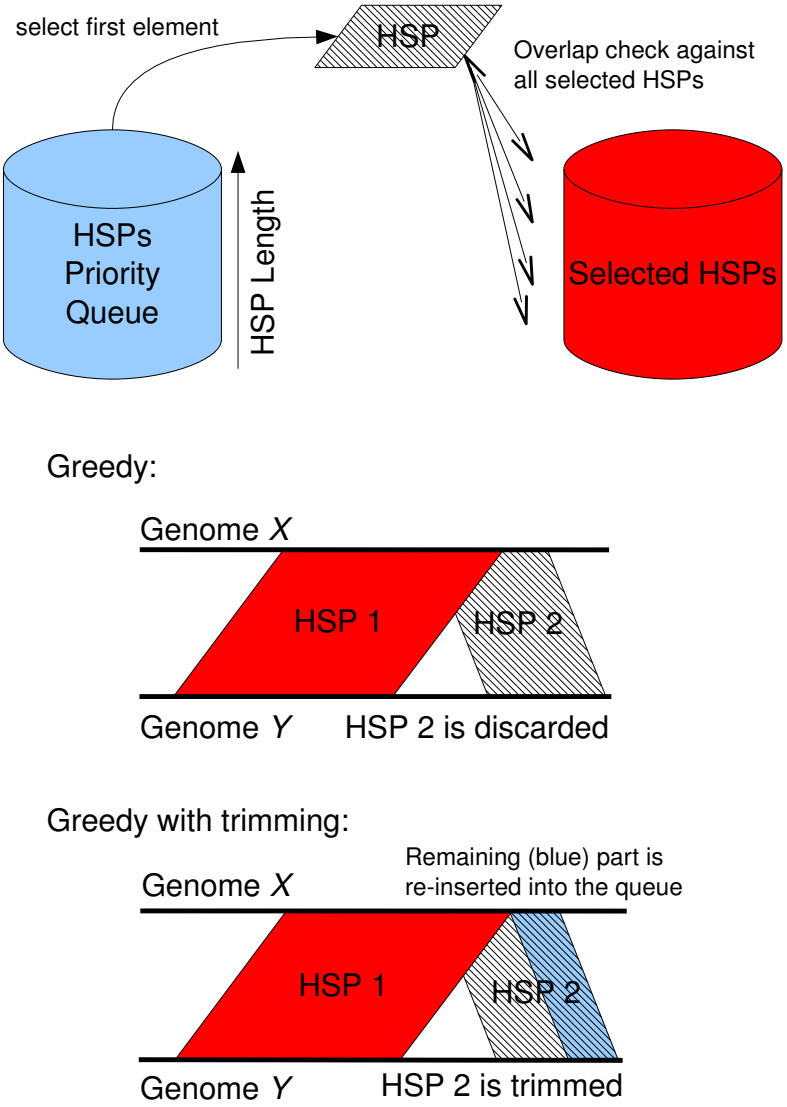


Figure 2.1: "Greedy" filtering strategy. A HSP is taken from the priority queue and checked against selected HSPs for overlapping regions.

Auch et al. (2006a) introduced a modification of the previous formula, counting only characters that are identical in both segments of an alignment. Here, ident(H) denotes the sum of the number of identical characters over all HSPs.

$$s_{idmatch}(X,Y) := \frac{2 \cdot \text{ident}(H)}{g} \tag{2.3}$$

Application of the identity function can be seen as using a rather simple scoring model, consisting of a substitution matrix with all diagonal entries set to 1 and all other entries set to 0. To further refine similarity estimation, we incorporated a more realistic scoring model to test whether this allows to improve the inferred distances:

$$\text{scoreratio}(q,s) := \max(|q|,|s|) \cdot \frac{\text{score}(q,s)}{\text{score}(s,s)} \tag{2.4}$$

$$s_{scorematch}(X,Y) := \frac{2 \cdot \text{scoreratio}(H)}{g} \tag{2.5}$$

The `scoreratio` function 2.4 thus is defined as the length of the larger segment multiplied by its normalized score. Score normalization in a range between 0 and 1 is done by dividing the score by the self-score of the respective subject sequence.

As substitution matrix, either pre-defined matrices or matrices based on observed frequencies can be used. Pre-defined matrices include, e.g., BLOSUM (Henikoff and Henikoff 1992) and PAM (Dayhoff et al. 1978) in case of amino acids, or Purine/Pyrimidine- and Transversion/Transition-based matrices for nucleotides. In Section 2.2.4 we exemplify a method for deriving empirical substitution matrices that are directly based on the `BLAST` aligment data.

Denominators: We introduced two different denominators in Henz et al. (2005). The second one (equation 2.7) performed better according to the data presented in Henz et al. (2005) for cases where one genome is actually a subset of another one. This is the case for the parasitic genome of *Buchnera aphidicola*, which is a subset consisting of approximately 14% of the genes from *Escherichia coli* (Moran and Mira 2001). Such a constellation would result in an overestimation of the distance between the two taxa when applying formula 2.6.

$$g_1 := |X| + |Y| \tag{2.6}$$

$$g_2 := 2 \cdot \min(|X|,|Y|) \tag{2.7}$$

Breakpoint similarity: We also implemented a similarity function that relies on the concept of breakpoints (Sankoff and Blanchette 1997; Blanchette et al. 1997; Sankoff et al. 2000; Wang et al. 2003). By applying the idea of breakpoints for homologous genes to our concept of homologous sequences based on HSPs, we define that a breakpoint occurs if a third, intervening HSP is found between two HSPs in genome X, but not between the two corresponding consecutive HSPs in genome Y (Auch et al. 2006b; Henz et al. 2005). An alternative definition would be that if HSP 1 is the direct neighbour of HSP 2 in X, then this must also be true for Y. Otherwise, a breakpoint is observed (see Figure 2.2).
Let B_X, B_Y be the number of observed breakpoints in X and Y respectively, and M_X, M_Y the amount of adjacent HSPs. We define the breakpoint similarity as follows (Auch et al. 2006b, equation 1):

$$s_{breakpoint}(X,Y) := 1 - \frac{B_X + B_Y}{M_X + M_Y} \tag{2.8}$$

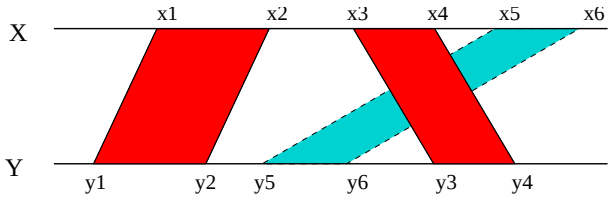


Figure 2.2: HSP-based Definition of breakpoints (Auch et al. 2006b). The Y segment of the turquoise HSP ($[y5, y6]$) is located between the two red HSPs ($[y1, y2, x1, x2]$ and $[y3, y4, x3, x4]$). This will be counted as a breakpoint.

Homology-based similarity function: In Auch et al. (2006b) we introduced a similarity function that is solely based on homologous sequence information, completely discarding any non-matching sequences. The idea behind this equation is analogous to single-gene phylogeny, where first a homology search is performed and afterwards tree reconstruction is only based on characters that are assumed to be homologous according to the applied alignment algorithm. We define ident(H) as the sum of the numbers of all identical character pairs in the obtained HSPs, and length(H) as the sum of the lengths of the larger interval for each HSP:

$$\text{length}(H) := \sum_{(q,s)\in H} \max(|q|, |s|)$$

Then we obtain the two homology-based similarity functions:

$$s_{hom} := \frac{\text{ident}(H)}{\text{length}(H)} \tag{2.9}$$

$$s_{scorehom} := \frac{\text{scoreratio}(H)}{\text{length}(H)} \tag{2.10}$$

The difference between these formulae and formulae 2.3, 2.5 lies in the definition of the denominator. Here, only the length of the sequences covered by a match are considered.

Dissimilarity conversion: All similarity functions introduced here are defined to be constrained between 0 and 1, and thus can be converted to dissimilarity values. Several approaches exist for converting similarity into dissimilarity values. Common options for conversion are (see Lefkovitch 1993; Legendre and Legendre 1998, p. 252):

$$d(X,Y) := 1 - s(X,Y) \tag{2.11}$$

and

$$d_{\log}(X,Y) := -\log(s(X,Y)) \tag{2.12}$$

The rationale behind formula 2.12 is to utilize a logarithmic scale to correct for saturation effects in the underlying data (e.g., see Felsenstein 2004, p. 158-159).

Note however that the so defined "distance" functions do not necessarily provide metric distances, i.e., distances that obey the triangle inequality (see e.g., Legendre and Legendre 1998, p. 274–275). Given three genomes X, Y, and Z, the triangle inequality states that $d(X,Y)$ has to be lesser or equal to $d(X,Z) + d(Z,Y)$. But other well-known distance transformations like Jukes-Cantor (Jukes and Cantor 1969) also do not obey the triangle inequality condition (Felsenstein 2004, p. 158). Moreover, Felsenstein (2004, p. 158) states that "most distance matrix methods do not absolutely require the Triangle Inequality to hold".

Matrix averaging

`BLAST` is not symmetric (Altschul et al. 1990). Thus, when using genome X as query and genome Y as subject the results can be different compared to using Y as query and X as subject. Therefore, after calculating dissimilarity values, symmetry of the dissimilarity matrix is ensured by averaging:

$$d_{avg}(x,y) = d_{avg}(y,x) := \frac{1}{2}\left(d(x,y) + d(y,x)\right)$$

Further possibilities would be to use the maximum or minimum of the asymmetric distances, but we have shown in Henz et al. (2005) that averaging produces better results when comparing the obtained phylogenetic trees to the NCBI taxonomy.

Tree and network reconstruction

As final step in phylogenetic inference, a distance-based phylogenetic reconstruction method is applied to the obtained distance matrices. Among tree-based algorithms, `UPGMA` (Sokal and Michener 1958), `NJ` (Saitou and Nei 1987; Studier and Keppler 1988), `BioNJ` (Gascuel 1997), `FastME` (Desper and Gascuel 2002; 2004), and `STC` (Vinh and von Haeseler 2005) are used. Additionally, network-based algorithms like Split Decomposition (Bandelt and Dress 1992b; Huson 1998) or `NeighborNet` (Bryant and Moulton 2004) can also be applied.

2.2.2 BLAST optimizations

Preliminary considerations

To identify local regions of high sequence similarity between two genomes, we used the popular tool `BLAST` (Altschul et al. 1990; 1997) in the implementation from Washington University (`WU-BLAST`), version 2.0MP-WashU, as well as the implementation from NCBI in version 2.2.18.

We did some preliminary tests to investigate memory consumption and run time of `WU-BLAST` and `NCBI-BLAST`. Generally speaking, the user has to decide either to optimize run time or memory requirements. Whereas `WU-BLAST` has a much lower memory footprint and is slightly slower, `NCBI-BLAST` tends to produce fewer HSPs in shorter time, at least when using standard settings. However, the difference in sensitivity between `WU-BLAST` and `NCBI-BLAST` did not lead to an observable effect on the quality of the `GBDP` phylogenies (see Section 2.3.1, p. 28). A detailed study of run time and memory consumption of different local alignment tools in view of the derived `GBDP` distances can be found in Auch et al. (2009b). The results clearly indicate that `NCBI-BLAST` provides a good balance between execution time and accuracy of the obtained `GBDP` distances.

Protein BLAST

In its original form, genetic data is available as nucleotide sequences. For this reason and because of run time considerations, our first approach consisted of using `BLASTN` to find similarities at the nucleotide level (Henz et al. 2005). But considering that bacterial genomes have a high density of gene coding regions, comprising 85% to 95% of the whole-genome (Saccone and Pesole 2003), using translated sequences for homology search seems to be

adequate. Furthermore, large difference in GC content as observed for many bacterial groups (Gregory 2005) can lead to a bias in nucleotide-based phylogenetic tree reconstruction, whereas reliability of amino acid sequences will be affected to a lesser degree (Hasegawa and Hashimoto 1993; Hashimoto et al. 1995). But it has to be considered that protein-based reconstruction may be hampered as well by differences in nucleotide frequencies (Foster et al. 1997; Singer and Hickey 2000). Even if such biases cannot be ruled out completely, protein sequences are believed to evolve more slowly than their corresponding nucleotide sequences. This is due to the fact that almost all amino acids are represented by more than a single codon. The last codon position carries the least significant signal, which is denoted as third base degeneracy in scientific literature (Lewin 2004, p. 168). Thus, most mutations in the third codon position are synonymous, i.e., they do not lead to a change in the corresponding amino acid sequence. From a phylogenetic perspective, this makes amino acid sequences ideal candidates to investigate deep evolutionary relationships.

However, using `TBLASTX` instead of `BLASTN` leads to a 36 fold increase in run time since six different reading frames (three for each of the two strands) have to be considered for each of the two genomes. Eventually, with growth of computing power and the availability of mid-sized and large Linux Clusters at the University of Tübingen, it became feasible to use `TBLASTX` as an alternative to `BLASTN`, even when dealing with large scale datasets.

Parameter selection

To avoid finding spurious homologies in sequence regions of low complexity, i.e., regions consisting of repeats of short patterns or even a single character, `BLAST` incorporates several filters designed to find such regions and to mask them out with ambiguity characters ("`X`" for protein and "`N`" for nucleotide sequences). According to our experience, low complexity filtering in the hit extension phase of the `BLAST` algorithm leads to HSPs that break apart in two or several smaller HSPs having a much lower bit score and thus, a higher e-Value. However, completely deactivating complexity filters leads to a huge increase of `BLAST` run time, so we decided to use soft masking. The soft masking mode constricts low complexity filtering to the seeding phase of the algorithm, leaving the hit extension phase unaffected and thus allows HSPs to remain intact while running faster. According to Korf et al. (2003, p. 120), this is the better option in most cases compared to using the default settings.

When using `TBLASTX`, an appropriate translation table for translating codons into amino acids has to be selected. For most Bacteria, translation table 11 (Bacterial and Plant Plastid Code, see NCBI 2008) can be used. The bacterial code differs from the standard code in using several additional

start codons. A specific code table exists for the group of Mollicutes (table 4), which reassigns the `UGA` codon serving as a stop codon in the universal code to Tryptophan (Trp). While there is evidence for this usage of `UGA` in Entomoplasmatales and Mycoplasmatales (Yamao et al. 1985; Bové 1993), a contrary observation was made for the group of the plant-pathogenic mycoplasmalike organisms (e.g, *Aster yellows*) and Acholeplasmataceae (Lim and Sears 1992), indicating standard usage as stop codon. Furthermore, even for different phyla, similar deviations concerning translation of the `UGA` codon as Trp could be observed, namely for *Bacillus subtilis* (Lovett et al. 1991; Matsugi et al. 1998) and *Escherichia coli* (Hatfield and Diamond 1993).

Since it remains unclear when to actually use translation table 4 instead of 11, we decided to use translation table 11 for all genomes. The impact on the `BLAST` similarity search should be negligible, since occurrence of an `UGA` codon falsely translated to a stop codon only leads to a small decrease of the score in case the other sequence contains no stop codon at that site. Furthermore, no irregularities regarding the placement of taxa belonging to the Mollicutes could be observed in the inferred phylogenies.

For nucleotide `BLAST`, we used NCBI `BLASTN` because of its run time properties. Parameters were `-F 'm D' -m 7 -S 3 -e 1E-2 -b 100000`, i.e., using the Dust filter for soft masking, XML output, usage of both strands, e-Value cut off of 10^{-2}, and a maximum amount of 100,000 HSPs per run.

When using translated `BLAST` (`TBLASTX`), we considered `WU-BLAST` because of its comparably small memory requirements. The NCBI version needed more than 8 GB memory when comparing large genomes, whereas the memory consumption of `WU-BLAST` almost always stayed below 1 GB. Options were `mformat=7 E=1E-2 wordmask=seg C=11 dbgcode=11 T=1000 W=3 hspmax=100000 hspsepSmax=50 hspsepQmax=50`, using XML output format, e-Value cut off 10^{-2}, SEG filter with soft masking, translation table 11 for query and subject sequence. To optimize performance, we added options to constrict maximum allowed separation between alignments along the query and subject sequence to 50, and to set the neighbourhood word threshold score to 1000. The first two options control which HSPs will be treated as candidates to be merged, which can have a huge impact on run time when finding many HSPs. We assume that there is no benefit in merging two HSPs having a distance of 50 sites, which means that they would have to bear the associated gap penalty, and thus would not be fused because of the declined score. The latter option forces `WU-BLAST` to require matching words in the seeding phase, leading to a huge increase in performance while having only a small decrease in sensitivity.

All sensitivity affecting options were tested against a small subselection of 20 taxa, representing the most important phyla as well as different genome sizes. We assured that the improvement in run time did not lead to a loss in

phylogenetic signal by comparing reconstructed trees with as well as without sensitivity affecting options to the NCBI taxonomy (Wheeler et al. 2008) using the c-score metrics (see Section 2.2.5, page 27).

2.2.3 Optimizations for large scale datasets

For an efficient handling of large datasets like the one presented in Section 2.3.4, we had to implement a solution based on a central data storage appliance as well as an aggressive scheme for data compression. This allowed us to handle the increasing amount of `BLAST` outputs, which grows quadratically with the number of genomes (due to the all-against-all comparison scheme).

Central Data Storage

The original `GBDP` program consisted of a set of scripts optimized for multiprocessor machines. Since the amount of single `BLAST` runs of this all-against-all approach increases quadratically, it was necessary to adapt this concept to a Cluster environment. For this purpose, a central repository was needed to deposit `BLAST` outputs of finished processes and to coordinate processes. Process coordination is needed to ensure that each job will be executed precisely once.

On the one hand, transaction integrity was an important factor to be considered, to guarantee that a file written to the server will be in a consistent state, even when the process crashes during writing its results. On the other hand, the central storage should be able to scale with the amount of concurrent processes, and it must be independent of the Cluster environment. A NFS or CIFS (Common Internet File System) server has to be mountable or at least a command line front end has to be present on the Cluster nodes, like `smbclient`. Using a network filesystem between several Clusters would involve alterations in the configuration of Cluster nodes that were not feasible due to administrative restrictions. Thus, we decided to implement a solution based on a SQL server instead of a network file system.

After careful consideration of drawbacks and benefits, we decided to use a PostgreSQL server (PostgreSQL 2008) as a file storage solution. This seems to be rather unconventional, since a SQL server is not designed to store such a large amount of BLOB (binary large object) based data (estimated database size for a 500 taxa project was more than 750 GB). But after an extensive evaluation phase, we conclude that this solution best matches our requirements regarding scalability and data integrity. Furthermore, by using data compression, we ensured that network usage remains within appropriate dimensions (see RFCs 1216 and 1925).

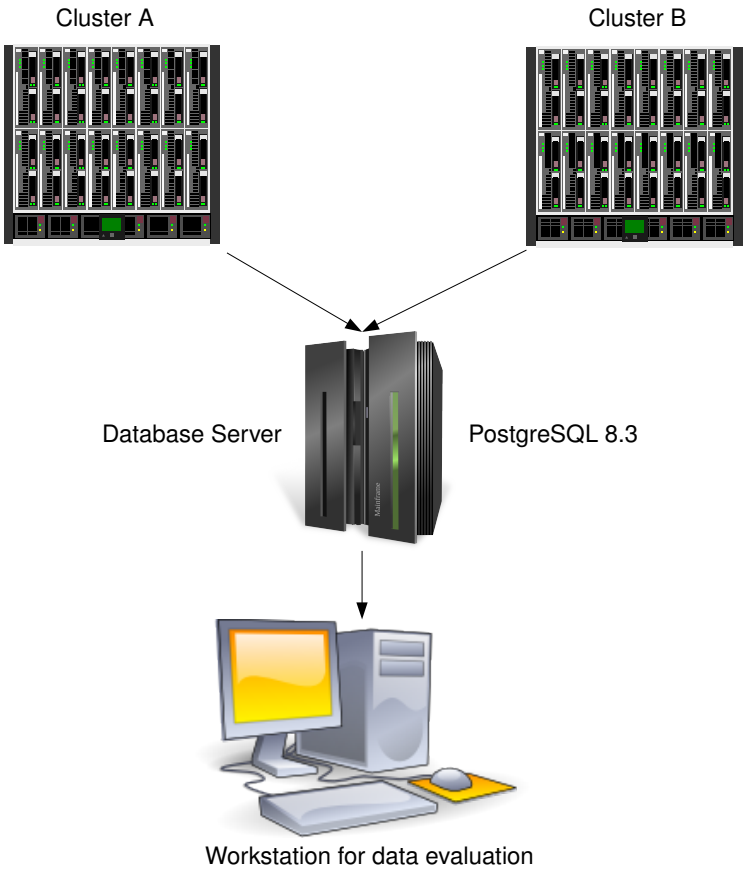


Figure 2.3: Data flow between Cluster nodes, database server, and the workstation where the final analysis is conducted.
Icons were derived from openclipart (2009).

The basic idea was to use a command line application on the Cluster nodes to handle data storage. The application provides elementary commands like "ls" (list directory), and "get" and "put" for reading and writing files to a repository. The repository can be a traditional file system directory, or a SQL database, whereas the location of the repository can be specified in a text-based configuration file. So, data storage can also be on a Cluster file system like Lustre or GPFS (Cope et al. 2005), or even NFS if performance and scalability are not critical. Small to medium datasets could thus be handled by using local facilities, whereas the computation of a large 500 taxa dataset (see Section 2.3.4) was done by configuring the application to submit the result files to the database server acting as a central data storage (see Figure 2.3). The frontend application for distance calculation was modified analogously to support data retrieval from the SQL server.

In the last two months of the 500 taxa project, three different Clusters were utilized, with more than 300 parallel `BLAST` runs producing approximately 700 GB of data. Thus, some modifications of the default configuration of PostgreSQL were necessary to handle a high amount of parallel sessions and to improve overall performance. We could clearly demonstrate that PostgreSQL is capable of handling such a scenario quite well. The SQL server hardware consisted of two quad core AMD Opteron 2.2 GHz processors with 32 GB of RAM and eight 500 GB disks in a RAID-6 ("Redundant Array of Inexpensive Disks") configuration. However, only a subset of these resources were actually employed since the server was run in a virtualized environment using Xen 3.2 and Linux Kernel 2.6.18 (Debian 4.0).

The schema of the SQL database is provided in Appendix C.1.

Data Compression

When dealing with large datasets, the quadratical growth of `BLAST` output has to be considered. The `BLAST` XML output used in our pipeline results in rather large files, but differences to the size of a `BLAST` text output diminish when a lossless compression algorithm like `bzip2` (Seward 2008) is applied. XML was favoured because it represents a modern, simple and easy to parse data exchange format.

However, dealing with data in the range of Terabytes forced us to look for a more space efficient solution. A first basic optimization step consisted of storing HSPs in a binary format, which also allows one to omit the effort to parse the output several times. But the decrease of file size was rather insignificant (see Table 2.1). This is due to the fact that the largest amount of data in a HSP is required by the actual alignment strings.

Since we also need alignment data for some distance calculations, using the condensed CGVIZ format (Delgado-Friedrichs et al. 2003), which only stores interval coordinates and statistical data like e-Value, was not applicable. This would be a substantial limitation of any further application of the

`BLAST` output, especially when using sequence information needed for substitution matrix estimation based on empirical frequencies (see Section 2.2.4). But a further improvement can only be achieved by losing information (or waiting for some new magic lossless compression algorithm).

Most algorithms that utilize alignment information are based on the assumption of independence between alignment columns, and thus will not be affected by a permutation of these columns. As an example, sequence-based phylogenetic inference methods like Maximum Parsimony or Maximum Likelihood are immune to such changes, at least theoretically. In practice, summing up floating point numbers is not associative, i.e., the order in which the numbers are summed up affects the outcome due to the limited precision of floating point numbers (Patterson and Hennessy 2009, p. 270-271). But such small discrepancies are artifacts without any biological meaning and thus, can be ignored in this context. In addition, estimation of an empirical substitution matrix also does not make use of any information based on the order of alignment columns.

Considering these facts, we implemented a lossy compression algorithm using a coding that disregards column order. In a first step, a matrix is built where each value represents the observed frequency of an unique character pair, by traversing the alignment column by column. Since score or distance calculation can be based on affine gap score models, we also have to consider the number of gap openings, which is assigned to an additional, usually illegal character pair consisting of two gap characters. Afterwards, the matrix is converted into a string representation by omitting character pairs that are not contained in the original alignment, i.e., having a corresponding value of 0 (see Figure 2.4). The resulting string is shorter than the original alignment, at least for all non-trivial cases. The chosen coding is not the most parsimonious one (compared to a binary encoding), but is optimized for readability. By applying a lossless compression algorithm afterwards, a convenient compression ratio can be achieved. Any gain that could have been achieved by using a binary encoding would have been almost completely counterbalanced by applying the compression algorithm to the string representation (even when applying the compression to the binary data as well).

An upper bound for the space complexity of a lossy compressed string is $O(\log n)$, which is a noticeable improvement over $O(n)$ for uncompressed strings. For calculating the upper bound, consider n_i to be the value of the ith matrix entry, then the number of characters to store this value as a string is $\lfloor \log_{10} n_i \rfloor + 1$. Thus, the upper bound is $O(\sum_i \log_{10} n_i) = O(c \cdot \log_{10} n) = O(\log n)$, where c denotes the maximum number of pairs (e.g., $21^2 = 441$ when using amino acids together with a gap symbol), which can be seen as an absolute term.

A large example using a `BLASTP` output is shown in Figure 2.5. By combining this algorithm with a lossless compression algorithm like `bzip2`, which

is based on the Burrows-Wheeler transformation (Burrows and Wheeler 1994), compression factors of almost 20 could be reached even when using medium-sized `BLAST` output files (see Table 2.1). The most time-consuming step of this approach consists of using the `bzip2` compression, but since this is executed in a highly parallelized environment, it was considered as negligible. In contrast, `bzip2`'s decompression is comparatively fast and no `BLAST` parsing has to be done afterwards, which suits the concept of analyzing the dataset on a single machine well.

Alignment:

```
query: GATTACAGATTACA
mid:    |||||    || |
hit:   -ATTAC--CATAGA
```

Matrix representation:

	A	**T**	**C**	**G**	**-**
A	4	1			
T		3			
C	1		1		
G			1		
-	1			2	2

String representation:

```
0/2;A-;AA/4;AC;CC;CG;G-/2;TA;TT/3;
```

Figure 2.4: Explanation of the lossy compression scheme. The count of each character pair is recorded in the matrix. Eventually, the matrix is converted to a string representation, including each observed pairing separated by a semicolon. If the number of observations is higher than one, the amount is specified after a slash symbol. The alignment contains two gap openings, which are expressed as `0` characters or (usually illegal) `-/-` pairings. The number of gap openings is needed when using an affine gap scoring model (see, e.g., Durbin et al. 1998, p. 16).

2.2.4 Substitution matrix estimation

The implementation for deriving a substitution matrix based on empirical frequencies follows in principle the log odds ratio calculation as outlined in Altschul (1991), Henikoff and Henikoff (1992), and Durbin et al. (1998, p. 14-15).

Original BLAST report:

```
Query: 10  YRNIGICAHVDAGKTTTTERILFYTGLSHKIGEVHDGAATMDWMVQEQERGITITSAATT 69
          YRNIGI AHVDAGKTTTTERIL  TG  H++GEVHDGA+TMD+M QE ERGITI SAATT
Sbjct: 7   YRNIGIFAHVDAGKTTTTERILKLTGKIHRLGEVHDGASTMDFMEQEAERGITIQSAATT 66

Query: 70  TFWRGMEAQFQEHRINIIDTPGHVDFTIEVERSLRVLDGAVVVFCGTSGVEPQSETVWRQ 129
           FW+G       HR N+IDTPGHVDFT+EV RSL+VLDG + VFCG+ GVEPQSET WR
Sbjct: 67  CFWKG-------HRFNVIDTPGHVDFTVEVYRSLKVLDGGIGVFCGSGGVEPQSETNWRY 119

Query: 130 ADKYGVPRMVFVNKMDRAGADFLRVVGQIKHRLGANPVPIQLNIGAEEEFKGVIDLIKMK 189
          A++  V R++FVNK+DR GADF RVV Q+K  LGANP+ + L IG E+EF GV+D++  +
Sbjct: 120 ANESEVSRLIFVNKLDRMGADFFRVVEQVKKVLGANPLVMTLPIGREDEFVGVVDVLTRQ 179

Query: 190 AINWNEADQGMSFTYEEIPADMLELAQEWRNHLVXXXXXXXXXLMEKYLEDGELSEVEIK 249
          A  W+++    +F  +E+PADM++  +E+R  ++         LM  Y+E  E +  +IK
Sbjct: 180 AYVWDDSGLPENFEVKEVPADMVDQVEEYREMMIETAVEQDDELMMAYMEGEEPTVEQIK 239

Query: 250 QALRQRTINNEIVLAACGSAFKNKGVQAVLDAVIEFLPSPTDV---PAIKGIDDRENSVE 306
            +R+ T +       CGSAFKNKG+Q VLDAV+++LPSPT+V   P
Sbjct: 240 ACIRKGTRDLAFFPTFCGSAFKNKGMQLVLDAVVDYLPSPTEVEPQPLTDPATGEPTGEV 299

Query: 307 RHADDNEPFSSLAFKIATDPFVGSLTFIRVYSGVVNSGDAVYNSVKQKKERFGRIVQMHA 366
               + P  +LAFKI  D F G+LTF+R+YSG + +  GD + NS   K ER GR+V+MHA
Sbjct: 300 ATVSVDAPLKALAFKIMDDRF-GALTFVRIYSGKIKKGDTILNSATGKTERIGRMVEMHA 358

Query: 367 NKRDEIKEIRAGDIAAAIGLKDVTTGDTLCDPNHVVILERMEFPEPVIQIAVEPRSKADQ 426
          N R+E++  +A DI A +G+K+V TG TLCDP H   LE M FP PVI IAV+P+ K
Sbjct: 359 NDRNEVESAQASDIIAIVGMKNVQTGHTLCDPKHECTLEPMIFPTPVISIAVKPKDKNGS 418

Query: 427 EKMGIALGKLAAEDPSFRVETDAETGQTLISGMGELHLDIIVDRMKREFGVDCNVGKPQV 486
          EKMGIA+GK+ AEDPSF+VETD ++G+T++ GMGELHLDI VD +KR +GV+  VG PQV
Sbjct: 419 EKMGIAIGKMVAEDPSFQVETDEDSGETILKGMGELHLDIKVDILKRTYGVELEVGAPQV 478

Query: 487 AYRETIRGKSEVEGKFVRQSGGRGQYGHVWLKIEPAEPGQGFVFVDAIAGGVIPKEFINP 546
          AYRETI    E      +QSGG GQ+G +  +I P E   GF F   + GG +PKEF
Sbjct: 479 AYRETITKAVEDSYTHKKQSGGSGQFGKIDYRIRPGEQNSGFTFKSTVVGGNVPKEFWPA 538

Query: 547 VAKGIEEQMNNGVLAGYPVLDVKATLFDGSFHDVDSSEMAFKIAGSMAFKKGALEAQPVL 606
          V KG +  M+ G LAG+PVLDV+  LFDG FH VDSS +AF+IA   AF++   +A P L
Sbjct: 539 VEKGFKSMMDTGTLAGFPVLDVEVELFDGGFHAVDSSAIAFEIAAKGAFRQSIPKAAPQL 598

Query: 607 LEPLMKVEITTPEDWMGDVVGDLNRRRGIIEGMDEGPAGLKIIHAKVPLSEMFGYATDLR 666
          LEP+MKV++ TPED +GDV+GDLNRRRG+I+  + G  G++ + A VPLSEMFGY   LR
Sbjct: 599 LEPIMKVDVFTPEDHVGDVIGDLNRRRGMIKDQEMGLTGVR-VKADVPLSEMFGYIGSLR 657

Query: 667 SATQGRASYSMEFAEYADVPKNIADAIIAE 696
          + T GR  +SMEF+ YA  P N+A+ +IAE
Sbjct: 658 TMTSGRGQFSMEFSHYAPCPNNVAEQVIAE 687
```

Packed representation:

```
0/4;-E;-P;-Q;A-;AA/29;AC;AE/2;AF;AG/3;AI/4;AL/2;AM/3;AN;AQ;A
R;AS/3;AT/4;AV/5;CC/3;CF;CL;DA;DD/25;DE/4;DG/4;DH;DN/3;DP;DS
/3;DT;DV;E-/2;EA/3;ED/6;EE/29;EH;EI;EK/5;EM/2;EP;EQ;ER;ES/3;
ET/2;EV/2;EY;F-;FF/22;FH;FI;FK;FL;FY/2;GA;GD;GE/3;GG/48;GK;G
N;GP/2;GS/2;GY;HH/9;HK/3;HM;HT;I-;IA/2;IF/3;II/19;IK;IL/3;IM
/3;IR;IT/2;IV/15;IW;IY;KA/3;KD/3;KE/4;KK/16;KN;KQ/2;KR/4;KT/
4;KV;LF;LI/4;LK;LL/22;LM/4;LP/3;LQ;LV/3;LY;M-;ME;MG;MI;ML/3;
MM/12;MQ;MR;MV;ND/4;NE/2;NK/2;NL;NN/9;NP/2;NT/2;NV;PA;PL;PP/
22;PQ;PR;PS;PV;Q-/2;QA/3;QE/3;QG;QK;QL;QM;QQ/7;QS/4;QT;QY;RA
;RE;RG;RI;RK/4;RP;RQ/2;RR/22;RS;RT;RV;SA/2;SD;SG/3;SI;SK/4;S
N;SQ;SS/13;ST/2;SV;TC;TD;TE/2;TF;TG;TQ/2;TS/2;TT/22;V-;VA;VC
/2;VD;VE/4;VF;VG;VI/8;VK/3;VL;VM;VN/2;VQ;VT/2;VV/31;WD;WF;WH
;WW/3;WY;XA;XD/2;XE/3;XQ;XT;XV;YF/3;YL/2;YS;YV;YY/6;
```

Figure 2.5: A BLAST text output of a HSP and its packed representation. Compression ratio increases with growing sequence length.

Format	Size in kB	compression factor
BLAST XML	43680	1.00
BLAST XML bzipped	9540	4.58
binary uncompressed	33776	1.29
binary bzipped	9792	4.46
lossy binary uncompressed	14300	3.05
lossy binary bzipped	**2192**	**19.93**

Table 2.1: Comparison of data formats, size and compression factor relative to the original BLAST XML output of a sample TBLASTX run (NC_000959 against NC_003888).

To determine the frequencies of character pairs, we use the alignments of the corresponding HSPs from all pairwise BLAST results and treat them as one large pairwise alignment. All pairs containing a gap character are discarded.

Let n be the length of this alignment, consequently the overall number of characters is $2n$. We count the observed frequencies $f_{a,b}$ for each pair of aligned residues a, b.

$$p_{a,b} = \frac{\frac{1}{2}(f_{a,b} + f_{b,a})}{n} \tag{2.13}$$

$$q_a = \frac{f_a}{2n} \tag{2.14}$$

$$s(a,b) = 2 \cdot \log_2 \left(\frac{p_{ab}}{q_a q_b} \right) \tag{2.15}$$

The term $p_{a,b}$ denotes the observed probability of occurrence for each pair a, b or b, a. Analogously, q_a denotes the probability of occurrence of a specific character in an alignment column. Since there is no specific order of the aligned sequences, i.e., whether one sequence is used as query or subject is of no importance for probability calculation, we ensured that $p_{a,b} = p_{b,a}$ by using the sum of the observed frequencies $f_{a,b}$ and $f_{b,a}$. This leads to a symmetrical scoring matrix, $s(a,b) = s(b,a)$.

In compliance with Henikoff and Henikoff (1992), we used a scaling factor of 2 in formula 2.15 to derive values in half-bit units being in the same range as the BLOSUM and PAM (Dayhoff et al. 1978) matrices.

Values for ambiguity characters were estimated by averaging over the values represented by the respective ambiguity character. For every pair involving the special character '*', indicating a stop codon when using amino acids, a value of $\min(s(a,b))$ is used. The score of a */* pair was hard-coded and set to 1, analogous to the BLOSUM matrices.

In the current stage, no explicit model for gap cost estimation is incorporated. Instead of this, the affine gap scoring model of `BLAST` is used, i.e., a gap open penalty of 9 and gap extension penalty of 2 for amino acids.

2.2.5 Evaluation of distances and phylogenies

Distance matrix evaluation

Reliability of distance-based tree inference algorithms is closely tied to the quality of the underlying distance matrix. Depending on the selected reconstruction algorithm, a distance matrix has to fulfill different mathematical properties. For instance, when using UPGMA, distances should obey the ultrametricity condition (Durbin et al. 1998, p. 168f).

Besides the strict claim of ultrametricity, which requires that the underlying sequences evolve under a molecular clock with a constant rate, a minimal requirement for reasonable tree reconstruction is that the distances should mostly be treelike (Felsenstein 1984). A mathematical definition of treelikeness is given by Buneman (1971) in form of the additivity condition. Holland et al. (2002) described an approach related to statistical geometry, which allows to measure the degree of deviation from treelikeness of a given distance matrix. It is based on inspection of distances between quartets of taxa (see Figure 2.6). For each quartet, a value $\delta_{Quartet} := \frac{q}{r}$ is computed, where q is the smaller one of both distances. When $r = 0$ (and thus, $q = 0$), $\delta_{Quartet}$ is defined to be 0. From this definition, it follows immediately that $\delta_{Quartet}$ is in a range between 0 and 1, whereas 0 indicates perfect treelikeness of the quartet under consideration. When this process is repeated for each quartet of taxa, an average δ value for the whole distance matrix can be computed. However, the number of quartets for n taxa is $\binom{n}{4}$, resulting in a time complexity of $O(n^4)$ when using all quartets. For this reason, we used an implementation that uses a random subsample to estimate the δ value for the whole matrix, as proposed by Holland et al. (2002).

Additionally, we defined $\epsilon_{Quartet}$, which differs from $\delta_{Quartet}$ only when $r = 0$, then $\epsilon_{Quartet}$ is defined to be 1. The ϵ value is calculated as the average over all $\epsilon_{Quartet}$ (or a subsample). The rationale behind this is that a distance of 0 for r and q has no biological meaning, and thus, should be treated like conflicting signal.

A similar approach is the Q criterion (Guindon and Gascuel 2002), which is simply defined as the sum of q over all quartets (or a subsample). A treelike distance matrix thus has a Q value of 0. We analogously defined R as the sum of r over all quartets. Furthermore, we used non-ultrametricity and non-additivity criteria based on the minimization formulae of Makarenkov and Legendre (2001) and De Soete (1986).

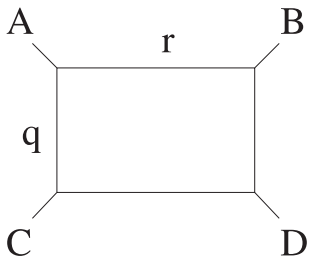


Figure 2.6: Quartet of taxa and their distances (Auch et al. 2006a). By convention, the smaller distance is labelled with q, so that $q \leq r$ holds.

Interior branch testing

In most studies, phylogenetic reconstruction is followed by an assessment of individual branches within the trees under consideration. The most commonly used approach is the bootstrap invented by Efron (1979) and adapted by Felsenstein (1985) to MSA-based phylogenies. The basic idea is to assess the uncertainty of the estimated tree topology by generating replicates from the original MSAs by using a sampling with replacement strategy. Here, the independence of alignment columns, which are treated as sampling points, is implicitly presumed. By conducting tree inference for each replicate, the fraction of how often a distinct branch (or bipartition) is observed in the trees can be seen as a confidence value for this branch.

However, statistical properties of the traditional bootstrap approach are subject of an ongoing controversial debate (see, e.g. Hillis and Bull 1993; Felsenstein and Kishino 1993; Zharkikh and Li 1995; Efron et al. 1996). Furthermore, there is doubt whether bootstrapping has to be considered to be consistent when taxon sampling is increased (Lecointre et al. 1993; Poe 1998).

Besides the bootstrap, there exist other methods for interior branch testing that have a different statistical background. Wróbel (2008) gives a comprehensive overview of this field (see also Anisimova and Gascuel 2006). A straightforward application of the underlying principle of bootstrapping (sampling with replacement of data points) to distance data is not possible. Using a resampling strategy based on alignment data would at least be computationally demanding, if not even impossible with current hardware. Since handling large alignment data in whole-genome phylogeny requires the usage of distance-based algorithms, consequently, bootstrapping cannot be considered as an option.

Sanjuán and Wróbel (2005) introduced a method based on a Weighted Least-squares Likelihood ratio test, which can be used for support value calculation for a given tree and distance matrix. The Weighted Least Squares (WLS) approach is based on the minimization of the difference between observed distances $d(i, j)$ from the matrix and the estimated, patristic dis-

tances $d_{patristic}(i,j)$ derived from the corresponding tree (Felsenstein 2004, p. 148).

$$\text{WLS} := \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\left(d(i,j) - d_{patristic}(i,j)\right)^2 \tag{2.16}$$

Here, $w_{i,j}$ denotes the weights assigned to the corresponding distances, which depend on the preferred WLS method. A minimization of the WLS can be obtained by adjusting distances as well as the tree topology. Distances can be adjusted by transforming the patristic distances expressed as the sum of individual path lengths into a set of linear equations obtained by differentiating the WLS with respect to one of the individual path lengths (see Felsenstein 2004, pp. 148-153). The tree topology can be optimized by applying heuristic search strategies (see Felsenstein 2004, Chapter 4).

When assuming that distances are normally distributed and independent, a log likelihood for the given tree can be derived from the above formula by considering the variances of the observed distances (for details, see Felsenstein 1984; Sanjuán and Wróbel 2005). This can be exploited for interior branch testing, by collapsing the considered branch, and recalculating lengths of the remaining branches. Then, the log likelihoods of the collapsed and the complete tree can be compared.

Sanjuán and Wróbel (2005) provide an implementation of their WLS Likelihood ratio test, called `WeightLESS`. It also includes computation of the Felsenstein (1984) F-test. This test allows for calculation of p-values for internal branches without the need to estimate variances beforehand.

Variance estimation

To be able to utilize `WeightLESS`, we need to estimate variances for the `GBDP` distances. Analogous to Kimura and Ohta (1972), we can obtain an estimate of the variance of the fraction $p_{X,Y}$ of nucleotide or amino acid sites that differ in the two genomes X and Y. The fraction $p_{X,Y}$ can be estimated by using one of the distance formulae defined above. Then, the binomial variance formula $\frac{p\cdot(1-p)}{n}$ can be applied (see Sokal and Sneath 1963; Felsenstein 2004, p. 214):

$$\text{Var}\left(d(X,Y)\right) := \frac{d(X,Y)\cdot\left(1-d(X,Y)\right)}{n} \tag{2.17}$$

Here, n represents the number of sites (denominator of the applied distance formula); n can be g_1, g_2 or length(H).

For logarithmic distances, we used the derivative of the distance correction formula 2.12, following the principle outlined in Felsenstein (2004, p. 214-215):

$$d'_{\log} = \frac{1}{1 - d(X,Y)} \tag{2.18}$$

The square of the derivative $d'_{\log}$ is then multiplied with the binomial variance formula:

$$\begin{aligned} \mathrm{Var}\left(d_{\log}(X,Y)\right) &:= \frac{d(X,Y)\cdot(1-d(X,Y))}{n}\cdot\left(d'_{\log}\right)^2 \\ &= \frac{d(X,Y)}{n\cdot(1-d(X,Y))} \end{aligned} \tag{2.19}$$

By applying formulae 2.17 or 2.19 to the entire distance matrix, we thus obtain a variance matrix, which is suitable for calculating support values using `WeightLESS`.

Comparison against a reference taxonomy

Several metrics exist for tree comparison. The most widely used metrics are the splits-based symmetric difference (Robinson and Foulds 1981, RF distance), weighted RF distance (considering the weights of the splits), and the Maximum Agreement Subtree metric (Goddard et al. 1994, MAST). All these metrics are applicable when dealing with binary trees. However, taxonomies tend to contain multifurcations since they are based on a small number of hierarchical levels of taxonomic units. The nomenclature introduced by the Swedish Biologist Carl von Linné in his work "Systema Naturae" (Linnaeus 1758) is a well-known example for such a hierarchical taxonomy. It provides the groundwork for the NCBI taxonomy (NCBI 2009c; Wheeler et al. 2008), which we use as reference taxonomy.

The NCBI taxonomy was chosen since it is updated at regular intervals, and thus, changes in taxonomy due to erroneous placements or inclusion of a new species are carried out in time. Furthermore, all species that are represented in Genbank (Benson et al. 2008) also have an entry in the taxonomy database of NCBI. During the study of Henz et al. (2005) we also evaluated the usage of 16S rRNA phylogenies as reference trees. Results were mainly in congruence with those obtained by using the NCBI taxonomy, whereas the data collection process was more labour-intensive and hard to automate even when using relevant databases like the Ribosomal Database project (Cole et al. 2009). To date, we successfully used the NCBI taxonomy

in many different projects to derive approximations for phylogenetic trees (Auch et al. 2006b; Huson et al. 2007b; Meier-Kolthoff et al. 2007).

As outlined in Henz et al. (2005), we developed a tree comparison metrics known as compatibility score ("c-score") as a refinement of the RF distance. It was specifically designed to compare fully resolved trees against multifurcating trees as obtained from taxonomical systems. To avoid over-counting of false positives, as it would be the case when applying the RF distance in this scenario, the c-score makes use of the concept of compatibility between splits. For a detailed explanation of splits and compatibility, the reader may be referred to Bandelt and Dress (1992a), as well as Huson (1998).

A split is called "non-trivial" if both partitions contain more than one single taxon, i.e., all splits that are not derived from a leaf edge are non-trivial splits. $\Sigma(T)$ denotes the set of all non-trivial splits in tree T, and $\Sigma(T_{compatible})$ is defined as all non-trivial splits of T that are compatible to the splits of the reference tree T_0. Thus, $\Sigma(T_{compatible})$ can be seen as a set of splits that are either already contained in $\Sigma(T_0)$ or are a refinement of T_0.

$$\text{c-score} := \frac{|\Sigma(T_{compatible})|}{|\Sigma(T)|} \tag{2.20}$$

The score is normalized to a range between 0 and 1, whereas 1 indicates perfect accordance with the reference tree. Although the current (naïve) implementation of the c-score algorithm has a time complexity of $O(n^3)$, calculating the c-score for trees comprising several hundreds of taxa can be done within seconds on a desktop PC.

2.3 Results and Discussion

Since there exist several variants of filtering strategies, nominators, denominators, and dissimilarity conversion formulae, we developed the nomenclature described in Table 2.2 to identify each specific distance algorithm. The nomenclature is used in the following sections.

2.3.1 Prokaryotic datasets

Species delineation with GBDP

In Microbiology, species classification is a challenging task compared to most eukaryotic phyla. Many animal, plant and fungal species can be distinguished by an abundance of morphological differences, behavioural traits, or by interbreeding barriers. Morphological features and metabolic peculiarities can be used to classify microorganisms to a certain degree, but the

Number	Nominator	Denominator	Dissimilarity conversion
0	2.2	2.6	2.11
1	2.2	2.7 (min)	2.11
2	2.2	2.6	2.12 (log)
3	2.2	2.7 (min)	2.12 (log)
4	2.9	2.9 (hom)	2.11
5	2.9	2.9 (hom)	2.12 (log)
6	2.3	2.6	2.11
7	2.3	2.7 (min)	2.11
8	2.3	2.6	2.12 (log)
9	2.3	2.7 (min)	2.12 (log)
10	2.5	2.6	2.11
11	2.5	2.7 (min)	2.11
12	2.5	2.6	2.12 (log)
13	2.5	2.7 (min)	2.12 (log)
14	2.10	2.10 (shom)	2.11
15	2.10	2.10 (shom)	2.12 (log)

Table 2.2: Nomenclature of distance functions names. The fully qualified name of a distance function is built by appending a prefix to the selected number. Prefixes denote the filtering approach, i.e. "`g`" for greedy, "`tr`" for greedy with trimming, or they denote a specific algorithm like "`bp`" for breakpoint distances and "`cov`" for the simple unfiltered coverage distance. Furthermore, a label is added to the prefix indicating if matrix averaging ("`a`"), minimum ("`min`") or maximum ("`max`") is used (see Section 2.2.1). For example, "`g_a10`" characterizes the function obtained by applying a greedy filtering strategy, followed by using nominator 2.5 with denominator 2.6 and dissimilarity conversion formula 2.11, as well as matrix averaging. A specific tree reconstruction can be referenced by appending the applied tree reconstruction algorithm, e.g., `g_a10_bionj` would refer to the tree reconstructed by using `BioNJ` with distance algorithm `g_a10`.

number of features and peculiarities that can easily be recognized is limited (Fraser et al. 2009). But even those recognizable features are not adequate to classify closely related species that share the same shape and metabolism. Consequently, species delineation nowadays is mainly based on DNA-DNA hybridization (DDH, see Rosselló-Mora 2006) experiments.

There exist several different techniques and standards for DNA-DNA hybridization. A common approach consists of cutting the genome of the test organism and the genome of a reference organism (type strain) into small fragments of 600-800 bp. Then, the mixture of fragments from both species is heated so that the DNA double-strand molecules dissolve. Afterwards, the temperature is decreased until the fragments form hybrid double-strands. The melting temperature depends on the degree of similarity between both

strains of a double-strand, thus by a stepwise increase of the temperature, the melting temperature can be determined and combined into a single DDH value. The DDH value usually is specified in percentage relative to the DDH value obtained by hybridizing the reference genome to itself. A value of 70% DDH or below is considered as an indication that the test organism belongs to a different species (Wayne et al. 1987).

Determining DDH values is an error-prone and labour-intensive process. Furthermore, there exist various DDH methods, which yield different results (Goris et al. 2007). This is considered as a major drawback of the hybridization approach. Consequently, several in silico methods were developed in recent years as an alternative to DDH (Konstantinidis and Tiedje 2005; Hanage et al. 2006; Goris et al. 2007; Martens et al. 2008; Deloger et al. 2009).

In Auch et al. (2009a;b) we proposed a new in silico method for determining species boundaries based on the `GBDP` method (see Figure 2.7). The new method correlates better with existing DDH data than the previously mentioned methods. Performance comparisons between the local alignment search tools `NCBI-BLAST` V. 2.2.18, `WU-BLAST` V. 2.0MP-WashU [04-May-2006], `BLAT` V. 34 (Kent 2002), and `BLASTZ` V. 7 (Schwartz et al. 2003) were conducted in regard to run-time and memory requirements, as well as in regard to the correlation between the obtained `GBDP` results and corresponding DDH values. Results indicated that `NCBI-BLAST` performs best considering run-time and memory requirements, while also providing an accurate correlation and error rate compared to DDH. The error rate was measured by comparing results of a classification of organisms close to the 70% DDH threshold for species delineation with the alternative classification provided by using `GBDP`. `NCBI-BLAST` was approximately two times faster than the second best program, `BLASTZ`, while its memory consumption was moderate ($\frac{1}{3}$ more than `BLASTZ`'s). However, these tests were only conducted using DNA-DNA-similarity search (`BLASTN`), thus no conclusions for protein similarity searches can be deducted.

In accordance with the test results, a web service was developed to facilitate utilization of the new method for the scientific community (see Figure 2.8). The web service can be accessed at http://www.gbdp.org/species.

Phylogenies

Previous work by Henz et al. (2003; 2005) covered a `BLASTN`-based similarity search and using distance algorithms `cov_a0` to `cov_a3`, `g_a0` to `g_a3`, `tr_a0` to `tr_a3` as well as `bp_a0` and `bp_a1`. Using a set of 91 prokaryotic genomes, we observed that distance algorithm `g_3` had the highest c-score (0.727, see Figure 2.9) by comparison to the NCBI taxonomy.

We extended the preceding work by using an enhanced taxon set comprising 97 bacterial and archaeal genomes, and by using two additional distance

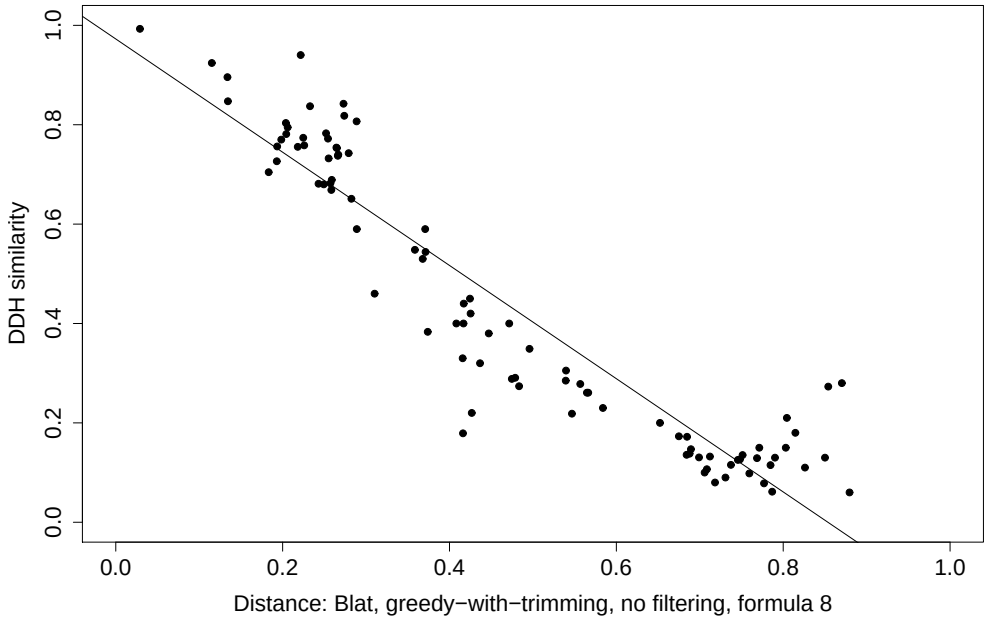


Figure 2.7: Comparison of GBDP distances and DNA-DNA hybridization data. The partial regression line shows that there is a strong linear correlation between GBDP distances and DDH data.

Figure 2.8: GBDP web service for computing DDH-like distances, α version (http://www.gbdp.org/species).

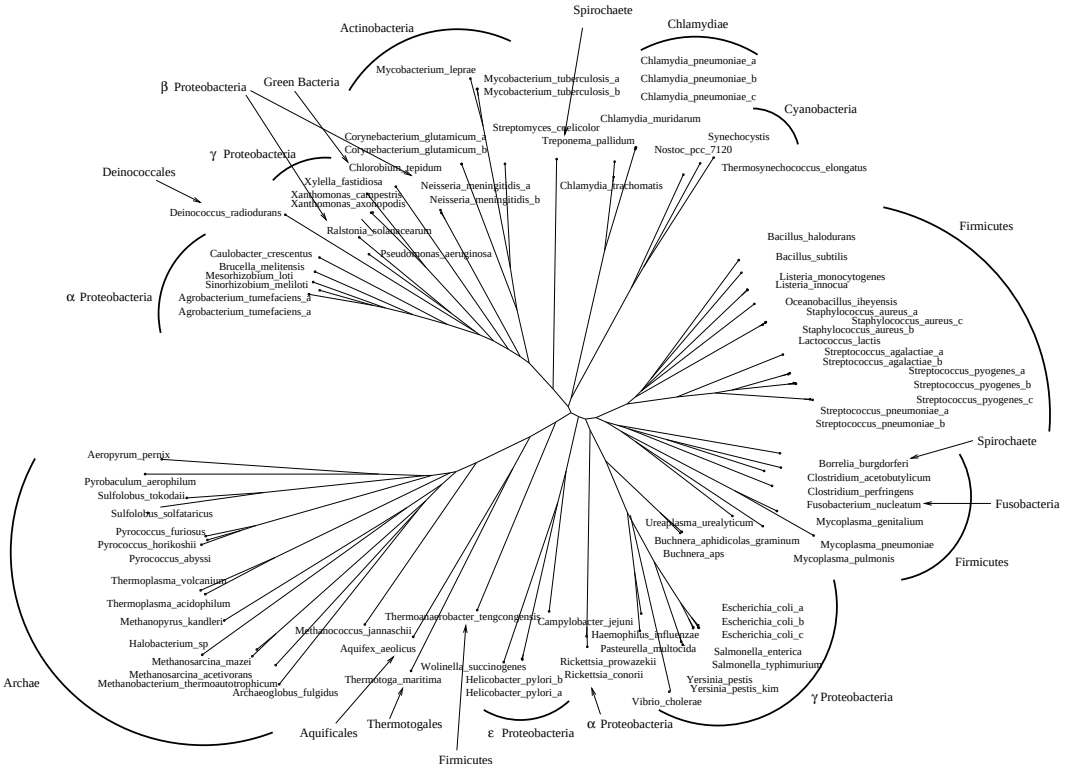


Figure 2.9: `BioNJ` tree reconstructed from a distance matrix calculated by using algorithm `g_a3` (see Table 2.2). Data from Henz et al. (2005). c-score is 0.727.

algorithms no. `4` and `5` based on homology (Formula 2.9, see Table 2.2). In this study, nucleotide-based as well as protein-based similarity search was carried out by using NCBI BLAST's `TBLASTX` algorithm. We also tested if calculating an average matrix between both matrices for each distance function, i.e., the one derived by using `TBLASTX`, and the second one by `BLASTN`, would lead to an improvement in accuracy.

Figure 2.10 shows boxplots of c-scores for different combinations of HSP selection and similarity search algorithms. Boxplots for greedy and "greedy with trimming" show a large interquartile range, because its first quartile lies below a c-score of 0.20. The reason for this is a low c-score of the homology-based functions `4` and `5`, which is in contrast to results obtained by using `TBLASTX`, where no similar observation could be made.

Overall, protein-based similarity search leads to higher c-scores than nucleotide-based, whereas "greedy with trimming" clearly outperforms the other approaches. Matrix averaging does not lead to changes in the c-score median values, but also diminishes performance of the best distance functions, especially when using functions `4` and `5`. When using trimming, not a single case could be observed when matrix averaging resulted in better c-scores.

The best tree was achieved by using distance algorithm `tr_6` with `BioNJ`, having a c-score of 0.8511. A cluster network representation created by Dendroscope (Huson et al. 2007c) is shown in Figure 2.11.

2.3.2 Mitochondrial and Plastidial datasets

Experimental setup

Taxon selection Completely sequenced genomes of plastidial as well as mitochondrial organelles were downloaded from NCBI (2005) and EBI (2005). In case of more than one genome belonging to the same species and having a different length, we randomly selected one sequence as proxy for all genomes having the same length and species affiliation. It is known that Apicomplexa, which are unicellular, parasitic eukaryotes, contain a special kind of extrachromosomal circular DNA that is considered to be derived from plastids (Köhler et al. 1997). We also included two genomes of this group, namely *Toxoplasma gondii* and *Eimeria tenella*.

For plastidial phylogenies, we used three cyanobacterial genomes, *Synechococcus* sp., *Synechocystis* sp., and *Thermosynechococcus elongatus* as outgroup. It is commonly believed that plastids originated from an ancestor of cyanobacteria (Gray 1989).

A multitude of studies indicate an α proteobacterial origin of mitochondria (e.g., Yang et al. 1985; Gray et al. 2001; Esser et al. 2004). Recent studies more precisely located the closest relative of mitochondria within the Rickettsiales (Lang et al. 1999b; Emelyanov 2003a;c). Correspondingly,

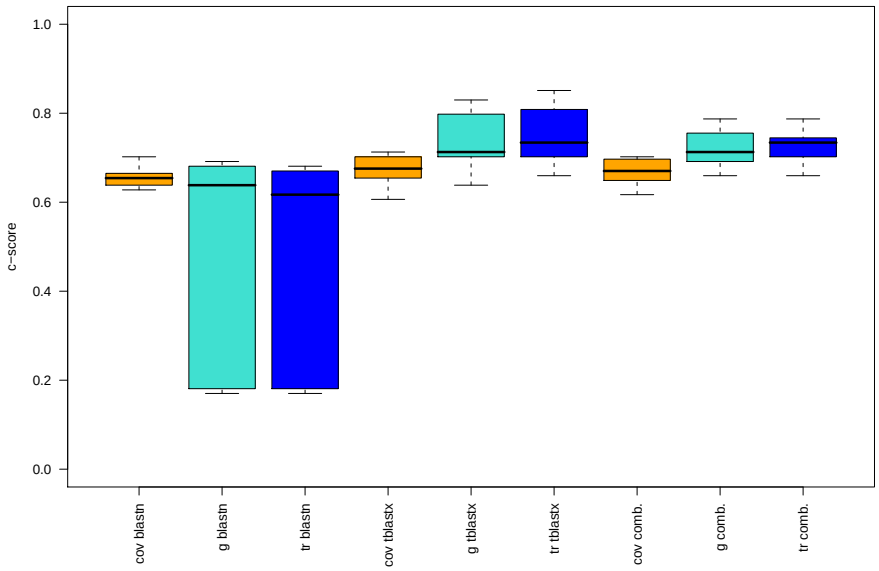


Figure 2.10: Boxplots showing c-scores for `GBDP` functions `cov_a0` to `cov_a3`, `g_a0` and `tr_a0` to `g_a6` and `tr_a6` (see Table 2.2) using tree reconstruction methods `UPGMA`, `NJ`, and `BioNJ`. A boxplot is shown for each combination of homology search (`BLASTN`, `TBLASTX`, and combined data) and filtering algorithm (`cov`erage, `g`reedy, and `tr`imming).

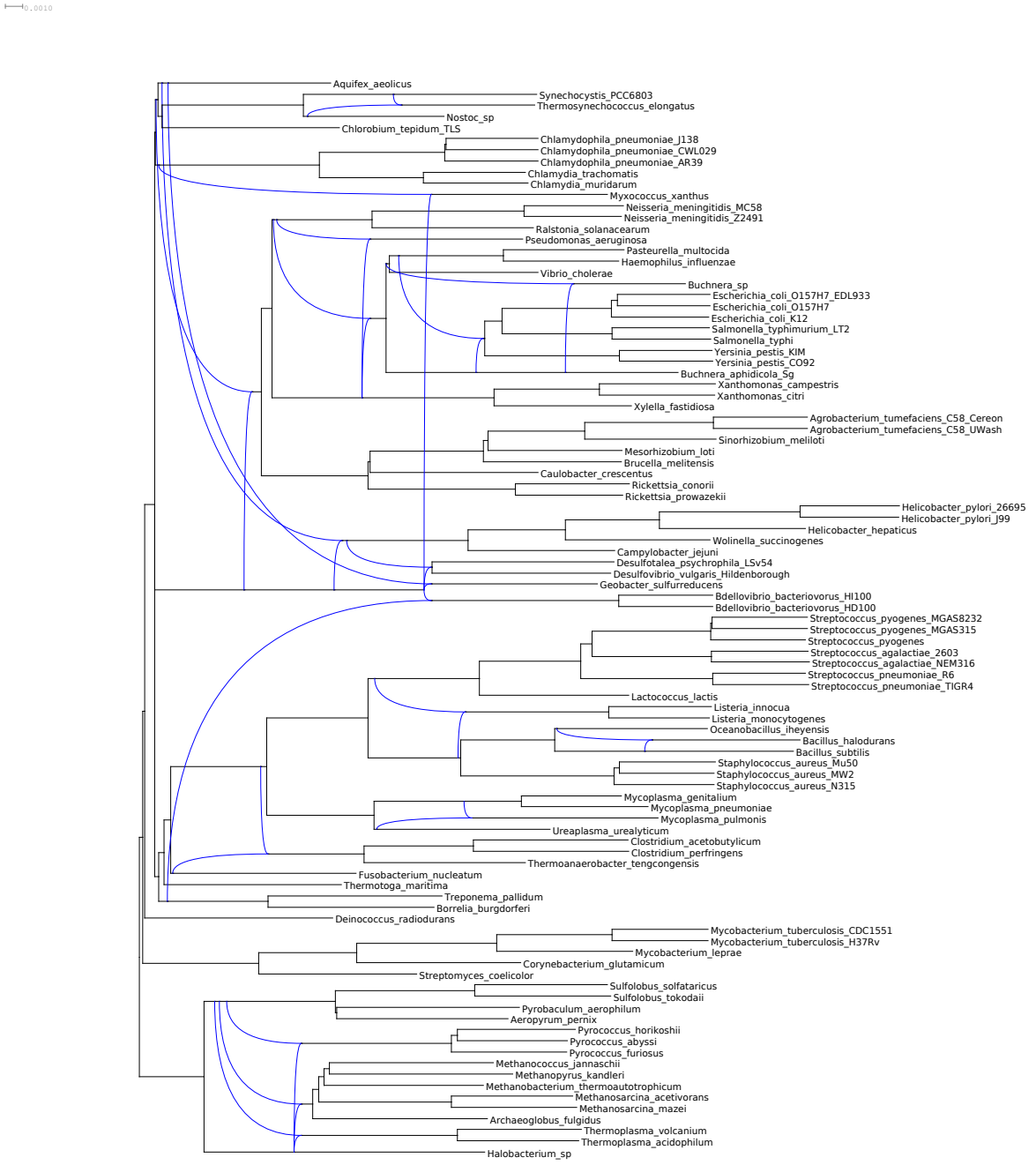


Figure 2.11: Cluster network view (see Huson et al. 2007c) of a `TBLASTX`-based tree using `GBDP` function `tr_6_bionj` and the corresponding NCBI tree. The blue edges indicate where a subtree is differently placed in both trees. The c-score of the tree was 0.8511.

as outgroup, we included genomes of two Rickettsia species as well as of two Wolbachia species, which also belong to the order Rickettsiales.

To get a balanced taxon selection, we decided to only include a single taxon in each case for the main lineages of the group Coelomata. This group includes the two phyla Arthropoda and Chordata, both represented by a high abundancy of sequences in the databases. The inclusion of more taxa from these phyla would have considerably increased run time, while leading to a bias of the mitochondrial dataset towards these metazoan lineages. This would have diminished its comparability with the plastidial dataset.

Altogether, we curated a plastidial dataset having 50 genomes, as well as a mitochondrial dataset comprising 125 genomes (including outgroup taxa).

`GBDP` functions and similarity search As variants of `GBDP`, we used distance functions `tr_0` to `tr_5` (see Table 2.2), i.e., matched distances (Equation 2.2), as well as homology based distances (Equation 2.9).

For similarity search, `BLASTN` and `TBLASTX` variants were used separately, as well as a combination of both approaches by merging resulting HSPs was produced. Additionally, combined matrices were derived by first applying a normalization function to each distance matrix, and afterwards calculating the average between the normalized `BLASTN` as well as `TBLASTX` matrices. Normalization is needed to fit the matrices to the same scale, which is of importance for logarithmic distance functions. For that purpose, the ranging method was applied (see Legendre and Legendre 1998, p. 38 and 252), by determining the minimum and maximum matrix values, and applying the following formula:

$$D_{normalized}(x, y) := \frac{D(x, y) - \min(D)}{\max(D) - \min(D)}$$

This leads to a distance in the range between 0 and 1.

Thus, two different approaches to combine nucleotide and protein level data were used. By combining datasets on the HSP level, the usually larger protein-based HSPs suppress the overall shorter, overlapping nucleotide HSPs. This can rather be seen as a slight enrichment of the `TBLASTX` hits, whereas in case of matrix averaging, both matrices are equally treated.

Regression analysis To quantify the influence of method selection on the c-score and δ value, we conducted a multiple linear regression using the `R` package (version 2.1.1, R 2008). Prior to the application of a regression analysis, we normalized the data to achieve a better comparability, by calculating z-scores for δ values and c-scores. The z-score is simply $\frac{x-\mu}{\sigma}$, where μ is the mean value, σ is the standard deviation, and x denotes the value to be normalized.

As explanatory variables, we used the following qualitative categories: Plastidial vs. Mitochondrial genomes; `BLASTN`, `TBLASTX`, combined HSPs, or matrix combining; minimum, maximum, or average of asymmetric distance values (see Section 2.2.1); non-logarithmic (Formula 2.11), or logarithmic distance conversion (Formula 2.12); applied distance functions 2.2 + 2.6, 2.2 + 2.7, 2.8, or the homology based function 2.9; and tree-reconstruction algorithms `UPGMA`, `NJ`, `BioNJ`, `FastME`, or `STC` for c-score as dependent variable.

The `R` package incorporates several important features. One such functionality is the automatic classification of qualitative variables into binary categories (see, e.g., Legendre and Legendre 1998, p. 46–47) that contain the same information, and are suitable for linear regression analysis. Additionally, `R` provides a step-wise elimination procedure based on the Akaike Information Criterion (AIC, Faraway 2002, p. 128–129). The elimination procedure starts with a model based on a complete set of describing variables, and calculates the according AIC score. In the following iterations, variables are successively removed and the AIC is re-calculated to find an optimal model. Optimality in the AIC sense is defined as a balance between model simplicity (i.e., the number of model parameters) and its likelihood (see also Section 3.2.1, page. 62).

Experimental results

Distance function assessment As a first step, the expressiveness of the δ value on phylogenetic reconstruction was examined by several linear regression analyses. The R^2 value of a linear regression represents the "percentage of variance explained" (Faraway 2002, p. 22), i.e., how much of the predicted variable's variance can be explained by the explanatory variables. Thus, a perfect prediction would induce a R^2 of 1.

To determine the influence of the δ value on the c-score, we made a regression using the δ value as explanatory variable. This yielded a R^2 of 0.617, which means that 61.7% of the variance in c-score can be explained by the δ value. Together with tree reconstruction methods, the δ value explains 62.1%, whereas including distance parameters, reconstruction methods and δ value explains 87.0% of the c-score variance. Thus, the largest part in the variance of the c-score can be explained by the δ value alone. According to this, the accuracy of reconstructed phylogenies is highly dependent on distance matrix quality as measured by the δ value. We conclude that the δ method is a valuable approach for distance quality assessment. In particular, it can be used in future research to select distance functions without having to rely on a reference tree or taxonomy.

Table 2.3 shows the results of a step-wise multiple linear regression with the c-score as dependent variable. Analogously, Table 2.4 shows the results of a regression analysis with the δ value as dependent variable.

	cscore (adjusted $R^2 = 0.775$)			
explanatory var.	coefficient	standard error	t value	$P(x > \|t\|)$
(Intercept)	0.032509	0.010895	2.984	0.00292
Plastids	0.123627	0.006290	19.655	$< 2 \cdot 10^{-16}$
BLASTN+TBLASTX	0.047477	0.008895	5.337	$1.18 \cdot 10^{-07}$
Matrix Averaging	0.024675	0.008895	2.774	0.00565
TBLASTX	0.044962	0.008895	5.055	$5.18 \cdot 10^{-07}$
Equ. 2.2 + 2.6	0.437622	0.008895	49.197	$< 2 \cdot 10^{-16}$
Equ. 2.2 + 2.7	0.380340	0.008895	42.757	$< 2 \cdot 10^{-16}$
Equ. 2.9	0.247183	0.008895	27.788	$< 2 \cdot 10^{-16}$
STC	-0.040611	0.009945	-4.083	$4.81 \cdot 10^{-05}$
UPGMA	-0.029744	0.009945	-2.991	0.00285

Table 2.3: Results of a step-wise multiple linear regression based on the AIC criterion (Auch et al. 2006b) for the c-score depending on all variables. Explanatory variables were: Plastidial vs. Mitochondrial genomes; `BLASTN`, `TBLASTX`, combined HSPs, or matrix combining; minimum, maximum, or average of asymmetric distance values (see Section 2.2.1); non-logarithmic (Formula 2.11), or logarithmic distance conversion (Formula 2.12); applying distance functions 2.2 + 2.6, 2.2 + 2.7, 2.8, or the homology based function 2.9; tree-reconstruction algorithms `UPGMA`, `NJ`, `BioNJ`, `FastME`, or `STC`. Only explanatory variables that were not eliminated by the step-wise optimization are shown.

	δ **value** (adjusted $R^2 = 0.888$)			
explanatory var.	coefficient	standard error	t value	$P(x > \|t\|)$
(Intercept)	0.533266	0.008668	61.522	$< 2 \cdot 10^{-16}$
Plastids	-0.172317	0.005779	-29.820	$< 2 \cdot 10^{-16}$
BLASTN+TBLASTX	-0.044311	0.008172	-5.422	$1.84 \cdot 10^{-07}$
TBLASTX	-0.041812	0.008172	-5.116	$7.82 \cdot 10^{-07}$
Equ. 2.12 (log)	0.043077	0.005779	7.455	$3.48 \cdot 10^{-12}$
Equ. 2.2 + 2.6	-0.184422	0.008172	-22.567	$< 2 \cdot 10^{-16}$
Equ. 2.2 + 2.7	-0.125118	0.008172	-15.310	$< 2 \cdot 10^{-16}$
Equ. 2.9	-0.088696	0.008172	-10.853	$< 2 \cdot 10^{-16}$

Table 2.4: Results of a step-wise multiple linear regression based on the AIC criterion (Auch et al. 2006b) for the δ value depending on all variables. Explanatory variables were: Plastidial vs. Mitochondrial genomes; `BLASTN`, `TBLASTX`, combined HSPs, or matrix combining; minimum, maximum, or average of asymmetric distance values (see Section 2.2.1); non-logarithmic (Formula 2.11), or logarithmic distance conversion (Formula 2.12); applying distance functions 2.2 + 2.6, 2.2 + 2.7, 2.8, or the homology based function 2.9. Only explanatory variables that were not eliminated by the step-wise optimization are shown.

For the δ value (Table 2.4), most influential parameters were using Equation 2.2, and plastidial genomes. This corresponds with the results obtained using the c-score as dependent variable. Here, Equation 2.9 also yields a positive effect. In contrast, breakpoint distances had the worst performance for both c-score as well as δ value. This is in agreement with the results in (Henz et al. 2005) based on prokaryotic phylogenies. We assume that the poor performance of breakpoint distances is caused by a deficiency in collinearity in the examined genomes. Breakpoint distances should only be used when there is a considerable amount of collinearity (Henz et al. 2005) between genomes, which only seems to be valid for closely related species.

Interestingly, using homology based distance functions based on Equation 2.9 leads to a marginal improvement of the c-score and δ value compared with functions based on Equation 2.2. This may be due to the fact that when comparing two distant genomic sequences, the similarity search algorithms may only find a small amount of highly conserved HSPs. In this case, the distance will be underestimated when applying Equation 2.9 because of the grade of conservation between the detected hits. On the other hand, more closely related taxa may also share regions having a lower grade of conservation, thus distance between those taxa will be overestimated.

In contrast to the results shown in Henz et al. (2005), using the length corrected denominator 2.7 together with Equation 2.2 lead to inferior c-scores and δ values for plastidial and mitochondrial genomes. However, using denominator 2.7 always led to a correct placement of the reduced genome of *Epifagus virginiana*. On the other hand, reconstruction based on Formula 2.6 incorrectly placed the taxon *E. virginiana* at the base of the Angiosperms. The species *Epifagus virginiana* is a parasitic plant that grows on the roots of beeches, and is commonly known as "beech drops". During its evolution, *E. virginiana*'s plastid lost all genes that are necessary for photosynthesis, but its preservation indicates that it still entails some important functionality for the organism (Krause 2008). *Epifagus* belongs to the phylum Streptophyta, subclass Asteridae, together with well-known plants like *Atropa belladonna* and *Nicotiana tabacum*. Accordingly, *E. virginiana* is correctly placed in Figure 2.12, unlike the incorrect placement in Figure 2.13.

The length corrected denominator 2.7 was specifically designed for cases where high rates of gene loss occurred in a lineage, as presumed for the evolution of *Buchnera* genomes (see Section 2.2.1, page 11). We conclude that uncorrected distances (i.e., based on denominator 2.6) are usually superior for reconstruction of plastidial and mitochondrial genomes. Denominator 2.7 should be preferred only when cases of extreme gene loss are observed.

Using a logarithmic distance transformation (see Section 2.2.1, page 13) leads to an increase of the δ value (with a moderate coefficient of 0.043, see Table 2.4). This means that the non-logarithmic distance transformation (Equation 2.11) leads to distances that deviate from the additivity condition

to a lesser extent. Usage of the logarithmic transformation (Equation 2.12) had no significant influence on the c-score. While logarithmizing has a small negative effect on the δ value, this effect seems to be too small to affect tree reconstruction. Actually, we assume that using logarithmizing may even improve topological accuracy in relation to taxa that underwent extreme genome modifications, like *E. virginiana* (Wolfe et al. 1992). Such effects can be seen when comparing Figure 2.12 (logarithmized) and Figure 2.13 (non-logarithmic) with regard to the placement and edge length of *E. virginiana*.

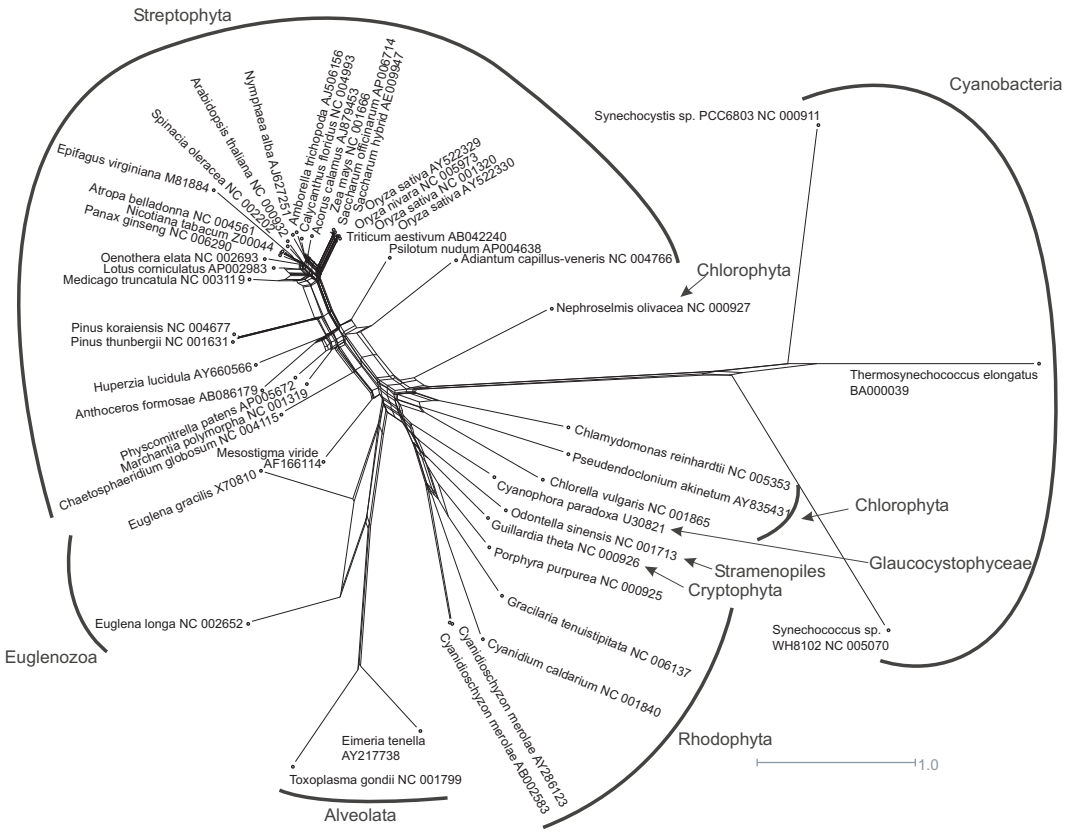


Figure 2.12: `NeighborNet` reconstruction using distance algorithm `tr_min2` (see Table 2.2) and `BLASTN` search within whole plastidial genomes. As outgroup, three cyanobacterial genomes were used.
Using a `BioNJ` reconstruction, this matrix resulted in the overall best tree according to the c-score (with respect to the NCBI taxonomy) of 0.8298 (Auch et al. 2006b). The corresponding δ value was 0.2013, indicating a high accordance with the additivity condition.

Distance correction by usage of minimum, maximum, or average matrix values due to the asymmetry of `BLAST` had no significant influence on the δ value, and only a rather small influence on the c-score (see Tables 2.3

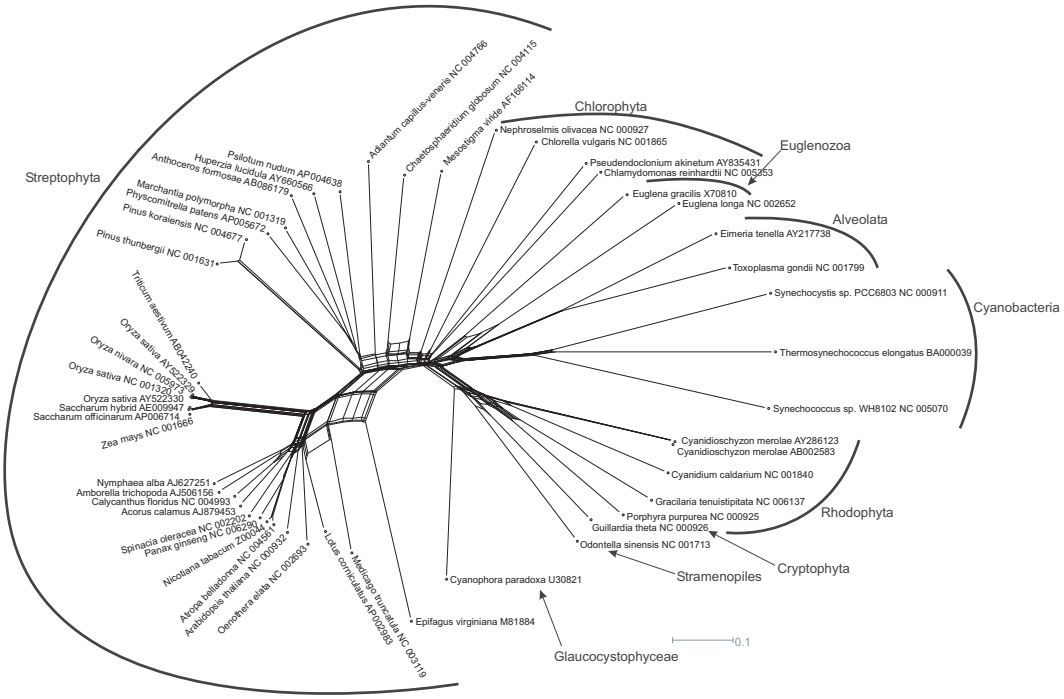


Figure 2.13: `NeighborNet` reconstruction using combined matrices derived by applying distance algorithm `tr_a0` (see Table 2.2) to `BLASTN` as well as `TBLASTX` within whole plastidial genomes.
This approach resulted in the lowest δ value (0.1629), thus giving the most tree-like distance data. However, the highest c-score was only 0.6596 using `BioNJ` or `STC` (Auch et al. 2006b). An interesting observation is the incorrect placement of *Epifagus virginiana* (see discussion).

and 2.4). Here, the coefficient for using matrix averaging was 0.025, which was the lowest observed coefficient of all significant variables. As already outlined, usage of the matrix averaging option thus seems to be sufficiently justified.

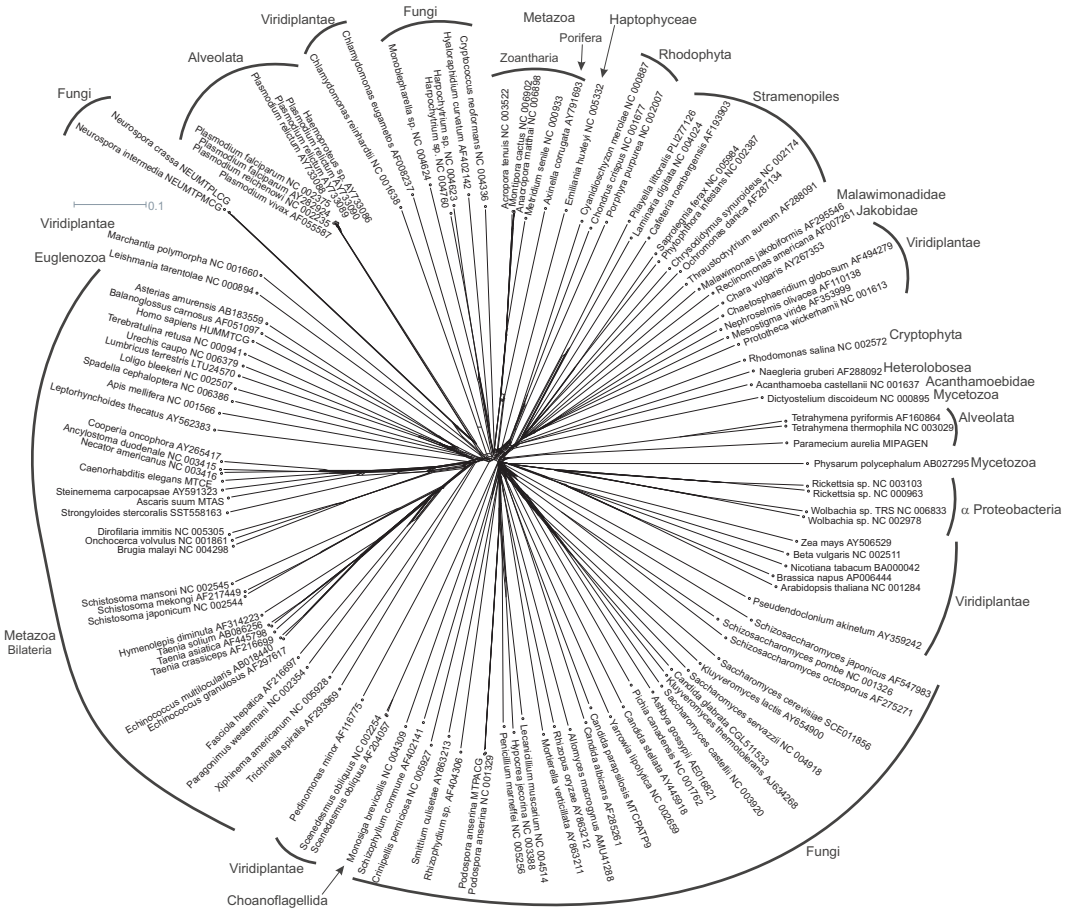


Figure 2.14: `NeighborNet` reconstruction using combined matrices derived by applying distance algorithm `tr_a0` (see Table 2.2) to `BLASTN` as well as `TBLASTX` within whole mitochondrial genomes.
This approach resulted in the highest c-score (0.5574) using `BioNJ`, whereas the δ value of the matrix was 0.2946, a relatively high quality compared to the other results using mitochondrial genomes. As outgroup, α proteobacterial genomes were used.

Influence of HSP search methods Table 2.3 indicates that using `TBLASTX` performs better than using `BLASTN` (coefficient 0.045), and combining nucleotide and amino acid data at the HSP level (`BLASTN` + `TBLASTX`, coefficient 0.047) slightly improves the results. Analogously, δ values also improve

when using HSP-combined data (see Table 2.4). As already mentioned in Section 2.2.2 (page 14), we assume that the difference between results obtained by using BLASTN and those using `TBLASTX`, originates from the greater sequence conservation at the protein level. But an enrichment of protein data by nucleotide data seems to lead to a slightly improved phylogenetic signal, as observed when using HSP-combined data.

However, the effect of HSP search methods on the δ value and c-score is far smaller than selection of the distance method or even the dataset itself (plastidial vs. mitochondrial). Considering run time differences, preference of the computationally feasible `BLASTN` algorithm, when dealing with large datasets, seems to yield a sufficient degree of accuracy.

Tree reconstruction methods Regression analysis indicates that `BioNJ` tree reconstruction produces the best trees according to their conformance with the NCBI taxonomy as measured by the c-score (see Table 2.5 and Figure 2.15). Regarding mean values of `BioNJ`, `FastME` and `NJ` reconstructions, the latter two algorithms do not perform considerably worse than `BioNJ`. Notably, whereas `UPGMA` is known to be sensitive to deviations from ultrametricity, this method performed best when distance quality was quite low. This can be observed for δ values above approximately 0.55 when regarding Figure 2.15.

Interestingly, performance of the `STC` tree reconstruction algorithm was severely impaired when using distance matrices having a bad δ value, whereas results obtained by `STC` were comparable to those obtained by using `BioNJ`, `FastME` and `NJ` when distance quality was high (see Figure 2.15).

c-scores	`BioNJ`	`FastME`	`NJ`	`UPGMA`	`STC`
mean	0.3899	0.3888	0.3790	0.3601	0.3601
best	0.8298	0.7660	0.7660	0.6170	0.8085

Table 2.5: Mean and best c-scores for applied tree reconstruction methods. Data from Auch et al. (2006b).

Phylogenies Overall, using plastidial data gave rise to better δ values and c-scores (see Tables 2.3 and 2.4). The c-score maximum was 0.8298 for plastidial (see Figure 2.12), and 0.5574 for mitochondrial data (see Figure 2.14).

It is well known that mitochondrial genomes dramatically differ in genome size and thus, in gene coding capacity (Gray et al. 1999). Whereas mitochondrial genome evolution in land plants tends towards an increase of genome size, the opposite can be observed for mitochondrial genomes in Metazoa (Lang et al. 1999a). Thus, a comprehensive taxon selection including plant as well as metazoan taxa, is hampered by insufficient homology between mitochondrial genomes. Furthermore, in metazoan mitochondria, a high

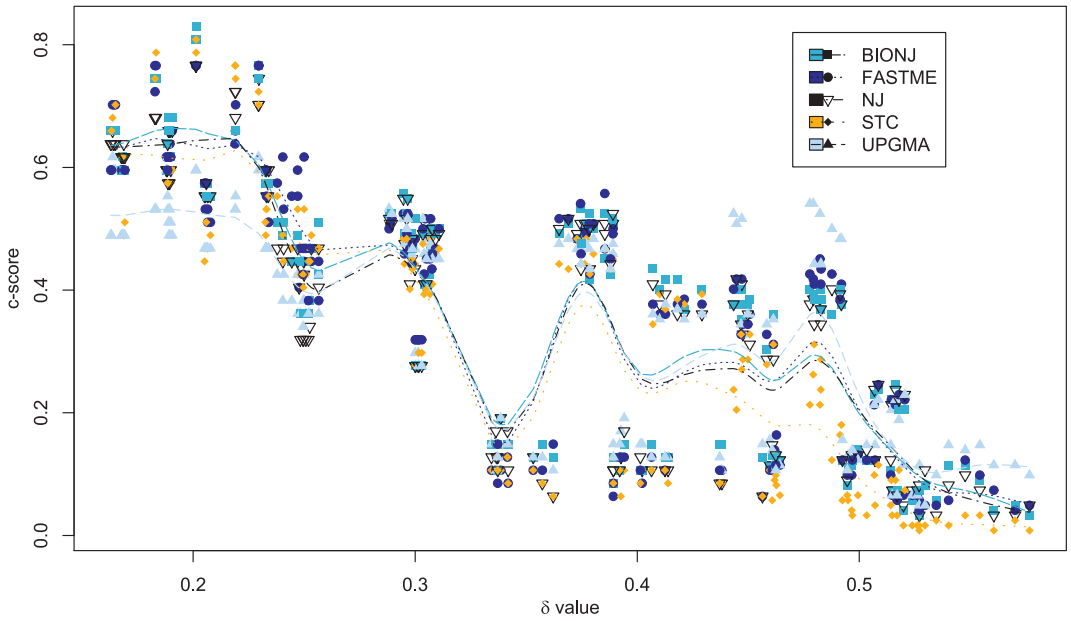


Figure 2.15: Comparison of distance functions and reconstruction methods (Auch et al. 2006b). The picture shows how δ values influence the corresponding c-scores when using different tree-reconstruction methods. Note that a low δ value indicates a high treelikeness of the underlying distance matrix, whereas a high c-score indicates a high level of correspondence of the accordant tree with the NCBI taxonomy. To illustrate characteristic trends of certain tree-reconstruction methods, cubic splines with 15 degrees of freedom were used to fit the data points onto a curve.

mutation rate leading to saturation complicates phylogenetic reconstruction within this group (Lang et al. 1999a).

On account of this, we conclude that the low backbone resolution, as seen in Figure 2.14, is no artefact of the GBDP method. This opinion is supported by the fact that reconstructed plastidial phylogenies highly resemble the NCBI taxonomy, which is usually based on nuclear genes.

A comprehensive analysis of the reconstructed phylogenies and deviations from the taxonomy is outlined in Auch et al. (2006b), to which the reader may be referred.

2.3.3 Extended Mitochondrial and Plastidial dataset

Experimental setup

Distance methods and similarity search Based on the study of Auch et al. (2006b), we extended the approach in several ways. First, we included the newly developed distance methods tr_a6 to tr_a9 derived from Equation 2.3 (see Table 2.2).

Second, we used BLAT (Kent 2002) in addition to BLASTN and TBLASTX. BLAT is several orders of magnitude faster than BLAST, but less sensitive.

Comparison with a reference taxonomy Additionally, we also incorporated an alternative to the c-score to measure the agreement of a tree with a reference taxonomy. Legendre and Lapointe (2004) introduced a test of congruence among distance matrices (CADM) that is based on the Mantel test of matrix correspondence (Mantel 1967; Mantel and Valand 1970). The test indicates whether matrices are congruent, i.e. whether they are related to each other in a way that permits combining them. CADM allows to calculate Spearman correlation coefficients (Legendre and Legendre 1998, p. 195–198) between distance matrices, which were used as agreement measure between matrices in this study. The Spearman coefficient varies between -1 and 1, whereas 1 indicates a perfect positive correlation. A coefficient of 0 would indicate absence of any correlation.

To apply the CADM test, the NCBI taxonomy, which was used as reference topology, was converted into a matrix of patristic distances. Then, a CADM score based on Spearman correlation was calculated between the distance matrix under investigation and the NCBI matrix. This method allows to directly measure agreement between the reference topology and the distances. Thus, it is independent of the chosen tree reconstruction method.

Topology-independent measures We also tested different topology-independent measures of distance quality, in addition to the δ value. Namely,

we used non-ultrametricity and non-additivity as described by Makarenkov and Legendre (2001), and De Soete (1986), as well as the ϵ value (see Section 2.2.5, page 24). Non-metricity is analogously defined as the square root of the sum of $(d_{ij} - d_{ik} - d_{jk})^2$ for all triplets of taxa i, j, and k in which $d_{ij} > d_{ik} + d_{jk}$, divided by the sum of all squared distances (SSD).

Experimental results

Distance functions and Homology search Table 2.6 shows the results of a step-wise regression analysis using the `R` package. The findings mainly correspond to the results of our previous study (see Section 2.3.2). All three metrics (c-score, ϵ value, and CADM score) coincide, that the most important factors for distance quality were usage of Plastidial data, and Equation 2.9. But in contradiction to previous results (see Tables 2.3, 2.4), Equation 2.9 led to distances that performed worse.

The newly included distance functions based on Equation 2.3 had a slight positive influence on distance quality as measured by the c-score (coefficient: 0.017). But admittedly, the result was not significant (P-value: 0.080).

Preferrence of using `BLAT` instead of `BLAST`, leads to a small but significant decrease in phylogenetic accuracy (see Table 2.6). While `BLAT` is much faster than `BLAST` (Kent 2002), it identifies only a subset of HSPs compared to the `BLAST` results. Thus we conclude that `BLAT` should not be applied if genomes are too distantly related, as is the case for mitochondrial genomes of the major eukaryotic groups (see also Section 2.3.2, page 44). Overall, we conclude that preference of `BLAT` over `BLAST` seems to lead to a trade-off between speed of similarity search and phylogenetic accuracy.

Figure 2.16 confirms the results of the previous study as shown in Figure 2.15. Although the `UPGMA` method performs supprisingly well in some circumstances, especially with low-quality distance matrices, results of the regression analysis indicate a negative impact on phylogenetic accuracy (see Table 2.6).

Distance quality metrics We compared several distance metrics according to their accuracy as measured by the c-score and CADM (test of Congruence Among Distance Matrices) score against the NCBI taxonomy (Table 2.7). Correlation between distance quality metrics and reference based metrics was tested by calculating the Pearson, as well as the rank-based Kendall and Spearman correlation coefficients (Legendre and Legendre 1998, p. 140, 199, and 195).

Results indicate that ϵ values, re-scaled Q values, and δ values performed best according to their correlation to the c-score, as well as to the CADM score. Whereas the difference between δ values and ϵ values consists only of the treatment of the case when both distances q and r are 0 (see Figure 2.6), a considerable improvement of correlation to the c-score and CADM was

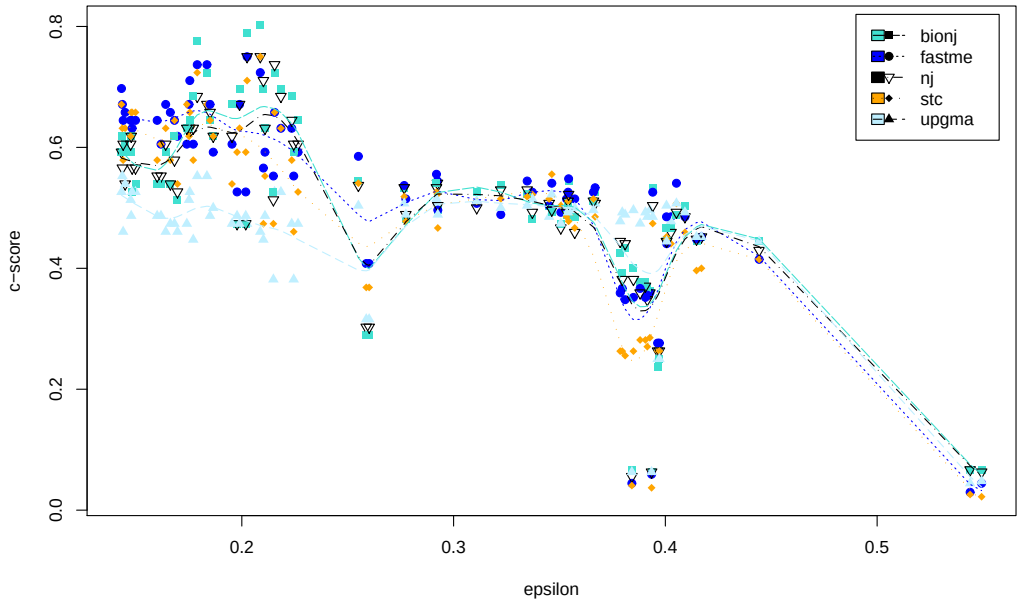


Figure 2.16: Comparison of distance functions and reconstruction methods (Auch et al. 2006a). Analogous to Figure 2.15, the picture shows how ϵ values influence the corresponding c-scores when using different tree-reconstruction methods. Note that a low ϵ value indicates a high treelikeness of the underlying distance matrix, whereas a high c-score indicates a high level of correspondence of the accordant tree with the NCBI taxonomy. To illustrate characteristic trends of certain tree-reconstruction methods, cubic splines with 15 degrees of freedom were used to fit the data points onto a curve.

explanatory var.	c-score (adj. $R^2 = 0.636$) coefficient	$P(x > \|t\|)$	ϵ value (adj. $R^2 = 0.859$) coefficient	$P(x > \|t\|)$	CADM-score (adj. $R^2 = 0.549$) coefficient	$P(x > \|t\|)$
Intercept	0.4510	$< 2 \cdot 10^{-16}$	0.3390	$< 2 \cdot 10^{-16}$	0.3366	$< 2 \cdot 10^{-16}$
UPGMA	-0.0540	0.0002				
Plastids	0.1371	$< 2 \cdot 10^{-16}$	-0.1744	$< 2 \cdot 10^{-16}$	0.1794	$1.59 \cdot 10^{-12}$
`BLAT`	-0.0303	0.0008	eliminated from model		eliminated from model	
translated	0.0843	$< 2 \cdot 10^{-16}$	-0.0347	0.0002	eliminated from model	
Equ. 2.12, (log)	not significant		0.0273	0.0027	eliminated from model	
Equ. 2.9	-0.2068	$< 2 \cdot 10^{-16}$	0.1003	$3.43 \cdot 10^{-12}$	-0.1506	$1.82 \cdot 10^{-6}$
Equ. 2.7, g_2	-0.0214	0.0324	0.0386	0.0002	not significant	

Table 2.6: Results of a step-wise multiple linear regression based on the AIC criterion for the c-score, ϵ value, and CADM score depending on all other variables (Auch et al. 2006a). Explanatory variables were: Plastidial vs. Mitochondrial genomes; `BLAST` or `BLAT`; using translated sequences or nucleotide sequences; non-logarithmic (Formula 2.11), or logarithmic distance conversion (Formula 2.12); applying distance functions 2.2, 2.8, the homology based function 2.9, or the newly included function 2.3; using denominators 2.6 or 2.7; minimum, maximum, or average of asymmetric distance values (see Section 2.2.1); tree-reconstruction algorithms `UPGMA`, `NJ`, `BioNJ`, `FastME`, or `STC`. Only explanatory variables that were not eliminated by the step-wise optimization are shown.

observed. According to that, we think that our definition of the ϵ value, with regard to the biological interpretation of distances, seems to be justifiable.

Overall, similarity between correlations with the c-score metrics and correlations with the CADM score in Table 2.7 could clearly be observed. This also shows that the influence of tree reconstruction methods is much lower than the influence of distance quality on the outcome of tree inference. Thus, the CADM score can be used to assess distance quality by comparing the distance matrices to a reference taxonomy, without having to compute trees from the matrices.

Phylogenies In accordance with the previous study, distances based on plastidial data led to better phylogenies according to the c-score (coefficient: 0.14) as well as the CADM score (coefficient: 0.18). Figure 2.17 shows a `NeighborNet` reconstruction based on the distance and HSP search method that achieved the highest c-score when using `BioNJ` tree reconstruction.

2.3.4 Large-scale study of 500 prokaryotic genomes

We downloaded all completely sequenced prokaryotic genomes that were available on NCBI (2009a) in July 2008. Strains of the same prokaryotic species were reduced to the type strain (in case that the type strain was indicated on the NCBI website), or a representative strain, which was selected

	c-score			CADM-score		
Metrics	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman
R^2	-0.420	-0.074	-0.129	-0.384	-0.022	-0.049
non-additivity; Q^2 [3]	-0.638	-0.351	-0.526	-0.710	-0.393	-0.565
$R - Q$ [0]	-0.028	0.220	0.295	-0.042	0.158	0.241
R [2]	-0.604	-0.212	-0.328	-0.601	-0.226	-0.356
Q [2]	-0.680	-0.399	-0.582	-0.751	-0.448	-0.649
R [1]	-0.631	-0.125	-0.189	-0.202	0.040	0.039
Q [1]	-0.730	-0.315	-0.451	-0.375	-0.164	-0.246
$(R - Q)^2$ [3]	-0.132	0.055	0.056	0.005	0.135	0.157
ϵ; Q/R	-0.701	-0.507	-0.712	-0.789	-0.510	-0.720
non-metricity [3]	-0.543	-0.302	-0.456	-0.634	-0.296	-0.432
δ; Q/R	-0.578	-0.404	-0.601	-0.746	-0.462	-0.664
R [0]	-0.062	0.194	0.263	-0.079	0.138	0.217
Q [0]	-0.147	-0.046	-0.079	-0.170	-0.106	-0.184
non-ultrametricity [3]	-0.459	-0.241	-0.379	-0.588	-0.268	-0.390
$R - Q$ [2]	-0.442	-0.132	-0.226	-0.370	-0.117	-0.201

Table 2.7: Correlation of Distance Quality Metrics (Auch et al. 2006a)
Non-metricity is defined as the square root of the sum of $(d_{ij} - d_{ik} - d_{jk})^2$ for all triplets of taxa i, j, and k in which $d_{ij} > d_{ik} + d_{jk}$, divided by the sum of all squared distances (SSD). $R - Q$ denotes the sum of $r - q$, and $(R-Q)^2$ denotes the sum of $(r-q)^2$. All other quality metrics are described in Section 2.2.5 on page 24.
Scaling formulae are [0], division by the total number of quartets; [1], division by the total number of quartets and the largest distance value in the matrix; [2] division by the square root of SSD; [3] taking the square root after division by SSD.

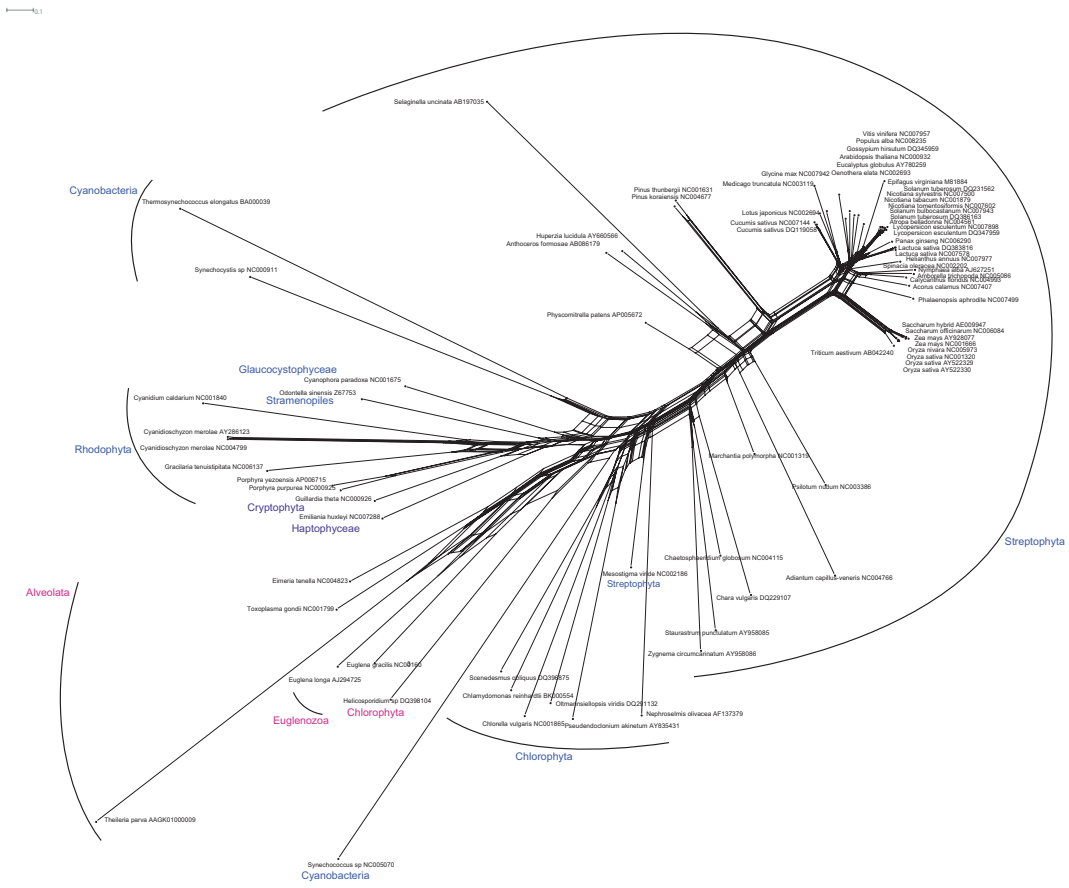


Figure 2.17: `NeighborNet` Plastid phylogeny based on distance method `tr_min9` (see Table 2.2) and a `BLASTN` search. This combination lead to the highest c-score of 0.790, using `BioNJ`. Data from Auch et al. (2006a).

manually. Eventually, a list of 515 prokaryotic genomes was assembled, consisting of 957 single files (chromosomes and plasmids).

`BLASTN` as well as `TBLASTX` runs were concurrently conducted on three different Cluster systems at the University of Tübingen, using the SQL Server infrastructure outlined in Section 2.2.3 (p. 17). Overall run-time was 6 months approximately, but not all Clusters or Cluster resources could be utilized over the whole time span. Most notably, the bwGRiD Cluster, which provided a noticeable speed-up, became available not until the last four weeks of the project.

To give an impression of the amount of data that had to be managed by the SQL server for the `TBLASTX` runs, we collected some statistical data about the database. The size of the `ta_file_stats` table containing the file names of the individual runs (see Appendix C.1) has 912,546 entries with a total size of 183 MB (index size included). The size of the BLOB data of the SQL table holding the data for the `TBLASTX` runs (`ta_file_streams`) is 674 GB, using the "lossy binary bzipped" encoding (see p. 19 and Table 2.1). Accordingly, using no compression at all would have resulted in more than 12 TB of data.

While data evaluation is still ongoing, some preliminary results are available. The best c-score (in regard to the NCBI taxonomy) obtained by using `BLASTN` was 0.6888 (`tr_a3 FastME`), whereas using `TBLASTX` resulted in a best c-score of 0.6426 (`tr_a6 FastME`). Interestingly, in contrast to prior observations, using `BLASTN` yielded better c-scores than using `TBLASTX`. A thorough analysis of the dataset is in preparation and will be published in the near future (Auch et al., in preparation).

2.4 Conclusions

In this chapter, we presented a framework for inferring whole genome phylogenies based on local alignment search tools like `BLAST`, `BLAT` or `BLASTZ`. Within this framework, several distance functions were implemented and evaluated using real-world datasets ranging from prokaryotic genomes to organelle genomes from mitochondria and plastids. Overall, a noticeable congruence between the inferred phylogenies and the NCBI taxonomy could be observed for the major groups of Bacteria, Archaea, plants, fungi, and animals. This may be seen as an indication for the presence of a predominant signal of vertical inheritance in these genomes, since the NCBI taxonomy is mainly based on the phylogeny of ribosomal DNA. Moreover, we can conclude that the `GBDP` approach is able to infer robust distance-based phylogenies based on organelle or prokaryotic genomes. The distance-based approach also allows to apply phylogenetic network algorithms like `NeighborNet` (e.g., see Figures 2.12, 2.13, and 2.14). Furthermore, by providing variance estimates, calculation of support values analogous to boot-

strap values can be accomplished using `WeightLESS` (Sanjuán and Wróbel 2005).

Additionally, we enhanced our framework for handling large-scale datasets. An efficient lossy compression algorithm was incorporated to deal with large amounts of `BLAST` outputs, and a SQL database was established for data storage. After providing these basic prerequisites, we were able to approach projects with several hundreds of taxa by utilizing Cluster resources, which are necessary to effectively parallelize local alignment searches. Hence, a large project with 515 prokaryotic genomes could be conducted successfully. Results will be published in the near future.

Besides phylogenetic reconstruction, distances derived by using the `GBDP` formulae can also be applied to other subjects. For example, `GBDP` distances have been successfully used as an alternative to DNA-DNA hybridization for species delineation. In principle, intergenomic distances could also be employed to detect horizontal gene transfer by comparing them to intergenic distances using Cook's distance, as proposed by Kanhere and Vingron (2009, see also Section 3.2.3, p. 65). Furthermore, an adaptation of `GBDP` for single loci (like, e.g., ITS or rRNA sequences) was developed and applied to Foraminifera SSU rRNA data (Göker et al., submitted to *BMC Biology*, see p. 140).

In the future, the `GBDP` approach may be enhanced by providing an interface to the SIMAP database (Similarity Matrix of Proteins, Rattei et al. 2008). Thus, the computational demanding step of `BLAST`ing translated genomic sequences against each other could be omitted by incorporating the protein alignment data as already provided by the SIMAP database (T. Rattei, pers. comm.).

Furthermore, improvements could be made in refining the variance estimation process by considering the asymmetry between distance values when interchanging subject and query sequences in a `BLAST` search. Also, an adaptation of MSA-based jackknifing (Felsenstein 2004, p. 339-342) to the HSP level could be used as a resampling strategy. Here, a random fraction of HSPs could be taken as a new sample. This would allow to calculate support values based on the generated HSP subsamples by calculating `GBDP` distances and inferring trees for each subsample.

Moreover, refinement of the distance functions may be the subject of future investigations.

Chapter 3

Detection of Horizontal Gene Transfer in Prokaryotes

3.1 Introduction

In this Chapter, we present an automated approach for the classification of genes into groups of orthologs and the detection of horizontally transferred genes within these groups. Detection of HGT (horizontal gene transfer) is performed using three methods that are based on phylogenetic tree inference but use different strategies for tree comparison: Two statistical approaches are incorporated, namely `AxParafit` (Stamatakis et al. 2007) and Cook's distance (Kanhere and Vingron 2009), as well as an efficient topological algorithm that was developed by our group.

The outcome of the different HGT detection methods is compared using a dataset of genes shared between 279 prokaryotic organisms. But first, some important biological concepts are outlined in the next sections.

3.1.1 Horizontal gene transfer and phylogenetic reconstruction

Many phylogenetic approaches are based on the study of a single gene or locus as phylogenetic markers such as 16S rRNA or other informational genes (i.e., genes that are involved in transcription and translation, see Jain et al. 1999). However, an analysis limited to a single source of phylogenetic signal disregards the fact that each gene has its own evolutionary history which can be different from the "true" species phylogeny (Rokas et al. 2003; Gadagkar et al. 2005). Furthermore, conflicts in the phylogeny of genes diverging from the species phylogeny are ignored. Such conflicts may be caused by horizontal gene transfer (HGT), gene duplication events or homologous recombination (Inagaki et al. 2006). Even for 16S rRNA there is

evidence for occasional horizontal gene transfer (Yap et al. 1999; Koonin and Wolf 2008), and recent findings indicate that a core set of prokaryotic genes may also be affected by horizontal gene transfer (Dagan and Martin 2007). Furthermore, the work of Sorek et al. (2007) indicates that there may be no gene family that is entirely untransferable. Thus, the question whether the true species phylogeny can be reconstructed at all is strongly associated with an estimation of the amount of horizontally transferred genes among species (Doolittle 1999b; 2000). Kunin et al. (2005) investigated the relative contribution of HGT to prokaryotic phylogeny, but the number of HGT events during prokaryotic evolution is still discussed controversially (Beiko et al. 2005; Ge et al. 2005; Lerat et al. 2005; Wiezer and Merkl 2005; Dagan and Martin 2006; Choi and Kim 2007; Dagan et al. 2008; Koonin and Wolf 2008).

In Chapter 2, we focused on whole-genome based phylogenetic reconstruction methods to infer reliable phylogenies that are not affected by the presence of horizontal gene transfer. Even in the light of HGT, this remains feasible as long as it can be assumed that HGT is no "rampant" factor shaping the evolution of prokaryotes (see, e.g., Kurland et al. 2003; Puigbò et al. 2009). In the previous chapter, we concluded that the phylogenetic signal derived from analyzing whole-genome data is dominated by vertical inheritance. Still, the question about the amount of HGT in prokaryotic genomes remains unanswered.

Basically, there exist three different approaches for detecting horizontally transferred genes. Compositional methods try to use signatures based on GC content, nucleotide or word count frequencies to detect genes that deviate from the average genomic composition (e.g., Dufraigne et al. 2005). In contrast, phylogenetic methods rely on the comparison between gene trees and organismal trees to detect deviations that can be interpreted as HGT (Ge et al. 2005; Poptsova and Gogarten 2007). A third method uses `BLAST` similarity searches to find the closest hits for a group of genes (Podell and Gaasterland 2007; Podell et al. 2008). If the closest hit has a larger taxonomic distance as expected, the corresponding gene is assumed to be derived via horizontal transfer. However, these methods all have strengths and weaknesses, and it has been assumed that these methods may detect different types of horizontal gene transfer events (Ragan 2001; Lawrence and Ochman 2002; Ragan et al. 2006). Compositional methods are better suited for the detection of recent genetic transfers, since genes derived via HGT assimilate to the composition of their host genome over time (see, e.g., Gophna et al. 2006). Though, some genomic regions have a different base composition due to functional constraints, which could be misinterpreted as HGT. However, phylogenetic methods are able to detect ancient gene transfers, but tree comparison is difficult to automate and a computationally demanding task. Furthermore, phylogenetic methods are computationally more challenging than compositional methods. The accurracy of similar-

ity search based methods mainly depends on the accurracy of the local alignment search program. However, using only the best `BLAST` hit can be misleading (Eisen 2000; Koski and Golding 2001).

3.1.2 The concept of homology

Reconstruction of phylogenetic trees is only reasonable if the features that are used to infer the phylogenies (i.e., morphological traits or sequence data) share a common ancestry. Consequently, distinguishing between homologies (common ancestry) and analogies (similarities without a relationship based on common descent) is a crucial step prior to phylogenetic inference. The concept of homology can be applied to morphological traits as well as to sequence data (e.g., genes). Accordingly, homologous genes are genes that originate from a common ancestor gene.

Homologous genes can further be divided into orthologous (orthologs) and paralogous genes (paralogs). Orthologs are genes that are derived from a single gene in a common ancestor, whereas paralogs emerge from gene duplication. It is assumed that orthologous genes fulfill the same biological functions (Koonin 2005) in the corresponding organisms. In contrast, paralogs, which are derived by gene duplication instead of vertical inheritance, do not share the same evolutionary constraints (like purifying selection, which may be constricted to a single gene copy) and thus, may be subject to positive selection, which may lead to the acquirement of new functions (Kondrashov et al. 2002; Koonin 2005). Consequently, phylogenetic inferrence should be based on orthologous genes. Otherwise, phylogenetic reconstruction may be misleading when searching for horizontally transferred genes (see Eisen 2000).

However, finding clusters of orthologous genes is a challenging task, especially the detection and distinction of out-paralogs and in-paralogs (Remm et al. 2001). Whereas in-paralogs, i.e. paralogs that arose after speciation, pose no problem for phylogenetic reconstruction, out-paralogs, gene duplications preceding speciation, can easily be misjudged as events of horizontal gene transfer. Hence, prior to phylogenetic reconstruction, a method has to be applied that is able to distinguish between orthologs (and in-paralogs) on the one hand, and out-paralogs on the other hand.

3.2 Methods

3.2.1 Detecting a common set of orthologous prokaryotic genes

One of the best known databases of orthologous genes is the COG database at NCBI (Tatusov et al. 1997; 2000; 2003). In March 2009, the database contains 192,987 proteins from 66 unicellular genomes (NCBI 2009b), compris-

ing only a small subset of currently sequenced prokaryotic genomes (approx. 700, see Section 2.3.4 for recent estimates).

Thus, to get a more recent sampling of prokaryotic diversity, we considered to build a set of orthologous genes on our own initiative.

In order to derive a set of orthologous gene clusters, we decided to use an approach that is mostly independent of any species tree hypothesis, and that is able to remove in-paralogs as well as out-paralogs. The latter requires using taxonomical information for discarding paralogs that conflict with a tree induced by the taxonomy.

Taxon sampling

At the beginning of this study (November 2006), more than 300 fully sequenced prokaryotic genomes were available. By selecting only a single representative for each species having more than one sequenced strain, we finally compiled a list of 279 taxa. Using different strains of the same species would not have brought any benefit regarding phylogenetic diversity, but a needless prolongation of computing time.

For all of these taxa, annotated protein sequences were downloaded from the NCBI genome database (NCBI 2006). Eventually, we compiled a library consisting of 856, 535 protein sequences, on average, 3070 genes per genome.

The smallest genomes in the study were the genome of *Nanoarchaeum equitans* (536 genes, 490 kbp), *Mycoplasma genitalium* (477 genes, 580 kbp) and *Buchnera aphidicola* (504 genes, 620 kbp). By including these parasitic organisms, the amount of potential common genes that can be detected will be considerably diminished. However, the aim of this study was to find a set of common genes large enough for proper phylogenetic reconstruction of a species phylogeny, as well as the detection of horizontally transferred genes within this set, but not to find a comprehensive core set of prokaryotic genes.

Furthermore, the phylogeny of parasitic organisms like Mollicutes, Enterobacteriales or the genus Wolbachia is a matter of particular interest (Sirand-Pugnet et al. 2007a;b; Herbeck et al. 2005; Bordenstein et al. 2009). Therefore, exclusion of species with a shrunken genome would greatly diminish the value of our investigation, especially when regarding HGT events in parasitic genomes.

Ortholog detection using OrthoMCL

The first step in orthology detection consists of a `BLAST` search (Altschul et al. 1990) for sequence similarity between all genes in the library. This task was parallelized by using a computing cluster available at the "Zentrum für Datenverarbeitung", University of Tübingen (ZDV 2009). We used `NCBI-BLAST` version 2.2.17, with an e-Value cutoff threshold of 10^{-5} and soft

filtering. The latter option prevents HSPs from breaking apart if a region of low complexity is detected (e.g., a repetition of the same short pattern or even of a single amino acid), without losing the benefit of shorter runtime due to low complexity filtering. Using these parameters, we obtained 184, 966, 875 High-scoring segment pairs.

For further analysis of the `BLAST` results, we applied OrthoMCL (Li et al. 2003), a state-of-the art method for unsupervised orthology assignment (Chen et al. 2007; Altenhoff and Dessimoz 2009). The current distribution of OrthoMCL comes as a Perl script and includes the MCL binaries (Markov clustering algorithm, see Enright et al. 2002). On an environment equipped with 64 MB RAM, we were able to execute OrthoMCL with such a large dataset. The calculation took approximately 4 weeks on an AMD Opteron 2 GHz processor, using a single core since no parallelisation was done in the Perl script.

After clustering the results with MCL, we selected all clusters having at least one gene included from each single species. We repeated this step using different values for the inflation parameter between the default value of 1.50, and 3.0. Since we could not detect a setting that enabled MCL to separate all paralogs from the clusters, i.e. by producing consistent clusters having exactly one gene from each species, we used the default value of 1.50. Using the default value for the inflation parameter is supported by recent studies indicating that this parameter has only a small influence on the accurracy of OrthoMCL (Chen et al. 2007; Li et al. 2003). The default may already constitute an optimal trade-off between sensitivity and selectivity (Li et al. 2003).

Multiple sequence alignment

Multiple sequence alignments (MSA) were generated using the version of Muscle 3.6 (Edgar 2004a;b) improved by using the "nralign" algorithm, which considers the alignment of neighbouring residues in its scoring function (Lu and Sze 2009). Lu and Sze (2009) showed that their algorithm improves alignment quality mostly independent of the level of sequence identity. This fits well when using a large dataset showing quite different kinds of degrees of relationships between organisms. We decided to use Muscle to maintain a reasonable balance between high accurracy and speed. Resarch of Edgar (2004c) indicates that the average accurracy of Muscle alignments is similar to that of T-Coffee (Notredame et al. 2000) and clearly outperforms ClustalW (Thompson et al. 1994) and MAFFT (Katoh et al. 2002), whereas Muscle's execution time is lower.

To further improve alignment quality, we used `Gblocks` (Castresana 2000) to automatically clean the resulting MSAs. `Gblocks` is a program that deletes ambiguous alignment positions from a multiple sequence alignment, while it tries to preserve columns with significant phylogenetic signal. This

is achieved by using several criteria like the level of sequence conservation in a column, the amount of gaps and minimal length for blocks of conserved positions. A recent study of Talavera and Castresana (2007) shows that the impact of MSA quality improvement by using `Gblocks` is significant for protein-based sequences, especially when the alignments are heterogenous. In the latter case, even Maximum Likelihood based reconstruction methods seem to be hampered by using information from misaligned sites. This supports our view to use `Gblocks` to purify the MSAs prior to phylogenetic reconstruction.

Parameters for `Gblocks` mainly followed recommendations of Talavera and Castresana (2007) for "relaxed" settings, which seem to be better suited for ML-based tree reconstruction than more stringent settings. Furthermore after cleaning, alignments must remain long enough to bear sufficient information for the phylogenetic reconstruction of deep relationships. Hence, this supports the use of more protective settings. In detail, we set "Minimum Number Of Sequences For A Conserved Position" and "Minimum Number Of Sequences For A Flank Position" to $\lfloor \frac{n}{2} \rfloor + 1$, whereas n denotes the amount of sequences (which can be different if more than one gene per species is included in a cluster). The parameter for "Maximum Number Of Contiguous Nonconserved Positions" was set to 12, a value slightly larger than the recommended setting of 10. "Minimum Length Of A Block" was changed to 4 (recommendation: 5). Additionally, we decided to retain all positions with gaps. The recommended setting was to only retain positions consisting of no more than 50% gap characters.

Our tests indicated that these settings allowed us to keep alignment length in an acceptable range (see Table 3.1 and discussion in Section 3.3.1). We decided to also include heavily gapped columns because we assume that such columns may bear a phylogenetic signal on lower taxonomic levels (i.e., closer to the leaves). This helps to enhance resolution between family or genus members sharing a homologous block that may not be present in more distant taxa.

Phylogenetic reconstruction

Phylogenetic reconstruction of diverse samples of species has to cope with problems like saturation and heterotachy (differences in evolutionary rates of certain positions among lineages), which may lead to long branch attraction (LBA) artifacts. LBA was first described by Felsenstein (1978) for Maximum Parsimony (MP), but it also affects other reconstruction methods, like Maximum Likelihood (ML, Huelsenbeck 1995; Brinkmann et al. 2005). However, simulation studies showed that ML is more robust against LBA than MP (Kuhner and Felsenstein 1994; Swofford et al. 2001) and distance-based approaches like NJ (Huelsenbeck 1995).

Even in case of heterotachy, recent work of Philippe and co-workers (Philippe et al. 2005) indicates that ML is superior to MP in analyzing real world datasets. It should be mentioned that this result is in contradiction with the work of Kolaczkowski and Thornton (2004). The authors explicate the intrinsic advantage of MP as a non-parametric method over approaches based on assumptions of a distinct evolutionary model, like ML and Bayesian methods. In contrast, Philippe et al. (2005) and Spencer et al. (2005) pointed to some limitations in the validity of this study, which seems to be solely valid for a particular simulation scenario. Furthermore, Steel (2005) argues that the type of heterotachy investigated by Kolaczkowski and Thornton (2004) might not correspond to any biochemical mechanism known to date and thus, it seems to be only a theoretical setting. His opinion was countered by a follow-up paper of Thornton and Kolaczkowski (2005), stating that heterotachy is not sufficiently explored in real world datasets to allow to completely refuse their scenario. The discussion of this topic is ongoing, though many scientists agree that ML seems to be relatively stable in handling heterotachy in real world datasets when choosing an appropriate evolutionary model (Gadagkar and Kumar 2005; Lockhart et al. 2006).

Combining these facts and since ML seems also to be more robust against alignment ambiguity than Maximum Parsimony or distance-based approaches (Talavera and Castresana 2007), we decided to use a Maximum Likelihood based reconstruction method for our study.

However, development of more realistic models of evolution remains a hot and controversial topic (Steel 2005; Thornton and Kolaczkowski 2005). Consequently, we base our work on what is currently available and well-established in literature. Basically, there exist two such solutions for modelling unequal rates of change at different positions. One approach is based on random drawing of rate values independently for each site from a distribution. For this purpose, a Γ distribution as introduced by Uzzell and Corbin (1971) and adapted by Yang (1993) for ML-based tree-inferrence is widely used, due to its mathematical tractability. By specifying the shape parameter α, the corresponding Γ distribution can be used to approximate many other distributions, like an exponential distribution ($\alpha = 1$) or even a Gaussian distribution for large values of α. But there is no biological reason for favouring this function so far (Felsenstein 2004, p. 219).

The other major approach consists of using a fixed number of rate categories, which is far smaller than the actual amount of positions in the alignment. Constricting the number of rate categories is meant to avoid over-fitting, which is a formidable problem in model selection (Yang 1996; Steel 2005). Whereas the Γ model has a high demand of memory and run time, there exists an implementation of a fixed rate category model explicitly optimized for large datasets (Stamatakis 2006a, referred to as CAT model). Stamatakis (2006a) shows by evaluating empirical datasets that CAT can be used as a replacement for the Γ model.

We decided to use RAxML 7.04 (Stamatakis 2006b) as one of the fastest ML implementations available to date. RAxML provides parallelized versions capable of utilizing multi-threading environments (Ott et al. 2007) as well as MPI (Message Passing Interface, see Gropp et al. 1999) for interprocess communication, which makes this tool well-suited for cluster environments. Additionally, it has some unique advantages like its incorporation of the (comparatively) computationally undemanding CAT model, and a new rapid bootstrapping algorithm (Stamatakis et al. 2008), which makes RAxML more than an order of magnitude faster than other popular applications.

For selection of an appropriate model, we used ProtTest 1.4 (Abascal et al. 2005). ProtTest uses a modified version of phyml (Guindon and Gascuel 2003) for inferring ML trees and compares their log likelihoods based on the well-known Akaike Information Criterion (AIC, Posada and Crandall 2001). The AIC (see also Legendre and Legendre 1998, p. 520-521) tries to penalize models having a large number of parameters and thus can be used to avoid over-fitting of the model. This approach is based on the principle of Ockham's razor, which is well-established in the philosophy of science. William of Ockham (who lived and worked most of his time in Munich actually) formulated the *lex parsimoniae*: "pluralitas non est ponenda sine necessitate" (plurality should not be posited without necessity), meaning that when one has to choose among different hypotheses, each explaining the facts (almost) equally well, then the most parsimonious hypothesis should be preferred, i.e., the explanation needing the fewest assumptions (or parameters).

The AIC is defined as follows (Felsenstein 2004, p. 316), whereby p denotes the number of parameters, and L denotes the likelihood of the tree under investigation:

$$\mathrm{AIC} = -2 \ln L + 2p \tag{3.1}$$

When LnL decreases or the amount of parameters is growing, the AIC increases and vice versa. Trying to find a minimal AIC is thus the goal for model optimization. The addition of $2p$ can be seen as a penalty for introducing parameters, leading to a balance between model simplicity and its likelihood.

Another established method for model selection is the Bayesian Information criterion (Schwarz 1978), with n indicating the sample size (Felsenstein 2004, p. 316):

$$\mathrm{BIC} = -2 \ln L + p \ln(n) \tag{3.2}$$

This formulation leads to an increased penalty for models with many parameters when the sample size is large.

We decided to use the AIC as selection criterion, because it is preferred in the relevant literature (Felsenstein 2004, p. 316). Furthermore, using the

BIC would require to determine sample sizes of the given MSAs, which is considered a controversial topic (Abascal et al. 2007).

Since the calculation of likelihoods is partly depending on the concrete implementation, we compared the ProtTest results with those obtained from using the Perl script `ProteinModelSelection.pl` published on the `RAxML` homepage (Stamatakis 2009). The script uses `RAxML` to find the model with the highest likelihood score by using the same starting tree for each invocation.

However, for the actual tree inference we used the PROTMIX mode of `RAxML`, which uses the CAT model for searching a good tree topology, but afterwards switches to the Γ model to determine tree likelihood scores. The study of Stamatakis (2006a) indicates that this approach performs better in many cases than a direct search under the Γ model, while remaining computationally feasible using large datasets.

AxParafit and removal of remaining out-/in-paralogs

Since it can not be ruled out that the OrthoMCL-based clusters contain out-paralogs to some extent, we needed a method to purify the datasets. Consider the situation when two (or more) genes from the same species are included in a single cluster. This implies that at least one of the genes must be derived either by gene duplication or via horizontal gene transfer. Affected genes can be detected by reconstructing a phylogenetic tree and by comparing the placement of the genes belonging to the same species. In case of in-paralogy, e.g., gene duplication after speciation, the two genes should be nearest neighbours in the tree. Selecting the best candidate can be accomplished by looking at edge lengths and taking the one with the shortest edge length. This is based on the hypothesis that one of the gene copies should remain under purifying selection, and thus it should have a smaller evolutionary distance to the remaining genes.

When the genes are not adjacent in the tree, we can compare the positions to a species phylogeny and choose the candidate that is correctly placed. This method is well-founded, since the whole definition of orthology is based on the assumption that the phylogeny of orthologous genes reflects the species phylogeny (Altenhoff and Dessimoz 2009; Koonin 2005), at least in absence of any HGT events. Certainly, the worst case would be that both genes deviate from the species phylogeny. In that case, we would have to discard the whole cluster.

In absence of a self-derived hypothesis about the species phylogeny (which cannot be established with multi-labeld trees due to ambiguities in some clusters), we used a tree based on the NCBI taxonomy (NCBI 2009c) to exclude paralogs from further investigation. Note however that this method is not a variant of tree reconciliation (Mirkin et al. 1995), since the species phylogeny is only used to remove ORFs (Open Reading Frames) that deviate

from the species tree. Horizontally transferred genes can thus be preserved when there exists no other homologous gene in the same genome that is more appropriately placed in the corresponding gene tree. Particularly, replacements of genes by a gene copy derived from another species (i.e., xenologous gene displacements) can thus be recognized by this method. In summary, our method combines clustering methods and phylogenetic methods in a way that allows to conserve HGT events leading to gene replacement.

We used a statistical test for host-parasite cophylogeny, named `ParaFit` (Legendre et al. 2002), to compare tree topologies of gene trees and species trees in case that there exists more than one single candidate for orthology. With this method a hypothesis of cophylogeny between two sets of species can be tested. This approach is outlined in more detail in Chapter 4. `ParaFit` is especially suited for comparing the trees, because it is able to handle one-to-many associations (i.e., more than one homolog in the same species). Furthermore, it calculates the overall significance as well as significance values for each association between gene and species. More precisely, the reconstructed gene phylogeny is used as parasite tree, whereas the species tree based on the NCBI taxonomy is used as host tree. For each gene, an association is assumed between the gene and the species to which it belongs in the species tree.

To subsume the main types of events affecting historical host-parasite associations, there exist four possibilities (see Figure 4.1, page 99), whereby parasites denote the cluster genes, and hosts refer to the species:

- *cospeciation*, which is comparable to orthology
- *duplication*, meaning speciation of parasites only, which corresponds to paralogy
- *lineage sorting*, the loss of a parasitic species and its link
- and *host switching*, which can be seen as a HGT event

Since there is an obvious correspondence between these events and orthology/paralogy (see Section 4.1.1, p. 98, as well as Page 1994; Page and Charleston 1998), we think that this method can be used to prune datasets from paralogs.

For this purpose, we used `AxParafit` (Stamatakis et al. 2007) and the recently developed command line enabled version of `CopyCat` (Stockinger et al. 2009) to infer significance values for the gene/species associations. For details, the reader may be referred to Sections 4.2.2 (p. 105) and 4.2.4 (p. 108). In case of more than one association for a given species, we selected the gene having the highest $F1$ score (Legendre et al. 2002, formula 4). This identifies the association that contributes to a greater extend to the global Host-Parasite (or rather Species-Gene) relationship.

3.2.2 Species phylogeny

In order to determine horizontal gene transfer in gene trees, many methods refer to an organismal (or species) phylogeny as comparison. Currently, two approaches are commonly used to infer an organismal phylogeny from single gene trees: the supermatrix method (e.g., Rokas et al. 2003; Ciccarelli et al. 2006; Smith et al. 2009) on the one hand, as well as the consensus tree methods (Swofford 1991; Bryant 2003) and supertree methods (Bininda-Emonds 2005) on the other hand. The first method is based on concatenating single gene alignments to a large alignment and tree reconstruction based on the whole concatenated alignment. In contrast, the consensus and supertree methods try to combine several gene trees based on single gene alignments into a single tree, whereas supertree methods are able to combine trees having only partially overlapping taxa sets.

In this study, we favoured the supermatrix approach, since the contribution of every single gene in the supermatrix implicitly depends on its alignment length. This means that genes having only a low amount of phylogenetic signal due to their limited sequence length contribute to a lesser extent to the resulting tree than long genes. Furthermore, a simulation study performed by Gadagkar et al. (2005) indicates that concatenation leads to more accurate phylogenies than the consensus tree method.

Single gene alignments were concatenated and a species tree was inferred using `RAxML` 7.04. For each alignment region (i.e., each single gene alignment), an individual substitution model was specified by using `RAxML`'s multiple model file feature.

As an alternative to the supermatrix tree, a reference tree based on the NCBI taxonomy was included in the analysis. Additionally, a refined (binary) reference tree was generated by using the NCBI tree as a constraint when applying `RAxML` to the concatenated alignment.

3.2.3 Detection of horizontal gene transfers

Cook's Distance

Kanhere and Vingron (2009) proposed a statistical measure based on the comparison of phylogenetic distances between taxa, which are commonly represented by corresponding sequences. The basic idea is to apply a distance calculation process (`protdist` or `dnadist` from Felsenstein's phylip package) to infer distances between all single gene sequences. After this, the distances are compared against genomic distances by calculating Cook's distance (Cook 1979) between both measures.

Cook's distance (CDISS) provides an estimate of the influence a data point has on a linear regression. The linear regression model is based on the pairing of corresponding intergene and intergenomic distances. For each taxon, a mean CDISS is calculated and compared against a given threshold.

Taxa having a mean CDISS above this threshold are considered as candidates of horizontal gene transfer. As cut-off, Kanhere and Vingron (2009) proposed a distance of $\frac{2}{D}$, where D denotes the number of paired data points.

Phylogenetic distances can be seen as estimates of the sum of the branch lengths between two species (Felsenstein 2004, p. 147). On this account, we used the patristic distances between taxa of the gene trees as phylogenetic distances. A comparison of the results obtained by using patristic distances with those obtained by directly using the ML distances, resulted in no significant differences (data not shown).

As a source for genomic distances between taxa, we used patristic distances obtained from the species trees based on the concatenated alignment, and the NCBI tree.

Parafit statistical test

As outlined in Section 3.2.1, a statistical test for host-parasite cophylogeny as implemented in `ParaFit` (Legendre et al. 2002) and `AxParafit` (Stamatakis et al. 2007) can be applied to remove paralogs from a set of homologous genes. Furthermore, a gene transfer leads to differences between gene tree and species tree. These differences correspond to the distortion of the cophylogenetic structure between parasite and host tree due to a parasite switching to another host. In that case, the null hypothesis of the `ParaFit` statistical test is that gene tree and species tree do not share a common evolutionary history.

Since this test can also be applied to individual taxa (1-to-1 associations), genes that contribute negatively to the correlation between gene and species tree can be detected. When a gene has a p-value above the significance threshold α, the null hypothesis is considered to be not rejected for this gene. Such genes may be regarded as candidates for horizontal gene transfer.

To determine a reasonable significance threshold, the HGT counts obtained with different threshold values were optimized on the basis of their Pearson and Kendall correlation coefficient with the ML-conflict measure.

A topology-based method

We implemented an additional HGT detection method that is based on comparing the gene tree topology with the NCBI taxonomy.

An important precondition of this method is a meaningful rooting of the gene trees to be tested. Therefore, trees having a monophyletic archaeal clade were rooted by placing the root between the two prokaryotic superkingdoms. In all other cases, midpoint rooting (Farris 1972) was applied. The midpoint rooting strategy can be considered as reliable in cases where a proper outgroup cannot be provided (Hess and De Moraes Russo 2007).

Prior to invoking the HGT detection, edges with a low bootstrap support (below 0.90) were removed from the gene trees.

In rooted trees, each inner node defines an individual subtree. Accordingly, there is a direct relation between an inner node and a clade comprising all leaves that belong to the inner node's subtree.

Each taxonomic rank defines a complete partitioning of the taxa contained in the taxonomy. For each partition, the corresponding clades (i.e., inner nodes) are recursively determined in the gene tree. In the absence of HGT or other aberrations, a one to one correspondence is anticipated between the taxonomic partition and an inner node (representing a clade) of the gene tree. However, HGT leads to a polyphyly, which means that there has to be more than one clade that represents all members of a taxonomic partition. This even leads to the detection of paraphyletic clades when searching for the taxonomic partition that harbors the donor organism of the displaced gene.

In the case of HGT, our proposed algorithm has to detect the "main" clade, i.e. the clade that is supposed to represent the majority of the taxa belonging to this partition. By applying Ockham's razor, the most parsimonious scenario would then be to assume that this is the clade that correctly reflects the evolutionary history, whereas diverging clades may be caused by horizontal gene transfer. But it has to be considered that a clade may be paraphyletic, which, in this context, means that it also can include alien taxa. Thus, when searching the main clade of a taxonomic partition, our method tries to achieve a balance between the overall clade size, and the amount of taxa in the clade that actually belong to the corresponding partition.

To address this problem, we defined a clade score, which can be optimized by recursively traversing the gene tree. Let C be a set of clades that are not nested and that contain at least one taxon of the partition P currently under investigation. Moreover, we demand that C has to comprise all taxa of P, i.e. $\bigcup_{c \in C} c = P$. Let $|C|$ denote the amount of clades in C, and $p_c = c \cap P$ be the number of leaves belonging to $c \in C$ that are members of P. Further, let $|c|$ denote the total number of leaves within clade c. The algorithm then tries to recursively find the maximum clade score as defined by:

$$\text{cladescore}(C) := \frac{\sum_{c \in C} \frac{p_c}{|c|}}{|C|^2} \tag{3.3}$$

The quadratically increasing denominator acts as a penalty that avoids getting a large number of distinct clades.

To optimize the clade detection process, the algorithm is allowed to resolve polytomies, which emerge due to the removal of edges with a low bootstrap support. This was done by combining clades that are directly connected to the same polytomous node.

The resulting set of clades that are not identical to the corresponding main clade were considered to be candidates for HGT. The algorithm is repeated for each partition belonging to the taxonomic groupings of order, class, phylum and superkingdom.

3.2.4 Statistical tests

Test for rejection of the species phylogeny

Poptsova and Gogarten (2007) showed recently that the AU test (Shimodaira 2002, approximately unbiased test) is a good method in order to test whether a dataset rejects the species phylogeny. Thus, this test can be used to assess the reliability of a reconstructed gene tree that is disagreement with the species phylogeny.

For each alignment, we tested the original best ML tree, a ML tree based on the alignment under consideration that was constrained by the NCBI taxonomy, as well as the three species tree candidates. Site-wise log likelihoods were calculated by `RAxML` after applying a branch length optimization of the input trees. Afterwards, `consel` (Shimodaira and Hasegawa 2001) was used to perform the AU test.

A dataset was considered as incongruent with the species phylogeny, if no single candidate of the potential species trees reached a p-value larger than the significance level $\alpha = 0.01$.

Empirical test for false positives in HGT detection

A second AU test was performed to examine whether the different methods missed to detect some HGT events. For each combination of HGT detection method and reference (species) tree, the presence of such false positives was tested as follows: A pruned reference tree was created by removing all leaves for which the current method reported a HGT event. Subsequently, `RAxML` was used to infer a tree using the the pruned reference tree as constraint. Afterwards, an AU test was conducted with the best (unconstrained) ML tree and the newly inferred tree. Thus, in absence of any further HGT events, the tree under constraint should be within the confidence interval computed by `consel`.

ML-based measure of conflict between gene and reference trees

In addition to the c-score (see Section 2.2.5, page 27), which provides a measure of discrepancy between gene trees and a reference topology, the ML-conflict measure as recently introduced by Galtier and Daubin (2008) can also be applied. The ML-conflict quantifies divergence of two trees based on their Maximum Likelihood scores. Here, $\mathrm{LnL}_{\mathrm{alignment}}(T)$ denotes the likelihood of tree T for the given alignment, by optimizing branch lengths

of T. The ML-conflict between the gene tree T_{gene}, and reference tree T_{ref} is calculated as follows:

$$\text{ML}_{\text{conflict}}(T_{\text{gene}}, T_{\text{ref}}) := \min \begin{pmatrix} \text{LnL}_{\text{gene}}(T_{\text{gene}}) - \text{LnL}_{\text{gene}}(T_{\text{ref}}), \\ \text{LnL}_{\text{ref}}(T_{\text{ref}}) - \text{LnL}_{\text{ref}}(T_{\text{gene}}) \end{pmatrix} \quad (3.4)$$

As long as the best ML trees are used for T_{gene} and T_{ref}, the resulting distance should always be a positive number. Thus, when applying this distance function to the reference tree that was constrained by the NCBI taxonomy, a negative distance may occur when the current gene tree may exhibit a higher LnL score for the concatenated alignment than the constrained reference tree does. In this case, we used $\text{LnL}_{\text{gene}}(T_{\text{gene}}) - \text{LnL}_{\text{gene}}(T_{\text{ref}})$ as a measure of ML-conflict.

3.3 Results and Discussion

3.3.1 Detected Orthologous Clusters

Using the dataset of 279 prokaryotic genomes we derived a set of 17 common genes (see Table 3.1). Most of these 17 genes are contained in the set of prokaryotic core genes detected by Charlebois and Doolittle (2004). Additionally, we also found Alanyl and Tyrosyl tRNA-Synthetase. Despite many differences in the translational apparatus between Archaea and Bacteria, Elongation factor G in Bacteria and its archaeal homolog, Elongation factor 2, are also conserved across the two superkingdoms. A considerable set of genes related to replication and translation was removed from the dataset due to lack of conservation across both superkingdoms.

Some alignments were heavily reduced when applying `Gblocks`. The impact using `Gblocks` has on the quality of tree inference is analyzed in the following chapter.

3.3.2 Prokaryotic gene phylogenies

Model selection

Table 3.2 shows the amino acid substitution models that were selected by ProtTest (based on `phyml`) and `ProteinModelSelection.pl`, which utilizes `RAxML`. Interestingly, only two substitution matrices are selected by both methods, the WAG model (Whelan et al. 2001) and the RtREV model (Dimmic et al. 2002). The only discrepancy in model preference was observed in the gene of Phenylalanyl-tRNA synthetase β subunit. In that case, the AIC criterion favoured the RtREV+G+F model slightly over WAG+G+F. Since the amount of parameters is the same for both models (20 + 557 branch length estimates), a rather small difference in log likelihoods was crucial for

Description	No. of Paralogs	No. taxa with par.	Alignment length	Length after `Gblocks`	Percentage remaining
Elongation factor G/2	52	46	1509	712	47
Threonyl-tRNA synthetase	17	16	1031	674	65
Tyrosyl-tRNA synthetase	16	16	713	320	44
DNA polymerase III subunits γ and τ / replication factor C small subunit	7	7	3689	261	7
Methionyl-tRNA synthetase	6	6	1680	484	28
Arginyl-tRNA synthetase	6	6	1056	358	33
GTP-binding protein, YchF family	5	5	557	431	77
50S ribosomal protein L11	3	2	268	156	58
Phenylalanyl-tRNA synthetase β subunit	1	1	1447	593	40
Alanyl-tRNA synthetase	1	1	1628	809	49
O-sialoglycoprotein endopeptidase	0	0	1247	314	25
30S ribosomal protein S3	0	0	524	244	46
Phenylalanyl-tRNA synthetase α subunit	0	0	725	319	44
50S ribosomal protein L1	0	0	376	252	67
Valyl-tRNA synthetase	0	0	2460	754	30
Translation initiation factor IF-2	0	0	1984	588	29
30S ribosomal protein S9	0	0	313	143	45

Table 3.1: Orthologous clusters found by OrthoMCL.
The last column shows the percentage of remaining sequence length after cleaning the MSAs with `Gblocks`. ORFs not contained in the set of core genes described by Charlebois and Doolittle (2004) are set in bold type.

the preference of RtREV+G+F (LnL: −151736.34, AIC: 304626.67) over WAG+G+F (LnL: −151766.37, AIC: 304686.74).

Based on this observation, we conclude that `ProteinModelSelection.pl` performs as well as ProtTest in most cases. Given that `RAxML` is much faster than `phyml`, the former approach may be preferred for middle sized to large-scale datasets.

To evaluate the performance of selecting an individual model for each gene, we also inferred trees using the well-known JTT model (Jones et al. 1992). Table 3.3 shows the log likelihoods and the Robinson-Foulds distance between the trees inferred using the model chosen by `ProteinModelSelection.pl` and the trees based on the JTT substitution model. Whereas there are considerable discrepancies in likelihood scores between the trees based on these different substitution models, differences in c-scores are moderate. There are even some trees having a better c-score when using the JTT model. This may be an indication that the substitution model choice seems to have a smaller influence on tree reconstruction accurracy as it is commonly assumed.

Gene length and average bootstrap support

Figure 3.1 shows the relation between gene length before and after applying `Gblocks`, and the median bootstrap support (MBS) of the corresponding re-

Description	ProtTest	LnL (model)	RAxML	LnL (best tree)
Elongation factor G/2	RtREV+G+F	-136976.48	RtREV+G+F	−136155.13
Threonyl-tRNA synthetase	WAG+G	-145608.29	WAG+G	−144907.99
Tyrosyl-tRNA synthetase	RtREV+G+F	-75530.66	RtREV+G+F	−74870.82
DNA polymerase III subunits γ and τ / replication factor C small subunit	RtREV+G+F	-59224.63	RtREV+G+F	−58745.12
Methionyl-tRNA synthetase	WAG+G+F	-97075.39	WAG+G+F	−96627.21
Arginyl-tRNA synthetase	RtREV+G+F	-87123.25	RtREV+G+F	−86740.49
GTP-binding protein, YchF family	RtREV+G+F	-84454.48	RtREV+G+F	−83958.78
50S ribosomal protein L11	RtREV+G+F	-25330.85	RtREV+G+F	−25025.49
Phenylalanyl-tRNA synthetase β subunit	**RtREV+G+F**	-151736.34	**WAG+G+F**	−151420.00
Alanyl-tRNA synthetase	WAG+G	-184472.91	WAG+G	−183889.11
O-sialoglycoprotein endopeptidase	WAG+G	-66102.17	WAG+G	−65786.31
30S ribosomal protein S3	RtREV+G+F	-36989.87	RtREV+G+F	−36692.64
Phenylalanyl-tRNA synthetase α subunit	WAG+G	-71681.65	WAG+G	−71342.27
50S ribosomal protein L1	RtREV+G+F	-49261.86	RtREV+G+F	−48851.05
Valyl-tRNA synthetase	WAG+G	-158019.16	WAG+G	−157475.06
Translation initiation factor IF-2	RtREV+G+F	-113089.23	RtREV+G+F	−112404.28
30S ribosomal protein S9	WAG+G	-24024.51	WAG+G	−23757.89

Table 3.2: Substitution models and log Likelihoods. Note however that log likelihoods of phyml and RAxML should be compared with caution. The Akaike weights calculated by ProtTest were always 1 for the selected models.

Description	model LnL	JTT LnL	model c-score	JTT c-score	normalized RF dist.
Elongation factor G/2	−136155.13	−137646.05	0.6232	0.6196	0.0942
Threonyl-tRNA synthetase	−144907.99	−146525.10	0.6232	0.6341	0.0290
Tyrosyl-tRNA synthetase	−74870.82	−75822.61	0.5725	0.5652	0.1159
DNA polymerase III subunits γ and τ / replication factor C small subunit	−58745.12	−59229.21	0.6558	0.6775	0.2283
Methionyl-tRNA synthetase	−96627.21	−97635.49	0.7065	0.7029	0.0652
Arginyl-tRNA synthetase	−86740.49	−87854.07	0.5652	0.5688	0.0761
GTP-binding protein, YchF family	−83958.78	−84722.13	0.6630	0.6304	0.1848
50S ribosomal protein L11	−25025.49	−25494.59	0.5833	0.6051	0.2319
Phenylalanyl-tRNA synthetase β subunit	−151420.00	−152890.52	0.7174	0.7138	0.1196
Alanyl-tRNA synthetase	−183889.11	−185687.88	0.7101	0.7029	0.0725
O-sialoglycoprotein endopeptidase	−65786.31	−66502.38	0.6377	0.6196	0.1159
30S ribosomal protein S3	−36692.64	−37022.16	0.6920	0.7174	0.2355
Phenylalanyl-tRNA synthetase α subunit	−71342.27	−72043.48	0.6667	0.6848	0.1703
50S ribosomal protein L1	−48851.05	−49277.75	0.7029	0.6739	0.1413
Valyl-tRNA synthetase	−157475.06	−159249.08	0.6775	0.6630	0.1341
Translation initiation factor IF-2	−112404.28	−113454.37	0.7717	0.7319	0.1377
30S ribosomal protein S9	−23757.89	−23950.54	0.6232	0.6486	0.2790

Table 3.3: Comparison of results obtained by preferring a specifically selected model over the JTT model. c-scores were obtained from the purified trees (after applying AxParafit to remove paralogs).

constructed tree. The original trees prior to applying `AxParafit` were used for this comparison. Above a gene length of approximately 350 residues, the median bootstrap support persistently is over 70%. In 15 out of 17 cases the MBS has become worse after applying `Gblocks`. Exceptions were 50S ribosomal protein L11 (52% against 53% after applying `Gblocks`) and Elongation factor G/2 where no difference was detected (88% in both cases). This seems to indicate that the truncated columns bear sufficient phylogenetic signal that is in agreement with the columns selected by `Gblocks`. In case of disagreement, the MBS values of the trees based on untrimmed alignments would be noticeably reduced, which would be in contradiction with our observation.

However, systematic error in the data, equally affecting columns selected or ruled out by `Gblocks`, could be a further explanation of this observation. Figure 3.2 shows a plot of gene length against the c-score using the NCBI taxonomy as reference. In most cases, a slight improvement of the c-score could be observed when using untrimmed sequences. Since the space of possible tree topologies is quite large for 279 taxa (approximately 10^{642} rooted trees!), an improvement of the c-score by chance only seems to be rather unlikely. Thus, using the untrimmed alignments here seems to lead to an actual improvement of the trees, at least in most instances.

We do not want to challenge the meaning of using `Gblocks` for alignment purification in general, but our results seem to coincide with Talavera and Castresana (2007) who concluded that ML is able to extract some signal even from problematic alignment regions.

Phylogenies

Based on the comparison of phylogenetic trees obtained with and without alignment trimming, we decided to use the trees inferred from the untrimmed alignments for HGT detection. Table 3.4 shows the ML-conflict measure and the c-scores of the obtained phylogenies. The Pearson correlation coefficient between c-score measure and ML-conflict is -0.5300 using the MLNC tree, and -0.5379 using the best ML tree (BML). Figure 3.3 depicts a plot of c-scores against the BML ML-conflict. The plot clearly shows a correlation between both measures, especially when both measures indicate a high agreement with the respective reference tree. This indicates a strong phylogenetic signal in the data, and also, a significant congruence between both, the individual gene trees and the supermatrix tree on the one hand, and the NCBI taxonomy on the other hand.

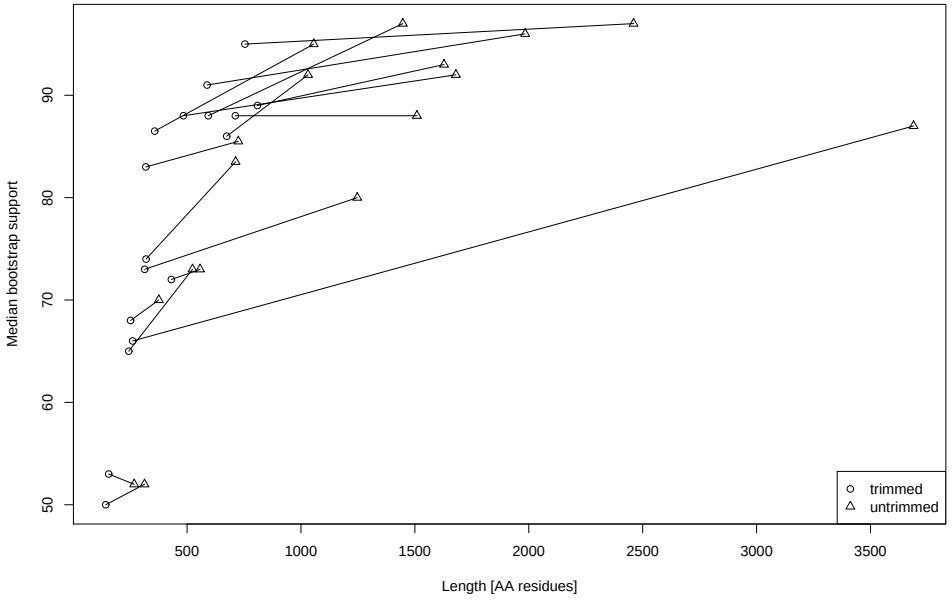


Figure 3.1: Alignment length vs. Median bootstrap support for trees based on trimmed (`Gblocks`) and untrimmed alignments. A connecting line is drawn between each pair of trimmed/untrimmed values.

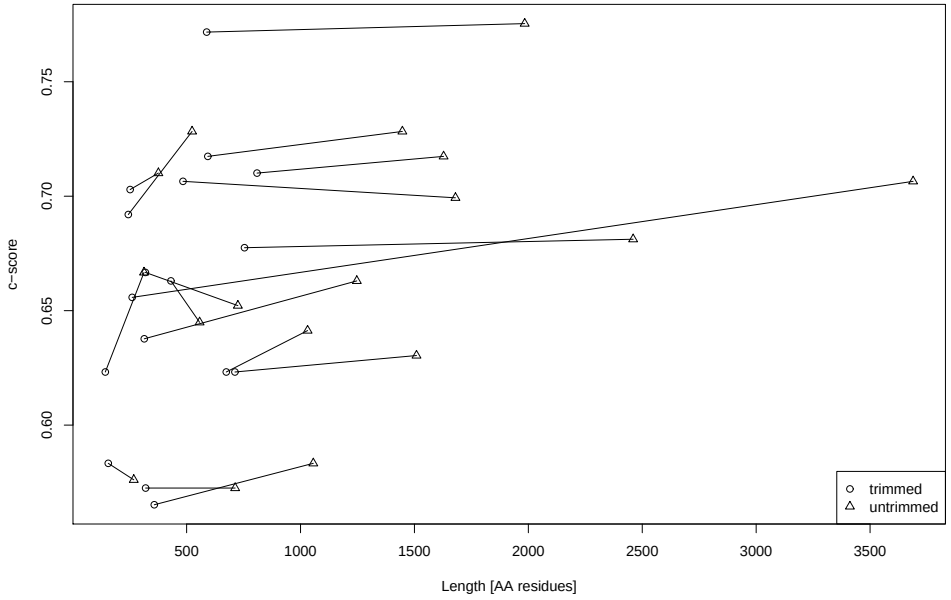


Figure 3.2: Alignment length vs. c-score against the reference taxonomy for trees based on trimmed (`Gblocks`) and untrimmed alignments. A connecting line is drawn between each pair of trimmed/untrimmed values.

Description	MLNC ML-conflict	BML ML-conflict	c-score
Elongation factor G/2	4900.49	4602.28	0.6304
Threonyl-tRNA synthetase	5190.79	4214.48	0.6413
Tyrosyl-tRNA synthetase	9733.19	6532.41	0.5725
DNA polymerase III subunits γ and τ / replication factor C small subunit	1478.44	972.85	0.7065
Methionyl-tRNA synthetase	6144.41	5462.72	0.6993
Arginyl-tRNA synthetase	14043.01	10645.44	0.5833
GTP-binding protein, YchF family	1485.22	1136.86	0.6449
50S ribosomal protein L11	890.93	732.59	0.5761
Phenylalanyl-tRNA synthetase β subunit	2224.14	992.74	0.7283
Alanyl-tRNA synthetase	2620.86	1409.40	0.7174
O-sialoglycoprotein endopeptidase	1114.01	903.36	0.6630
30S ribosomal protein S3	803.53	529.33	0.7283
Phenylalanyl-tRNA synthetase α subunit	1485.13	923.58	0.6522
50S ribosomal protein L1	985.02	691.95	0.7101
Valyl-tRNA synthetase	4549.25	3028.55	0.6812
Translation initiation factor IF-2	1984.28	1321.65	0.7754
30S ribosomal protein S9	810.06	572.70	0.6667

Table 3.4: Agreement of the gene trees measured against reference trees using the Maximum Likelihood distance, as well as the c-score.
NCBI: results obtained by directly using the NCBI taxonomy.
MLNC: results based on the reference tree inferred by using the NCBI taxonomy as constraint.
BML: best ML tree as determined by `RAxML`.

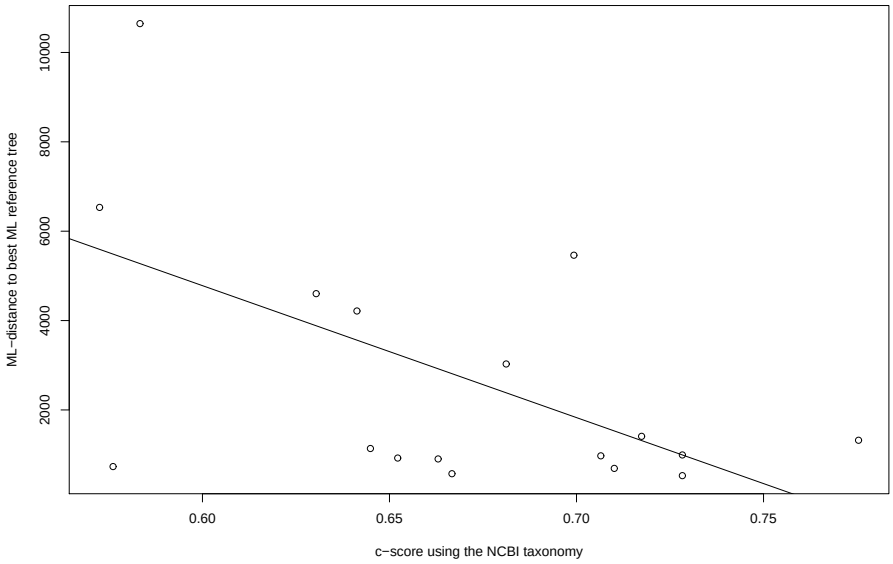


Figure 3.3: Plot of c-scores based on the NCBI taxonomy against ML-conflict between the gene trees and the best ML reference tree. The Pearson correlation coefficient is -0.5379.

3.3.3 Parameter optimization for HGT detection

Selection of an appropriate α threshold for Parafit

To detect horizontal gene transfer, topological and distance-based methods try to determine deviations between a gene tree and a reference tree, based on their topology or patristic distances. It can be assumed that there should be an interrelation between measures of discrepancy among gene and reference trees on the one hand, and the amount of HGT that can be detected by these methods, on the other hand. That plausible connection may depict a suitable and empirical strategy for optimizing sensitivity influencing parameters of the concerned methods.

In particular, `ParaFit` allows to select a p-value threshold, which indicates when an association is considered as significant, i.e., when the null hypothesis stating that there is no cophylogenetic relationship can be rejected. To determine an optimal p-value threshold, we computed the HGT counts for each p-value between 0.05 and 1.0, and we calculated the corresponding Pearson and Kendall correlation coefficients between the amount of HGT and the corresponding ML-conflict value for each gene tree. Additionally, we included the correlation to the c-score between gene tree and reference tree to provide a second discrepancy measure that is solely based on the tree topology.

Figures 3.4 (Pearson) and 3.5 (Kendall) show a plot of p-value thresholds against correlation coefficients. Overall, the correlation remains high for values between 0.05 and 0.80, having its maximum at a p-value threshold of $\alpha = 0.025$, when regarding the Pearson correlation. Using the Kendall correlation coefficient, the best value could be obtained at $\alpha = 0.020$, but there is only a small observable difference to the maximum when using $\alpha = 0.025$. Correlation to the c-score measure is considerably lower, but it shows a trend similar to the ML-conflict's correlation.

Based on the initial assumption of correlation between the amount of detected HGT and discrepancy measures, we recommend a threshold of $\alpha = 0.025$ for HGT detection based on `ParaFit`.

Threshold for Cook's distance

Kanhere and Vingron (2009) proposed a threshold of $\frac{2}{D}$ for the Cook's distance (see Section 3.2.3, p. 65) based on the results of a simulation study. To determine whether the observed discrepancy between the correlation of the HGT counts using the CDISS and the correlation of the `ParaFit` HGT counts (see Table 3.6) represent a general tendency, we applied the proposed optimization method to the Cook's distance by investigating correlations for HGT counts obtained by applying cut-off values between $\frac{0.5}{D}$ and $\frac{3}{D}$.

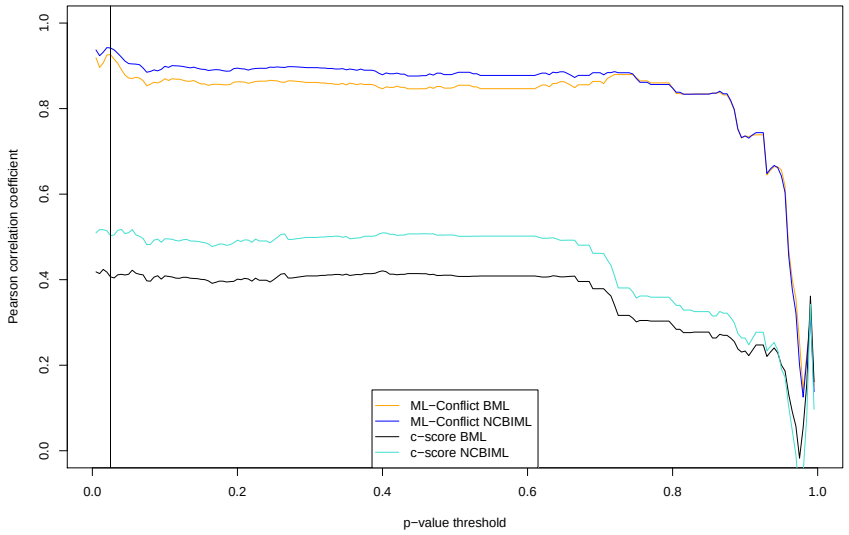


Figure 3.4: Pearson correlation between `ParaFit` HGT counts and ML-Conflict, as well as c-score to both ML-based reference trees, for different p-value thresholds. The signum of the c-score correlations was changed to fit in the plot.

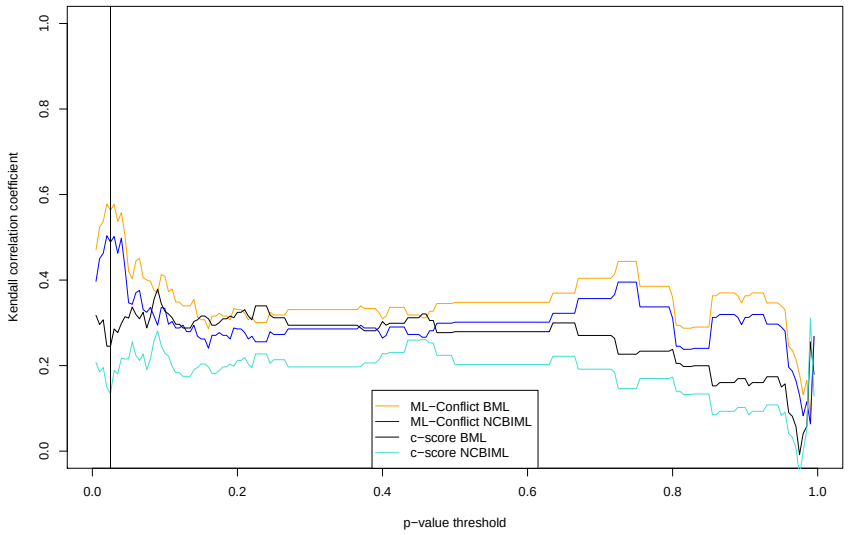


Figure 3.5: Kendall correlation between `ParaFit` HGT counts and ML-Conflict, as well as c-score to both ML-based reference trees, for different p-value thresholds. The signum of the c-score correlations was changed to fit in the plot.

Results are shown in Figure 3.6 (Pearson) and Figure 3.7 (Kendall). Clearly, the maxima are far from the threshold proposed by Kanhere and Vingron (2009), and maxima obtained by using both correlation coefficients do not show coincidence. This may indicate that there is no strong linear relationship between HGT counts obtained by using the CDISS method and the ML-conflict measure. Overall, both correlation coefficients indicate an inferior interrelation, compared to results obtained by using `ParaFit` (Figures 3.4 and 3.5).

Using thresholds that are optimal considering the correlation to the ML-conflict would lead to a pronounced increase of HGT counts. For example, when using a threshold of $\frac{0.9}{D}$ as proved optimal when using the Pearson correlation, the number of detected occurrences is more than fivefold in comparison to using the default cut-off. Furthermore, by lowering the CDISS threshold, the rate of false positives increases to a level above 5% (Kanhere and Vingron 2009).

Bootstrap cut-off for the topological method

We also applied the optimization strategy to the topological method for comparison purposes. Figures 3.8 (Pearson), and 3.9 (Kendall) show the results. The bootstrap cut-off determines the edges to be removed from the gene tree under consideration prior to applying the topological HGT detection method. Overall, above a bootstrap cut-off of 0.50, the correlation tends to improve when increasing the cut-off value. There are two interesting observations: a local maximum at 0.80, as well as a global maximum at a cut-off of 1.0, meaning that all edges having a lower support will be removed from the tree. Between these values, there is a small decrease towards the cut-off of 0.90 that was preferred in our study.

There is neither agreement in current literature concerning a feasible bootstrap cut-off value, nor how bootstrap values should be interpreted. The latter was already touched on in Section 2.2.5 (p. 25). Considering cut-off selection, statistical properties of the bootstrap procedure play an important role. Some authors consider the bootstrap method as biased towards the underestimation of support values (see discussion in Wróbel 2008; Felsenstein 2004, pp. 346). They argue that even a support of 70% might be regarded as sufficient (Hillis and Bull 1993). In contrast, Taylor and Piel (2004) found no evidence for a systematic bias. Hence, they proposed a conservative cut-off as high as 95% for the yeast dataset they surveyed.

Bringing all this together, we recommend a cut-off of at least 90% when searching for HGT occurrences in a phylogenetic tree. Avoiding the detection of spurious events should be the primary goal, even though this increases the amount of false negatives, i.e., undetected events. A low support value indicates that the branch is not sufficiently backed up by the underlying data, thus this branch should not be considered for HGT detection by a

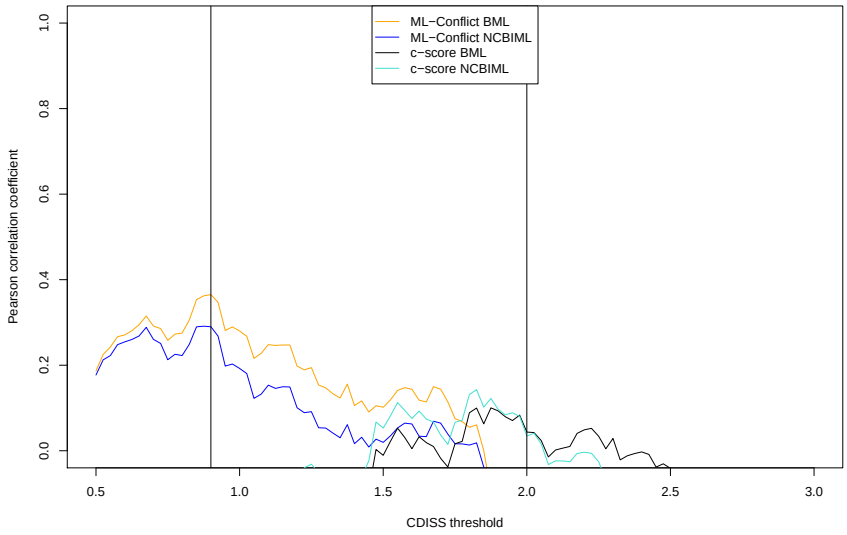


Figure 3.6: Pearson correlation between CDISS HGT counts and ML-Conflict, as well as c-score to both ML-based reference trees, for different thresholds n/D. The signum of the c-score correlations was changed to fit in the plot.

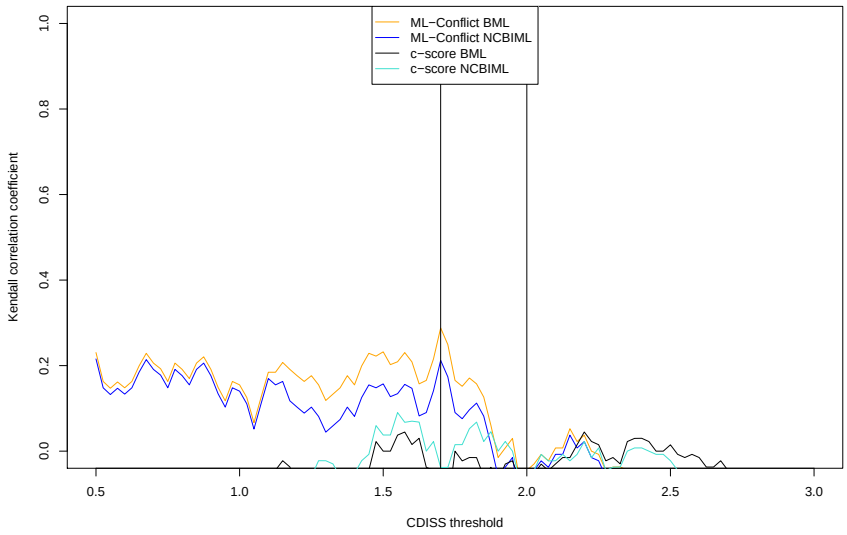


Figure 3.7: Kendall correlation between CDISS HGT counts and ML-Conflict, as well as c-score to both ML-based reference trees, for different thresholds n/D. The signum of the c-score correlations was changed to fit in the plot.

topology-based method. Removing a branch allows the algorithm to optimize its clade detection step by combining nodes belonging to the same clade while resolving polytomies. Consequently, inclusion of branches having low bootstrap values leads to recognition of more occurrences, especially those that are poorly supported by the underlying data. Here, choosing a cut off of 90% seems to represent a balanced trade-off between sensitivity and accurracy, whereas there is only a small loss of correlation against the ML-conflict measure (see Figures 3.8, and 3.9).

3.3.4 Comparison of the results

Amount of detected HGT events

Table 3.5 shows the amount of hypothetical HGT occurrences as detected by the different methods. In the following, we do not use the term "HGT event" in a stringent biological sense, meaning that an actual gene transfer happened between a distinct donor and recipient, in contrast to vertical gene transfer by clonal reproduction. Here, vertical inheritance of a gene followed by speciation leads to the recognition of one "event" for each descendant that is represented in the dataset, even if, in a stringent sense, only one actual transfer happenend during the lifetime of a common ancestor of the affected group. This results in an overestimation of the objective amount of HGT events, since we cannot distinguish between a recent event of HGT affecting a single species on the one hand, and a past event preceding cladogenesis on the other hand. In the following, the HGT numbers may be seen as numbers indicating how many taxa may be affected by HGT, instead of a count of individual gene transfers.

The amount of detected occurrences range between 291 (NCBI-`AxParafit`) to more than 500 (BML-`AxParafit` and topological method). Dagan and Martin (2007) tried to assess the amount of HGT by estimating the size of ancestral genomes that would be necessary to explain the observed amount of today's gene families. Based on this, they concluded that at least 65%, and perhaps even all gene families may be affected by at least one HGT event during their evolutionary history. In a later study, they even stated that at least $81 \pm 15\%$ of the genes within the probed genomes were involved in HGT (Dagan et al. 2008). Considering the subsequent vertical inheritance of these aquired genes, an accumulation of such horizontally acquired genes seems to be inevitable. Further, genes that are present in all recent organisms, must also have been present in the last universal common ancestor, and thus, have had plenty of time to undergo genetic transfers. In the light of this thought, the amount of HGT occurrences detected by the different methods may be credible.

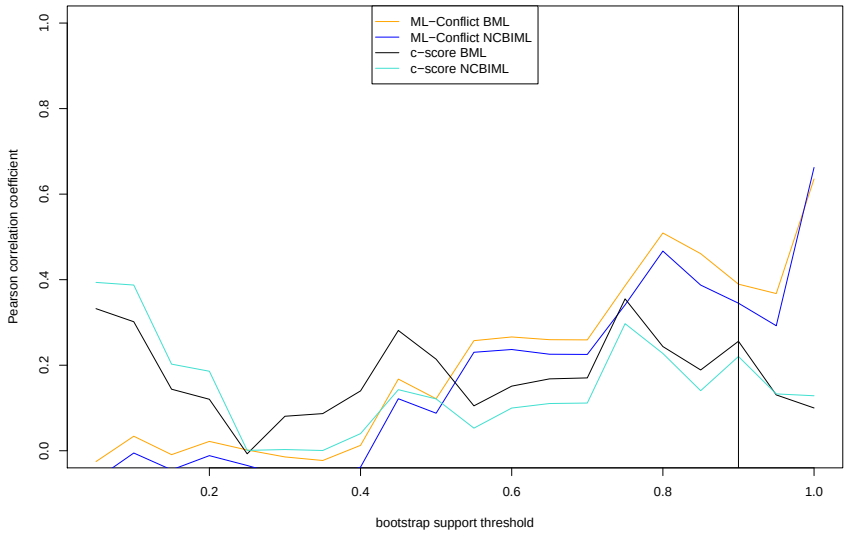


Figure 3.8: Pearson correlation between the topological method's HGT counts and ML-Conflict, as well as c-score to both ML-based reference trees, for different bootstrap significance thresholds. The signum of the c-score correlations was changed to fit in the plot.

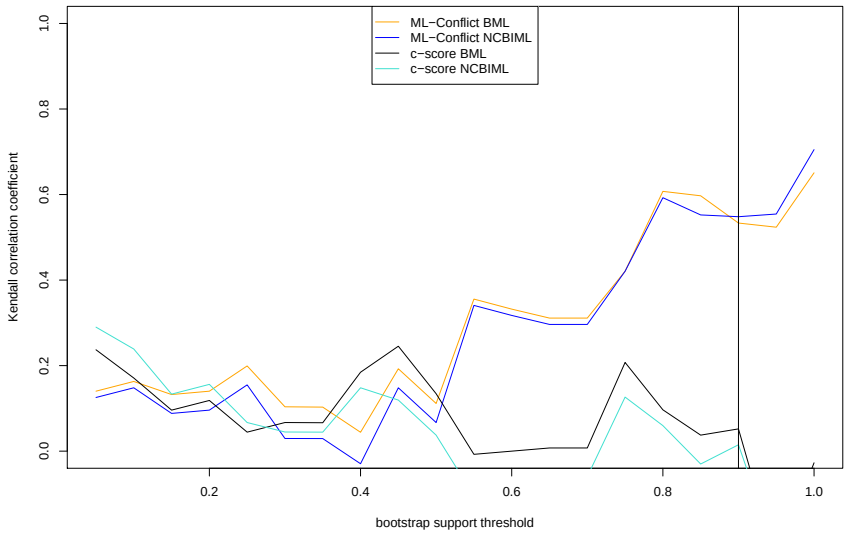


Figure 3.9: Kendall correlation between the topological method's HGT counts and ML-Conflict, as well as c-score to both ML-based reference trees, for different bootstrap significance thresholds. The signum of the c-score correlations was changed to fit in the plot.

Effect of reference tree selection

Three possible candidates for a species tree were used for HGT detection based on the CDISS and `AxParafit` method. One is solely based on the NCBI topology, and thus, its branch lengths reflect taxonomical distances rather than the number of site changes. The other alternatives are based on a concatenation of all single gene alignments (supermatrix). One tree was inferred using `RAxML` to search for the best ML tree (BML tree), whereas the other was reconstructed by `RAxML` using the NCBI topology as constraint (MLNC tree). This leads to a binary tree that is completely in agreement with the NCBI taxonomy, but having branch lengths reflecting rates of site changes.

Table 3.6 shows the correlations between the amount of predicted occurrences and the ML-conflict as well as the c-scores. Interestingly, a high correlation between both ML-conflict values (BML and MLNC) can be observed (Pearson $r_{ps} = 0.9871$). This indicates that gene trees deviating from the reference tree based on the concatenated alignment, also deviate in a similar magnitude from the constrained reference tree that resembles the NCBI taxonomy.

Using the CDISS method, correlation of the amount of HGT occurrences is high ($r_{ps} = 0.9217$) for the MLNC (constrained) and the best ML tree (BML, unconstrained), but considerably lower between MLNC and the NCBI taxonomy ($r_{ps} = 0.4023$), as well as between BML and the NCBI taxonomy ($r_{ps} = 0.3970$). In contrast, the correlations are far higher when comparing results based on `AxParafit` showing $r_{ps} > 0.9$ in each case. A main difference between the NCBI taxonomy tree and the two trees derived by using `RAxML` consists in the meaning of the distances between leaves. Using the NCBI taxonomy, this distance resembles taxonomical units, while in the other cases, the distances are estimated by applying the Maximum Likelihood function. While the `ParaFit`-based method was already successfully applied to patristic distances solely based on taxonomical information (Meier-Kolthoff et al. 2007; Stamatakis et al. 2007), the CDISS distance was tested on distances that were derived in a manner that allows to assume a linear correlation between inter-gene and inter-genomic distances. This condition may possibly be violated when comparing topological distances with ML-based distances.

Nevertheless, for both distance-based methods, more than 75% of HGT occurrences are consistently detected regardless of which ML tree was used as reference (see Table 3.7, MLNC $\cap$ BML). The agreement between results based on the best ML tree, as well as the constrained tree indicates that the supermatrix method can be used here to infer a reliable species tree. Furthermore, the c-score of the BML tree is 0.7536 when using the NCBI taxonomy as reference. This means that the phylogenetic congruence between both trees is in the same range as the percentage of overlap between

the detected HGT occurences. Moreover, the agreement between both trees indicates that there is a strong signal of vertical inheritance in the derived set of genes, despite being exposed to a considerable amount of HGT.

Description	NCBI CDISS	MLNC CDISS	BML CDISS	NCBI AxParafit	MLNC AxParafit	BML AxParafit	Topo
Elongation factor G/2	25	33	40	16	23	48	117
Threonyl-tRNA synthetase	25	31	31	9	21	27	121
Tyrosyl-tRNA synthetase	10	18	18	62	91	93	71
DNA polymerase III subunits γ and τ / replication factor C small subunit	18	12	20	17	22	27	2
Methionyl-tRNA synthetase	8	3	4	29	30	49	23
Arginyl-tRNA synthetase	23	13	11	65	135	139	28
GTP-binding protein, YchF family	26	33	30	19	29	30	109
50S ribosomal protein L11	17	26	23	10	25	29	1
Phenylalanyl-tRNA synthetase β subunit	23	24	23	8	18	21	12
Alanyl-tRNA synthetase	23	23	23	1	12	18	1
O-sialoglycoprotein endopeptidase	17	35	31	9	14	16	5
30S ribosomal protein S3	31	19	21	11	27	31	3
Phenylalanyl-tRNA synthetase α subunit	26	20	19	14	19	20	17
50S ribosomal protein L1	20	18	20	4	10	15	0
Valyl-tRNA synthetase	24	22	22	2	38	32	53
Translation initiation factor IF-2	27	18	17	5	24	28	7
30S ribosomal protein S9	23	21	18	10	15	16	0
Sum	366	369	371	291	553	639	570

Table 3.5: Amount of detected HGT events using different detection methods. Note that each gene considered as conflicting by the accordant method is counted as one individual event, thus leading to an overestimation if an entire monophyletic group is affected by the same HGT event due to vertical inheritance.
NCBI: results obtained by directly using the NCBI taxonomy.
MLNC: results based on the reference tree inferred by using the NCBI taxonomy as constraint.
BML: best ML tree as determined by `RAxML`.

Correlations	NCBI CDISS	MLNC CDISS	BML CDISS	NCBI AxParafit	MLNC AxParafit	BML AxParafit	Topo	MLNC ML-conflict	BML ML-conflict	c-score
NCBI CDISS		0.1776	0.1401	-0.1303	0.0380	0.0230	0.1456	-0.0837	-0.0837	0.2291
MLNC CDISS	0.3970		0.7786	-0.2105	-0.2090	-0.1504	0.1504	-0.1045	-0.0896	-0.1423
BML CDISS	0.4023	0.9217		-0.2349	-0.2256	-0.1515	0.1591	-0.0902	-0.0451	-0.1132
NCBI AxParafit	-0.4272	-0.3735	-0.3960		0.5185	0.5075	0.2687	0.2222	0.2667	-0.4089
MLNC AxParafit	-0.1998	-0.3235	-0.3819	0.9136		0.8445	0.4000	0.3529	0.4265	-0.3026
BML AxParafit	-0.2483	-0.3428	-0.3540	0.9324	0.9759		0.4478	0.4889	0.5630	-0.2900
Topo	0.1153	0.4731	0.5443	0.2099	0.1661	0.2265		0.5482	0.5333	-0.3346
MLNC ML-conflict	-0.2989	-0.3180	-0.3224	0.8571	0.9077	0.9417	0.3450		0.9265	-0.2878
BML ML-conflict	-0.3033	-0.3064	-0.2931	0.8350	0.8665	0.9264	0.3895	0.9871		-0.3026
c-score	0.3165	-0.2264	-0.1430	-0.6223	-0.5742	-0.5755	-0.4242	-0.5300	-0.5379	

Table 3.6: Correlations between the amount of HGT events detected by the tested methods (CDISS, `ParaFit`, and the topological method, see Table 3.5), and different tree metrics (c-score, and ML-conflict, see Table 3.4).
The lower triangle contains the Pearson correlation coefficients (r_{ps}), whereas the upper triangle contains the Kendall correlation coefficients (r_{kn}).
NCBI: results obtained by directly using the NCBI taxonomy.
MLNC: results based on the reference tree inferred by using the NCBI taxonomy as constraint.
BML: best ML tree as determined by `RAxML`.

Degree of congruence between different HGT detection methods

Table 3.7 shows a quantification of the overlap between HGT detection methods. Clearly, the intersection between predictions of `AxParafit` and the topological method shows the largest overlap, both in absolute counts (152 for BML) and relative amount (24%). However, many hypothetical events detected by `AxParafit` or the topological method could not be confirmed by the Cook's distance based approach.

Whereas the intersection of all methods (CDISS ∩ `AxParafit` ∩ Topo) contains only a small fraction of hypothetical events, we could clearly demonstrate that the events predicted by all methods are established occurrences of HGT (see Section 3.3.5). We also tried the CDISS method using a smaller threshold of $1.3/D$ (Kanhere and Vingron 2009, relaxed cut-off setting). Although this lead to a great increase in the amount of postulated events from 371 to 946 (best ML tree as reference), representing more than a 2.5 fold change. Even when using this large set, the relative amount of intersection sizes stayed almost the same (ranging from 10 to 12 percent).

By adjusting the thresholds for the `AxParafit` method, a larger intersection size may be obtainable, though this would also lead to an increase of false positives. Since all three methods use different approaches for HGT detection, we conclude that a combination of two or more methods may lead to a corroborated set of hypothetical HGT candidates.

	CDISS	CDISS %	`AxParafit`	`AxParafit` %
NCBI ∩ MLNC ∩ BML	159	43	226	35
NCBI ∩ MLNC	181	49	234	42
NCBI ∩ BML	176	47	238	37
MLNC ∩ BML	319	86	493	77

	NCBI	NCBI %	MLNC	MLNC %	BML	BML %
CDISS ∩ `AxParafit` ∩ Topo	4	1	22	4	18	3
CDISS ∩ `AxParafit`	16	4	52	9	53	8
CDISS ∩ Topo	55	10	65	11	71	12
`AxParafit` ∩ Topo	91	16	135	24	152	24

Table 3.7: Intersection sizes
NCBI: results obtained by directly using the NCBI taxonomy.
MLNC: results based on the reference tree inferred by using the NCBI taxonomy as constraint.
BML: best ML tree as determined by `RAxML`.

Empirical test for false positives

All 119 AU-tests indicated the presence of false positives. This may imply that all methods used for HGT detection underestimate the real amount of HGT. A more pessimistic interpretation would be that the confidence interval provided by the AU test narrows with increasing taxon count. Thus, small discrepancies between trees would lead to an exclusion of all trees except the best ML tree from the confidence interval. But this remains speculative, since we did not explicitly test this assumption.

3.3.5 HGT events detected by all methods

In the following, we will focus on the hypothetical HGT events that are predicted by all three proposed methods. This is the most conservative setting that allows to assess whether these methods are appropriate to detect real occurrences of horizontal gene transfer. Table 3.8 shows these 19 events. Thereof, 16 events could be found when using the best ML tree (BML) as reference, as well as when using the constrained reference tree (MLNC). In the remaining cases, either the `ParaFit` p-value lay above the threshold, or the Cook's distance (CDISS) was below the defined threshold.

Valyl tRNA synthetase

Strikingly, 19 occurrences of HGT have been detected for the gene of Valyl tRNA synthetase (see Figure 3.10). All of these have in common that the origin of the transfer seems to be located within the Euryarchaeota, whereas the recipients were Bacteria, more precisely Rickettsiales (α Proteobacteria) and Actinobacteria.

Previous studies already indicated the presence of a gene transfer of this gene from Euryarchaeota to Rickettsia (Farahi et al. 2004; Emelyanov 2003b; Woese et al. 2000), hence, raising the question whether other representatives of Rickettsiales were also affected. Among this order, all representatives of families Rickettsiaceae (*Wolbachia*, and *Rickettsia*) and Anaplasmataceae (*Anaplasma*, *Ehrlichia*, and *Neorickettsia*) that were included in this study, seem to share an archaeal copy of the Valyl tRNA synthetase gene. Only Candidatus *Pelagibacter ubique* seems to be unaffected and is correctly placed within the remaining α Proteobacteria. A possible scenario that could explain this observation is based on the assumption that the SAR11 clade to which *P. ubique* belongs to (Rappé et al. 2002), diverged from the remaining Rickettsiales before Rickettsiaceae and Anaplasmataceae separated. A common ancestor of these two families then may have received the Valyl tRNA synthetase gene from an Euryarchaeon. This would also correspond to the subtree of Rickettsiales located within the Euryarchaeota that is mostly in agreement with the species tree and the NCBI taxonomy.

Also, the supposed branching order of the SAR11 clade and the clade consisting of Rickettsiaceae and Anaplasmataceae is in accordance with the species phylogeny based on the concatenated alignment.

Emelyanov (2003b) investigated the evolutionary history of Valyl tRNA synthetase and detected analogous deviations in the phylogeny of Rickettsiaceae (comprising *R. prowazekii*, *R. conorii*, *Wolbachia*, *Ehrlichia chaffeensis* and *Cowdria ruminantium*). They concluded, "that acquisition of the archaeal enzyme by the family Rickettsiaceae or the order Rickettsiales shaped the evolutionary history of the rickettsial lineage". Our findings strongly support this view.

The remaining hypothetical occurrences of HGT concerned six Actinobacteria, three of them being either human pathogens (*Tropheryma whipplei* TW08/27 and *Propionibacterium acnes* KPA171202), or normal inhabitants (*Bifidobacterium longum*) of the human body. *Frankia alni* and *Leifsonia xyli xyli* are plant pathogens, whereas *Thermobifida fusca* is a hay and organic waste decomposing bacterium. For all observations within Actinobacteria, no indication could be found in the current literature. However, Raoult et al. (2003) detected that Valyl tRNA synthetase gene of *T. whipplei* strain Twist may be derived from Euryarchaeota. Since strain Twist is up to 99% identical to strain TW08/27 at the nucleotide level (Raoult et al. 2003), we assume that both strains are equally affected. Furthermore, the original analysis of *Tropheryma whipplei* TW08/27 focused on the detection of recent genetic transfer by investigating nucleotide composition (Bentley et al. 2003), whereas Raoult et al. (2003) relied on phylogenetic methods.

There is no evidence in literature for HGT between Archaea on the one hand, and the remaining Actinobacteria on the other hand. But certainly, HGT is not uncommon between Actinobacteria, as well as between Actinobacteria and other phyla (Ventura et al. 2007).

Figure 3.10 shows the subtree containing Archaea and the concerned Actinobacteria and Rickettsiales. Strikingly, all Bacteria are concentrated in one cluster showing 100% bootstrap support. Thus, the observed clustering of Actinobacteria within Archaea cannot be explained by one single HGT event. One may assume that one single HGT event between Archaea and these Bacteria occurred, whereas at least one further gene transfer between both bacterial clades must have emerged after the initial transmission. Even then, a simple gene loss scenario after a common ancestor received the archaeal gene, may not be the most parsimonious explanation, because only a small fraction of all Actinomycetales included in this study seem to be affected. No single *Streptomyces* species, nor Corynebacterineae (comprising *Mycobacterium*, *Corynebacterium*, *Rhodococcus*, and *Nocardia*) share the archaeal Valyl tRNA synthetase gene. Nevertheless, the archaeal version is also found in *Bifidobacterium longum*, which belongs to a sister group of Actinomycetales, namely the Bifidobacteriales.

At least, human pathogens (*T. whipplei*, and *P. acnes*) and commensals (*B. longum*) within the phylum Actinobacteria share a common habitat, the human body. These microbes may occasionally meet in the same host, and perhaps, may have had the chance to share genes in an ancestor of our species, a long time ago. This may also explain the clustering with Rickettsiales, since this group also includes many human and animal parasites.

The concerned Bacterial clades both group with Thermoplasmatales (*Thermoplasma acidophilum*, *T. volcanium*, and *Picrophilus torridus*), and two Halobacteria (*Haloquadratum walsbyi*, and *Haloarcula marismortui*), showing 100% bootstrap support for this cluster. These Euryarchaeota occupy rather extreme habitats, like hydrothermal vents (*T. volcanium*), salterns or even the Dead Sea (*H. marismortui*). Despite being well supported by the clustering, these organisms may not be the most likely donors, considering that most of the concerned Bacteria are animal parasites or commensals. Among Euryarchaeota, other possible donor candidates may be found within Methanobacteria, since this phylum includes many inhabitants of the human gastrointestinal tract (Gregory 2005, p. 613), like, e.g., *Methanobrevibacter smithii* (Huson et al. 2009).

In Appendix C.2 (p. 147), two Archaeal consensus networks of a previous study are illustrated (Figures C.1, and C.2). Both networks show a high amount of uncertainty in reconstruction of the branching order of archaeal clades. Interestingly, that uncertainty is even present in the most conserved and least conflicting genes (Figure C.2). Difficulties in the reconstruction of archaeal branching orders have been also experienced by other workgroups (Soria-Carrasco and Castresana 2008), in particular regarding the placement of Thermoplasmatales and Halobacteria (Gophna et al. 2005). Thus, alternatives to the observed clustering may also be possible.

Threonyl tRNA synthetase

For the gene of Threonyl tRNA synthetase, two hypothetical HGT events were detected. An inter-domain transfer from Bacteria to the Archaeon *Aeropyrum pernix* could be observed, which is in agreement with the findings of Farahi et al. (2004).

Furthermore, the Cyanobacterium *Prochlorococcus marinus* strain CCMP-1375 seems to have aquired the Threonyl tRNA synthetase gene from Proteobacteria. This is a well-known example of HGT that seems to affect many *P. marinus* strains (Luque et al. 2008; Zhaxybayeva et al. 2006).

GTP-binding protein, YchF family

The YchF family represents a group of conserved, hypothetical proteins that contain a GTP-binding domain. Recent studies indicate that this is

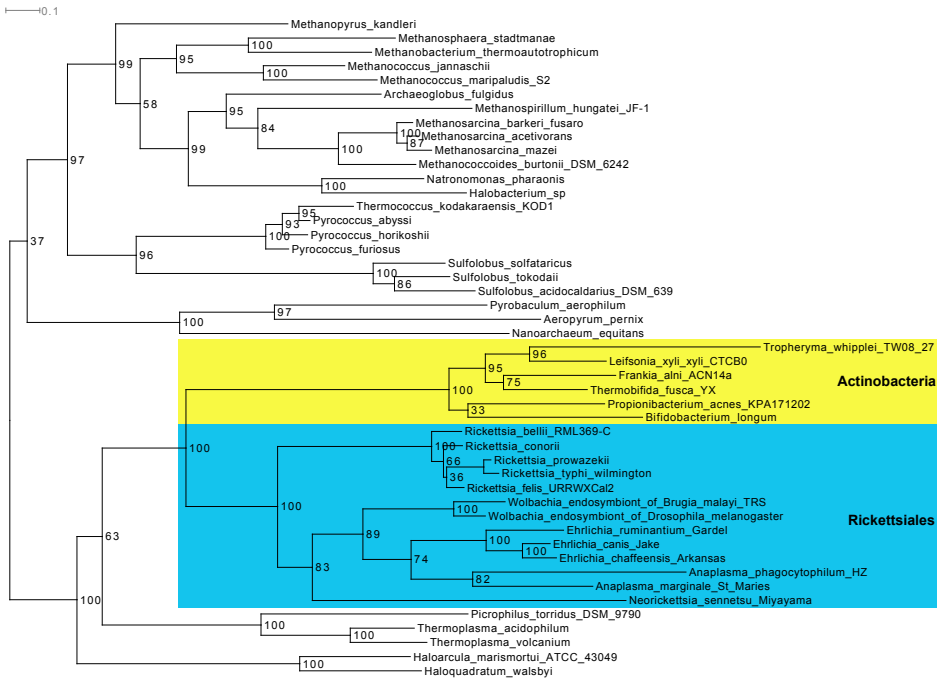


Figure 3.10: Section of the Valyl tRNA synthetase tree showing the position of certain Rickettsiales and Actinobacteria within the Archaeal clade. The figure was created using Dendroscope (Huson et al. 2007c).

an ubiquituous protein functioning as a GTP-dependent translation factor (Galperin and Koonin 2004; Caldon et al. 2001).

The YchF gene is often used to infer a species phylogeny based on supermatrix approaches (e.g., Bapteste et al. 2008; Marri et al. 2007; Ciccarelli et al. 2006, equates to COG0012). Anyhow, in current literature, no information could be found that confirms our findings.

Missed events

Kanhere and Vingron (2009) tested the Cook's distance using a dataset of Aminoacyl tRNA synthetases by comparing the results against well-known cases of HGT (Kanhere and Vingron 2009, Table 1). In addition to the events described here, they could detect the transfer of Phenylalanyl tRNA synthetase from Archaea to Spirochaetes. This HGT event was previously described by Woese et al. (2000) for both, α and β subunit of this protein.

In our data, the gene trees for both subunits clearly show these events, consequently leading to the detection of this event by the topological method, as well as by the Cook's distance. However, the `ParaFit` p-value obviously was below the threshold ($p = 0.001$ for *Treponema pallidum* and *T. denticola*).

3.4 Conclusions

We downloaded gene sequences of 279 prokaryotic organisms and tried to find a common set of genes that can be used for phylogenetic inference. Due to the application of a strict protocol to remove paralogs and the requirement that each gene has to be present in all included genomes, only a small set of 17 common genes could be detected unequivocally, consisting mostly of genes coding tRNA synthetases and ribosomal proteins. Individual gene sequences were aligned and afterwards filtered by applying `Gblocks`. We tested the effect of alignment filtering by comparing both, trees based on filtering, as well as trees based on the original alignments, to the NCBI taxonomy. Generally speaking, it could be shown that the pruning process resulted in trees that were more distant to the NCBI taxonomy. We suppose that this is an indication that the Maximum Likelihood reconstruction method is able to extract information even from poorly aligned sites, an observation also reported by Talavera and Castresana (2007).

Inferred trees were compared to a species tree derived by concatenating all single gene alignments into a supermatrix, and to the NCBI taxonomy. A considerable agreement between single gene trees and the NCBI taxonomy could be observed, as well as between the unconstrained supermatrix tree and the NCBI taxonomy. Considering that the NCBI taxonomy is broadly based on 16S rRNA trees, this indicates the presence of a strong coherent phylogenetic signal in all genes. With exception of the T-IF 2 gene, the

Threonyl-tRNA synthetase						
			BML		MLNC	
Recipient	Donor clade	rank	AxParafit *p*-value	CDISS	AxParafit *p*-value	CDISS
Aeropyrum pernix	Proteobacteria	phylum	1	+	0.999	+
Prochlorococcus marinus CCMP1375 †	Proteobacteria	phylum	0.958	+	0.966	−

GTP-binding protein, YchF family						
			BML		MLNC	
Recipient	Donor clade	rank	AxParafit *p*-value	CDISS	AxParafit *p*-value	CDISS
Moorella thermoacetica ATCC 39073 †	Proteobacteria	phylum	0.108	+	0.102	−
Syntrophomonas wolfei Goettingen	Proteobacteria	phylum	0.099	+	0.096	+
Thermotoga maritima	Proteobacteria	phylum	0.060	+	0.056	+

Valyl-tRNA synthetase						
			BML		MLNC	
Recipient	Donor clade	rank	AxParafit *p*-value	CDISS	AxParafit *p*-value	CDISS
Anaplasma marginale St Maries	Euryarchaeota	phylum	0.224	+	0.189	+
Anaplasma phagocytophilum HZ	Euryarchaeota	phylum	0.210	+	0.199	+
Ehrlichia canis Jake	Euryarchaeota	phylum	0.175	+	0.142	+
Ehrlichia chaffeensis Arkansas	Euryarchaeota	phylum	0.180	+	0.170	+
Ehrlichia ruminantium Gardel	Euryarchaeota	phylum	0.187	+	0.143	+
Neorickettsia sennetsu Miyayama	Euryarchaeota	phylum	0.753	+	0.729	+
Rickettsia bellii RML369-C	Euryarchaeota	phylum	0.371	+	0.264	+
Rickettsia conorii	Euryarchaeota	phylum	0.400	+	0.289	+
Rickettsia felis URRWXCal2	Euryarchaeota	phylum	0.394	+	0.284	+
Rickettsia prowazekii	Euryarchaeota	phylum	0.410	+	0.301	+
Rickettsia typhi wilmington	Euryarchaeota	phylum	0.435	+	0.300	+
Wolbachia endosymbiont of Brugia malayi TRS	Euryarchaeota	phylum	0.226	+	0.224	+
Wolbachia endosymbiont of Drosophila melanogaster	Euryarchaeota	phylum	0.249	+	0.214	+
Bifidobacterium longum ∗	Euryarchaeota	phylum	**0.020**	+	0.053	+
Propionibacterium acnes KPA171202 ∗	Euryarchaeota	phylum	**0.018**	+	0.065	+
Tropheryma whipplei TW08/27 ∗	Euryarchaeota	phylum	**0.023**	+	0.091	+
Frankia alni ACN14a ∗	Euryarchaeota	phylum	**0.007**	+	0.05	+
Leifsonia xyli xyli CTCB0 ∗	Euryarchaeota	phylum	**0.016**	+	0.04	+
Thermobifida fusca YX ∗	Euryarchaeota	phylum	**0.018**	+	0.04	+

Table 3.8: Hypothetical HGT events detected by all three methods. A "+" in the CDISS column means that the corresponding CDISS value was above the HGT detection threshold.
† HGT events that were not detected when using the MLNC reference tree (constrained by the NCBI taxonomy).
∗ HGT events that were not detected when using the best ML tree.

supermatrix tree had a higher resemblance to the NCBI taxonomy than any other single gene tree. Given that vertical inheritance overweights any other signal, this shows that the supermatrix method can even be used when there is a distinct amount of disagreement between single gene trees. This finding is endorsed by the fact that more than 75% of predicted HGT occurrences could be found using the NCBI-constrained, as well as the unconstrained reference tree.

The amount of disagreement that may be caused by horizontal gene transfer (i.e., the proportion of leaves affected by HGT) was analyzed by applying three different methods. We used the Cook's distance (CDISS) introduced by Kanhere and Vingron (2009), as well as a statistical test for host-parasite cophylogeny (Legendre et al. 2002). Additionally, a method based on the comparison of single gene trees with a taxonomy was developed by our group, that directly analyzes clades defined by the topology of an individual gene tree. The method was tested using the NCBI taxonomy.

Interestingly, there was a high number of predicted occurrences by the different methods, ranging from 250 to more than 500 hypothetical occurrences of HGT over all 17 genes, whereas the overlap between all three tests was restricted to 19 occurrences only. Most of these common occurrences were observed for the gene of Valyl tRNA synthetase and are well established in current literature. In contrast, no indication could be found for two hypothetical HGT events predicted for the ychF gene. But overall, we could show that the intersection between all three methods produces a robust set of HGT candidates. This observation could be made with both, the unconstrained as well as the NCBI-constrained reference tree based on the supermatrix of all genes.

By restricting the analysis to a common, and thus, essential set of genes, only xenologous gene displacements (XGD) can be recognized, which involve HGT on the one hand, and loss of the original gene on the other hand. Thus, XGD may be less frequent than the adoption and retention of a new gene by an organism (Koonin and Wolf 2008). Additionally, the complexity hypothesis as postulated by Jain et al. (1999) predicates that informational genes, whose products are involved in many complex interactions with other proteins, are less frequently transferred as, e.g., metabolic genes. Although assigned to the group of informational genes, tRNA synthetases only interact with a narrow area of the ribosome and also show a substantial pattern of HGT (Woese et al. 2000; Woese 2002; O'Donoghue and Luthey-Schulten 2003; Farahi et al. 2004; Beiko et al. 2005). But even ribosomal proteins are occasionally affected by HGT (Koonin and Wolf 2008), and so far, there is no evidence for an absolutely untransferable gene (Sorek et al. 2007). Yet it remains controversial whether informational genes are less affected by HGT than operational genes, as predicted by the complexity hypothesis (Kanhere and Vingron 2009; Koonin and Wolf 2008; Choi and Kim 2007; Beiko et al. 2005; Nakamura et al. 2004). Therefore, the number of occurrences may

reflect the real amount of genes that were affected by HGT in this set, even when considering that the composite of all three methods contains more than 500 possible occurrences. This high amount of predicted occurences may coincide with the findings of other groups (Woese 2002; Zhaxybayeva et al. 2006; Dagan and Martin 2007; Dagan et al. 2008).

On the other hand, there also exist methodological biases that may lead to the detection of spurious incidents. The Cook's distance (CDISS) was invented to detect outliers in a correlation based on linear regression (Cook 1979). Thus, a linear relation between gene and species distances is an important precondition for the application of this method. Consequently, Kanhere and Vingron (2009) explicate that protein families evolving at inconstant rates cannot be analyzed using the Cook's distance method. Hence, genes showing a considerable fraction of heterotachous sites may lead to the detection of outliers that are wrongly considered as candidates for HGT. Such occurrences of accelerated evolutionary rates for translational proteins are known for certain taxa and can arise by adaptation to an extreme environment (Cavalier-Smith 2002). But, considering the presence of long branches between clades, a large amount of outliers may also lead to a decrease of the Cook's distance when the Mean squared error grows (see Equation 2 in Kanhere and Vingron 2009).

Strikingly, there is a negative correlation between ML-conflict and the amount of observed HGT using the CDISS method. On the whole, this means that a lower amount of HGT is detected in trees showing a larger disagreement with the reference tree, and vice versa. A positive correlation between ML-conflict and CDISS HGT counts could only be achieved using a reduced cut-off value of $\frac{1.7}{D}$ or below (see Figures 3.6, and 3.7). But lowering the threshold leads to a considerable increase of HGT counts, and unavoidably, to an increase in the rate of false positives. We assume that the negative correlation indicates that the CDISS measure seems to be more affected by the presence of heterotachous sites than the `ParaFit` method, precisely because the former method is developed to detect outliers.

When regarding the `AxParafit` results, it has to be considered that the `AxParafit` method may miss HGT occurrences when a whole clade of a considerable size may be misplaced, but the topology of the clade's subtree resembles the equivalent subtree in the reference tree. In that case, some, if not all, of the concerned taxa may significantly contribute to the overall cophylogenetic structure, even if the position of the whole clade deviates from the reference. But this is an assumption that has to be tested yet by conducting a simulation study. Furthermore, since `AxParafit` focuses on distances rather than tree-topology, non-proportional deviations of edge lengths between gene and species tree may also affect the outcome. One potential source of such deviations may be Heterotachy. Certainly, as a result of the high correlation between ML-conflict and `AxParafit`'s HGT counts ($r_{ps} > 0.9$), we conclude that the influence of heterotachy to the

outcome of the `AxParafit`-based HGT detection method is considerably lower, compared to the CDISS method's results.

Nonetheless, the correlation between ML-conflict and `AxParafit` results indicates that `AxParafit` can be used to detect taxa that negatively affect congruence between gene and species tree. A taxon missing significance for cophylogenetic association thus may be a reasonable candidate for HGT.

The significance threshold was determined by optimizing correlation between determined HGT counts and ML-conflict. A threshold of $\alpha = 0.025$ was found to be optimal using the Pearson correlation coefficient. Using Kendall's τ, an $\alpha = 0.02$ performed slightly, but insignificantly better, allowing us to propose a threshold of 0.025 that is in a typical range for a significance threshold in applied statistics. However, the null hypothesis of the `AxParafit` test states that there does not exist a cophylogenetic structure between an individual gene and its organism, represented by the corresponding taxon in the reference tree. Hence, the concerned leaf is considered to be a candidate for HGT if the null hypothesis cannot be rejected. The statistical power of the test is defined as the rate of correct rejections of the null hypothesis when the very same is wrong. Correspondingly, a high power of the `AxParafit` test indicates a low amount of false positives in HGT detection. Thus, the statistical power is a crucial measure of the suitability of this test for HGT detection.

Legendre et al. (2002) conducted several simulation studies to determine the power of the `ParaFit` test. In one test, trees were generated that share a common part, whereas the remaining taxa were placed differently. This scenario resembles the empirical situation when a large part of the tree shows a pattern similar to the reference tree, while some taxa deviate due to HGT. In that case, results of Legendre et al. (2002) indicate that the statistical power mainly depends on the fraction of species showing a cophylogenetic structure. The c-score measure clearly shows that the gene trees are mostly in agreement with the reference tree represented by the NCBI taxonomy (see Table 3.4, page 75). Hence, it can be assumed that the cophylogenetic structure is prevailing in all gene trees. But the `ParaFit` test itself provides a well-suited method to clarify whether the fraction of cophylogenetic species is adequate: the global significance test for cophylogeny. Consequently, if the global test shows a significant cophylogenetic structure between gene tree and reference tree, the individual tests should provide sufficient statistical power to ensure a low rate of false positives in HGT detection. Using our empirical dataset, all global tests were significant using $\alpha = 0.01$, which corroborates our conclusion based on the gene trees' c-scores.

In addition to these two distance-based methods, we proposed an approach depending on the comparison between a gene tree and a taxonomy. For each taxonomic group, a corresponding clade in the gene tree is located. Taxa that are not correctly placed in the associated clade are identified as potential HGT candidates. Thus, clade identification is the most crucial

part of the algorithm, which we tried to accomplish by searching a node that provides the best balance between clade tightness and completeness. Likewise, quality of the taxonomic source affects the outcome of the algorithm, considering that taxonomic ranks are, at least to some level, arbitrary, and undergo frequent modifications and refinements. This may also explain that correlations between ML-conflict and HGT counts were considerably lower than results obtained by using `ParaFit`, while they were noticeably better than those of the CDISS method. Furthermore, the method is focused on tree topology, whereas edge lengths are not considered. Nevertheless, restricting the HGT search process to the tree topology has its strength, since the method is not affected by branch length differences, possibly caused by heterotachy. At least, this is valid as long as the underlying tree reconstruction method is unsusceptible against heterotachy. Of course, distance-based methods suffer from such differences, which makes this a prominent feature of topology-based methods. While this omits detection of false positives due to heterotachy, there exist candidates for false negatives that are not easily mastered. Taxa that could not be successfully assigned to a specific clade, e.g., due to polytomies, are simply overlooked by the current implementation.

Consequently, future work has to be focused in refining this selection process to avoid both, the detection of false positives and the oversight of HGT events. Furthermore, combination of the topological method with a statistical method like `ParaFit` may also be a possibility to enhance accurracy of our HGT detection method, considering the overlap between both methods. On the one hand, this omits susceptibility to heterotachy, while on the other hand, it allows to compensate for ambiguous taxon classification due to polytomies. Moreover, run time efficient implementations of the `ParaFit` algorithm are available (Stamatakis et al. 2007; Stockinger et al. 2009), whereas the topology-based method is computationally undemanding.

Chapter 4

Cophylogenetic studies

4.1 Introduction

4.1.1 Biological background

Terminology

In phylogeny, extant species are treated as independent entities. While trying to reconstruct relationships between species based on common ancestry, dependencies between communities of species are only regarded in the light of horizontal gene transfer. However, in nature, organisms are embedded in a tight web of dependencies with cohabiting life-forms and their environment, thus forming an ecosystem. Such dependencies are: the food chain, host and parasite, as well as commensal or symbiotic relationships. It can be assumed that evolutionary forces shaping one species will also affect other species that are either directly coupled as, e.g., in a predator-prey or host-parasite relationship, or due to intermediate factors like the competition for shared resources. Consequently, such mutual evolutionary dependencies can be described as coevolution. More precisely, coevolution is a process of mutual influence between participating species, where one change in a species induces changes in another species, which eventually causes the first species to adapt again to the changed environment (Janzen 1985), thus depicting a recurring scheme.

A related concept is cospeciation, denoting "the joint speciation of two or more lineages that are ecologically associated" (Page 2003). It has to be considered that there is no perfect accordance between these two terms (Göker 2003). On the one hand, coevolution does not necessarily involve speciation, and thus, cospeciation. On the other hand, joint speciation of two lineages may be driven by other processes like, e.g., genetic drift of a host's immunity gene leading to unilateral adaptational pressure on the parasite. Unfortunately, both concepts are not deducible by phylogenetic methods, since an actual mutual evolutionary dependency or an ecological association

cannot be verified in this way (Göker 2003). When using phylogenies, historical dependencies of speciations cannot be established. Host speciation followed by colonization of host species by a parasite that quickly adapts to the different hosts may lead to congruent host and parasite phylogenies, while not actually depicting a cospeciation event. In contrast, cospeciation should induce a congruent phylogeny, i.e., a cophylogenetic structure. Thus the lack of cophylogenetic structure may be seen as contradicting a cospeciation hypothesis.

Consequently, we use the term "cophylogeny" instead of "cospeciation" in the following, when referring to congruent phylogenies.

Historical associations between hosts and parasites

In the broader sense, historical associations between entities in biology can be classified into the following groups (Page and Charleston 1998): organisms and organisms, like hosts and parasites; genes and organisms, e.g., when looking for deviations between gene and species phylogenies (see Section 3.2.1, p. 63); and organisms and areas (Biogeography and Paleobiogeography). All three can be described analogously, since there is a direct correspondence. Hence, in the following, the terms host and parasite also represent the associations between organisms and their genes, as well as between areas and their inhabitants.

Associations between hosts and their parasites may change during time. These changes basically can be modeled by four events (see Figure 4.1):

- *Cospeciation.* Host speciation is closely followed by speciation of the parasite.

- *Duplication.* Speciation of the parasite only, without a corresponding cladogenesis concerning the host.

- *Lineage sorting.* Extinction of the parasite while the host persists. This is in contrast to extinction of the host, which also leads either to extinction of the parasite, or host switching (see below).

- *Host switching.* The parasite colonizes a new host species, while either leaving the original host (complete host switching) or colonizing both, the old as well as the new host (partial host switching). Complete host switching is a particular case of partial host switching followed by extinction in the old host.

It is assumed that cospeciation may be the predominant factor shaping the history of host-parasite associations (Fahrenholz' rule, see Eichler 1948). But to date, this rule could not be thoroughly investigated by conducting

large-scale studies of host-parasite associations covering major phyla. Coevolution and cospeciation are not directly approachable by using cophylogenetic methods, but a falsification of the Fahrenholz rule may be contrivable if there is no evidence for cophylogenetic structure in a large-scale analysis.

In recent years, many phylogenetic reconstruction tools were developed that are performant and adequately parallelized to be used with large datasets (see, e.g., Stamatakis 2006b; Stamatakis et al. 2008), whereas there was no such tool for computational analysis of cophylogenetic relationships. Accordingly, our objective was to provide computational tools making large-scale cophylogenetic analyses feasible.

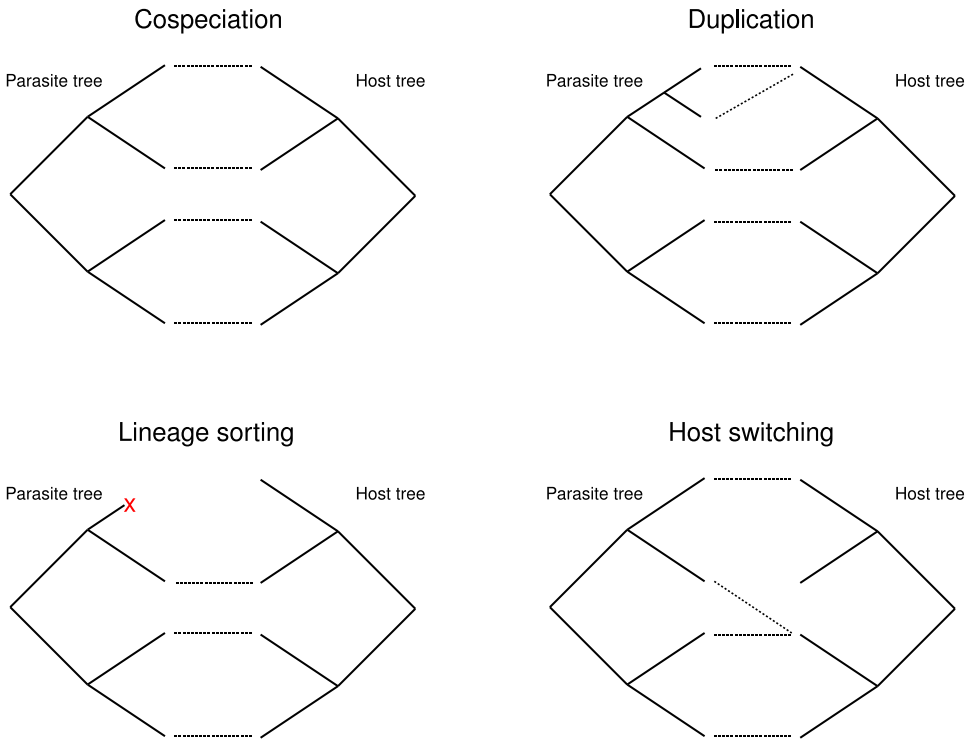


Figure 4.1: Four basic kinds of events that may occur during the history of host-parasite associations: cospeciation, duplication, lineage sorting and host switching. A thorough model can be found in Begerow et al. (2004, Fig. 2).

4.1.2 Technical background

Computational methods for host-parasite cophylogenies

Though it seems feasible to compare small host and parasite phylogenies by hand, computational methods are needed to study deep cophylogenetic relationships comprising many different families, orders, or even phyla. A plenitude of such computational methods exist, of which the best-known methods are described in the following (see also reviews in Stevens 2004; Charleston and Perkins 2006).

One of the oldest and most known methods is the Brooks Parsimony (BPA, Brooks 1981; 1990). Here, the nodes of the parasite tree are coded as a binary matrix representation (see Felsenstein 2004, p. 541 for an example). Afterwards, the characters of the matrix are reconstructed on the host tree by parsimony, allowing to optimize the host tree in the light of parasite phylogeny. Other parsimony approaches use an event cost model, which is based on assigning a cost value to each possible event affecting historical host-parasite associations (see Figure 4.1), as implemented in `TreeFitter` (Ronquist 2001).

Another popular group of algorithms depend on reconciled tree analysis. The main idea behind tree reconciliation is to reconcile several incongruent trees into a tree showing a minimum of conflict. Several different heuristics exist to address this time-consuming problem. Common implementations are `Component` (Page 1990), `TreeMap` (Charleston 1998; Charleston and Page 2002), and `Tarzan` (Merkle and Middendorf 2005).

Furthermore, statistical frameworks are available to test for congruence of host and parasite phylogenies, which are based on Maximum likelihood (Huelsenbeck et al. 1997), or Bayesian analysis (Huelsenbeck et al. 2000). But these tests are specifically designed to handle bijective relations only. Here, bijectivity means that one-to-one associations are merely allowed, i.e., each parasite can only be associated to a single host, and vice versa.

Fast methods for tree comparison, like I_{cong} (de Vienne et al. 2007) are also limited to bijective associations between parasite and host taxa.

However, all of these methods are not applicable to large-scale cophylogenetic datasets, due to memory consumption or run time considerations, due to the fact that their statistical properties are simply unknown in these circumstances (Stamatakis et al. 2007; Stockinger et al. 2009), or because they do not support multiple associations between the same parasite and different hosts, or vice versa. Especially the latter restriction diminishes practicability of these method, since associations of parasites that can colonize a multitude of hosts become hard to model. For example, approximately 45% of european smut fungi are known to colonize more than a single host species (Begerow et al. 2004), which underlines that allowing bijective associations only, is a severe limitation for large-scale empirical studies.

In contrast to the aforementioned methods, `ParaFit` (Legendre et al. 2002) is explicitly designed to allow for parasites colonizing several host species, or for hosts that are colonized by more than one parasite species. `ParaFit` has been successfully applied in several empirical studies (Hansen et al. 2003; Ricklefs et al. 2004; Meinilä et al. 2004; Refrégier et al. 2008; Garamszegi 2009). Furthermore, the method's statistical properties were thoroughly tested in several simulation studies. Legendre et al. (2002) showed that the `ParaFit` test yields acceptable rates for type I and type II errors. Furthermore, they demonstrated that the power of the test (1 - type II error rate) increases with the size of the dataset, which makes this tool well-suited for large-scale analyses.

Therefore, we decided to focus on the `ParaFit` algorithm, which is outlined in the next section.

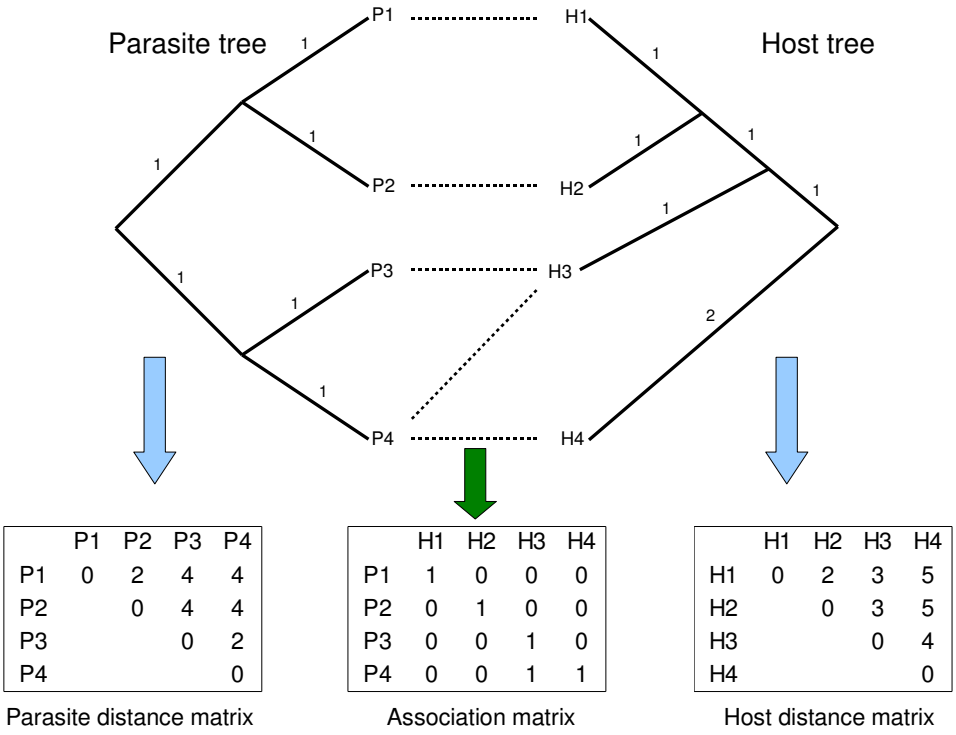
Description of the Parafit test

Legendre et al. (2002) introduced `ParaFit`, a program that implements a statistical test for host-parasite cophylogeny. The null hypothesis of the statistical test is that the evolution represented by the corresponding phylogenies of hosts and parasites has been independent.

Prior to conducting the `ParaFit` test, phylogenies have to be transformed into distance matrices, either based on patristic distances or topological distances. Patristic distances simply reflect the sum of all edge lengths of the path between two taxa (e.g., see distance matrices in Figure 4.2). When using topological distances, all edge lengths are set to 1, so that the corresponding patristic distance matrix reflects tree topology only. Additionally, genetic distances can also directly be used with `ParaFit`.

Before the `ParaFit` test is conducted, host and parasite distance matrices have to be transformed into rectangular matrices by Principal Coordinate Analysis (PCoA, see Legendre and Legendre 1998, p. 424-426). This also reduces the size of the matrices, which leads to reduced space and run time demands of the subsequent `ParaFit` analysis. In contrast to the widely known Principal Component Analysis, PCoA can also be applied to distances derived from semi-quantitative variables (Legendre and Legendre 1998, p. 388), such as topological distances based on taxonomical rankings. An implementation of the PCoA algorithm called `DistPCoA` is provided by Legendre and Anderson (1998).

After providing host and parasite principal coordinate matrices, as well as a 0-1 encoded association matrix (comparable to an adjacency matrix) representing the host-parasite associations (see Figure 4.2), the `ParaFit` test can be performed. It consists of two matrix multiplications to obtain the fourth-corner parameters (Legendre et al. 2002; Legendre and Legendre 1998, p. 565-574) between host and parasite phylogenies. The new matrix simply is a cross-product of the principal coordinate matrices weighted by



	P1	P2	P3	P4
P1	0	2	4	4
P2		0	4	4
P3			0	2
P4				0

Parasite distance matrix

	H1	H2	H3	H4
P1	1	0	0	0
P2	0	1	0	0
P3	0	0	1	0
P4	0	0	1	1

Association matrix

	H1	H2	H3	H4
H1	0	2	3	5
H2		0	3	5
H3			0	4
H4				0

Host distance matrix

Figure 4.2: Data preparation steps for a `ParaFit` run. Phylogenetic trees of hosts and parasites are converted to patristic distance matrices, whereas the associations are coded into a 0-1 matrix, analogous to an adjacency matrix. Drawing corresponds to Legendre et al. (2002, Figure 1).

the corresponding association matrix. It can further be converted into a single value by taking the sum of squares of the matrix values, which is referred to as `ParafitGlobal` (Legendre et al. 2002). The `ParafitGlobal` value gets larger in case of higher congruence between the trees, given the associations. Additionally, for each host-parasite association, its contribution to the `ParafitGlobal` value can be calculated by removing the association and re-calculating `ParafitGlobal`. This allows to obtain a value describing the individual contribution of the corresponding association to the global value (named `ParafitLink1` and `ParafitLink2`, whereas the latter is scaled appropriately). The test is then repeated by permuting the values within each row of the association matrix, resulting in an estimate of the distribution of the `ParafitGlobal` value. Permuting the association matrix leads to randomized associations, allowing to test whether the original `ParafitGlobal` is significantly better than values obtained from random associations. Eventually, the fraction of all values that are larger than the original `ParafitGlobal` is determined, which can be interpreted as the p-value of the statistical test under the null hypothesis (stating that both phylogenies are independent). Thus, if the original `ParafitGlobal` value is larger than or equal to $1 - \alpha$ of the obtained values (e.g., 99% for $\alpha = 0.01$), the null hypothesis is rejected. Likewise, individual significances are calculated by obtaining estimates of the distributions of `ParafitLink1` and `ParafitLink2`.

Since `ParaFit` does not allow to provide taxon names with the input distance matrices, its output consists of a list of associations, having the taxa labelled in input order (e.g., "Host 1", "Parasite 1"). These labels have to be mapped by hand to the corresponding host and parasite, which constitutes a rather error-prone procedure. In addition, `ParaFit` only provides a basic console menu complicating the handling for users. Hence, we decided to develop an easy-to-use graphical frontend (named `CopyCat`, see Meier-Kolthoff et al. 2007) that allows biologists to prepare their data in a convenient way, and to conduct the analysis in a user-friendly manner. Furthermore, the application to large-scale datasets required optimizations of the `ParaFit` and `DistPCoA` programs, including adaptation to Cluster and Grid environments (see Stockinger 2006; 2007, for an introduction into this topic).

4.2 Methods

4.2.1 Large-scale cophylogenetic studies with CopyCat

As outlined in the previous Section, we decided to design an easy-to-use graphical frontend on top of the command line applications `ParaFit` and `DistPCoA`, which we named `CopyCat` (Meier-Kolthoff et al. 2007).

For a typical cophylogenetic analysis using `ParaFit`, the user has to prepare three different matrices (see Figure 4.2). These are the host and parasite distance matrices, as well as the association matrix. The first two can be obtained either by directly using genetic distances or by computing patristic distances from a phylogenetic tree. In many cases, tree inference is done by collecting specific marker sequences, aligning them into a Multiple Sequence Alignment, and applying a tree reconstruction algorithm. Frequently used markers are 16S rRNA (Woese 1987), ITS (Internal Transcribed Spacer, Göker et al. 2009), and ribosomal proteins (Henz et al. 2004).

However, the necessity to provide phylogenetic data for a set of associated hosts and parasites may constitute an obstacle to the scientist who may be primarily engaged in gathering the available association data. In practice, collecting all sequences needed to infer a host or parasite tree can be a rather challenging, error-prone, and above all, time consuming task. `CopyCat` greatly reduces workload by offering the option to derive host and parasite distances from taxonomical data based on the NCBI taxonomy (NCBI 2009c). The user can decide to provide a distance matrix, a phylogenetic tree, which is automatically converted to a patristic distance matrix, or whether distances shall be inferred from taxonomical information.

In the latter case, the user only has to prepare the association file, which can be provided as a simple adjacency list containing the scientific names of associated host and parasite species. `CopyCat` tries to match these names by comparing them to the NCBI taxonomy (NCBI 2009c). Non-matchable entries are highlighted so that the user can change those and re-run the matching process. Additionally, a NCBI taxon ID can also be specified directly. To support the handling of large datasets, identified taxa can be filtered by narrowing them to certain systematic divisions as defined by the NCBI (like Bacteria, Invertebrates, Mammals etc.), or by performing a taxonomic reduction to genera or families (see Figure 4.3). Afterwards, a taxonomic tree is generated comprising all selected taxa.

However, taxonomical trees mostly contain polytomies, in contrast to binary trees as obtained by using phylogenetic tree reconstruction algorithms. Since this may influence the outcome of a `ParaFit` analysis, `CopyCat` computes several tree statistics prior to conducting the time-consuming analysis, allowing the user to decide whether it is necessary to further improve taxon sampling. These statistics comprise tree resolution and balance (Colless 1982), cherry count (McKenzie and Steel 2000) and cladistic information content (Thorley and Page 2000). Furthermore, to reduce bias in taxon sampling, `CopyCat` implements a model based on the broken stick distribution (Legendre and Legendre 1998, p. 244) to detect species that are over-represented in respect of the proportion of associations in the dataset.

After filtering, `CopyCat` automatically invokes `DistPCoA` (or `AxPcoords`, see Section 4.2.2), which computes the principal coordinates of the distance matrices. The user can then decide to either invoke `ParaFit` (or `AxParafit`)

on the local machine, or whether CopyCat shall generate a zip file containing all necessary files to start the analysis on a remote machine.

Eventually, CopyCat merges the ParaFit output with the original taxon names, and displays the associations together with the corresponding significance values. Significant associations are highlighted in compliance to a threshold, which can be altered by the user (see Figure 4.4).

CopyCat was developed in Java using the highly performant SWT library (Standard Widget Toolkit) as graphics engine (see, e.g., Guojie 2005). Versions for Linux, Windows and MacOS are provided at http://www-ab.informatik.uni-tuebingen.de/software/copycat. Use of the program is free for academic purposes.

4.2.2 AxParafit and AxPcoords

As a first step prior to conducting large-scale cophylogenetic studies, we established CopyCat as a user-friendly frontend for the command line driven tools DistPCoA and ParaFit. As described above, CopyCat also allows to automatically infer distances between hosts and parasites by using taxonomical data. This simplification should not be underestimated, since it relieves the user from the burden to collect marker sequences for inferring trees, which becomes unfeasible for large taxon sets.

After establishing adequate prerequisites, we focussed our work on performance improvements of DistPCoA and ParaFit, since the execution time of those applications dominates the overall runtime of the CopyCat analysis pipeline. As a first step, the original Fortran sources were ported to C. The code was carefully optimized with regard to runtime and memory consumption (for details, see Stamatakis et al. 2007).

A considerable performance improvement was accomplished by using highly optimized matrix multiplication routines from the BLAS (Basic Linear Algebra Package) library, and eigenvector/eigenvalue decomposition routines from the LAPACK (Linear Algebra PACKage) library. For this purpose, processor-independent implementations of BLAS and LAPACK were used, in particular the ATLAS Library (Whaley and Petitet 2005) and the GNU scientific Library (GSL). Furthermore, specifically optimized versions of BLAS and LAPACK for AMD and Intel processors were incorporated and evaluated in regard to potential runtime benefits. Namely, these were the AMD Core Math Library (ACML), as well as the Intel Math Kernel Library (MKL), which both are freely available for academic use. Using the LAPACK routines lead not only to considerable runtime improvements but also to improved numerical stability compared to the original implementation (Stamatakis et al. 2007).

While AxParafit and AxPcoords are specifically designed to interact with CopyCat, they can also be used independently. Both programs provide command line options, which are mostly equivalent to the functionality of

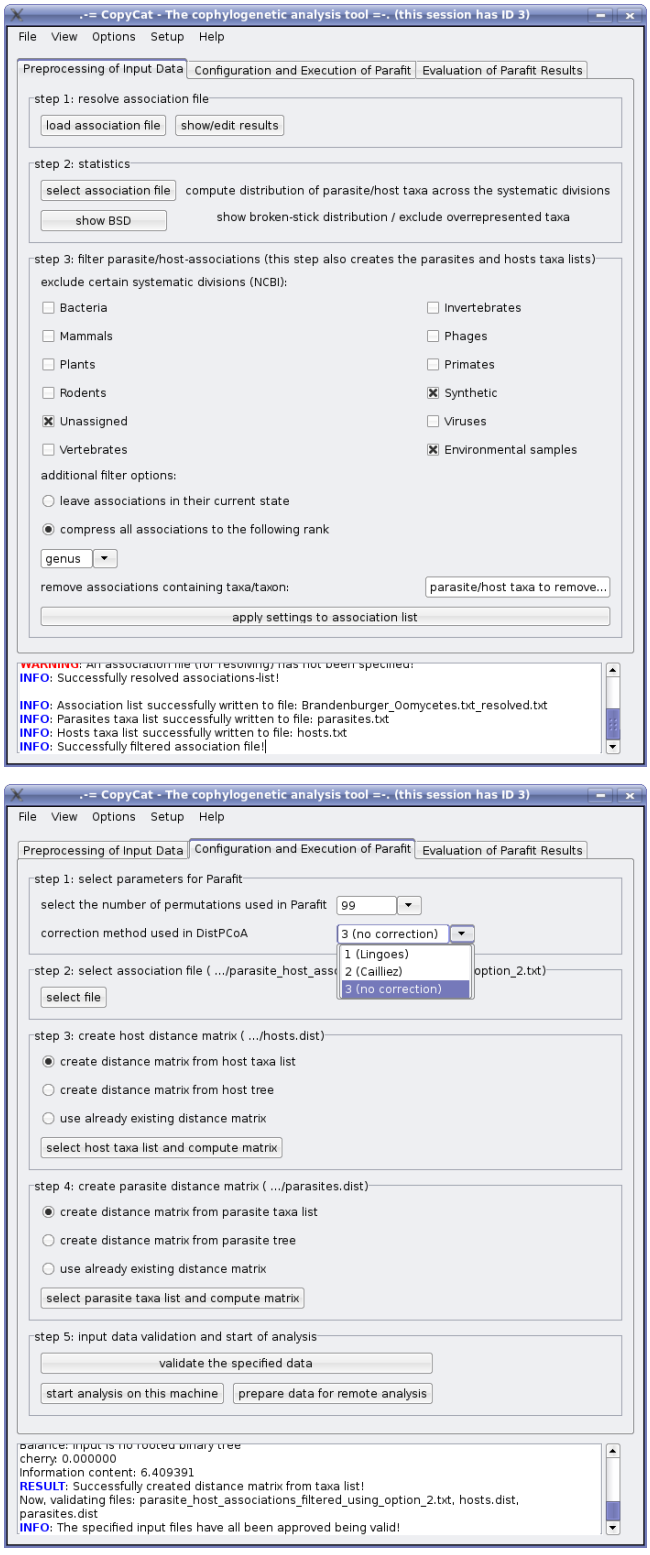
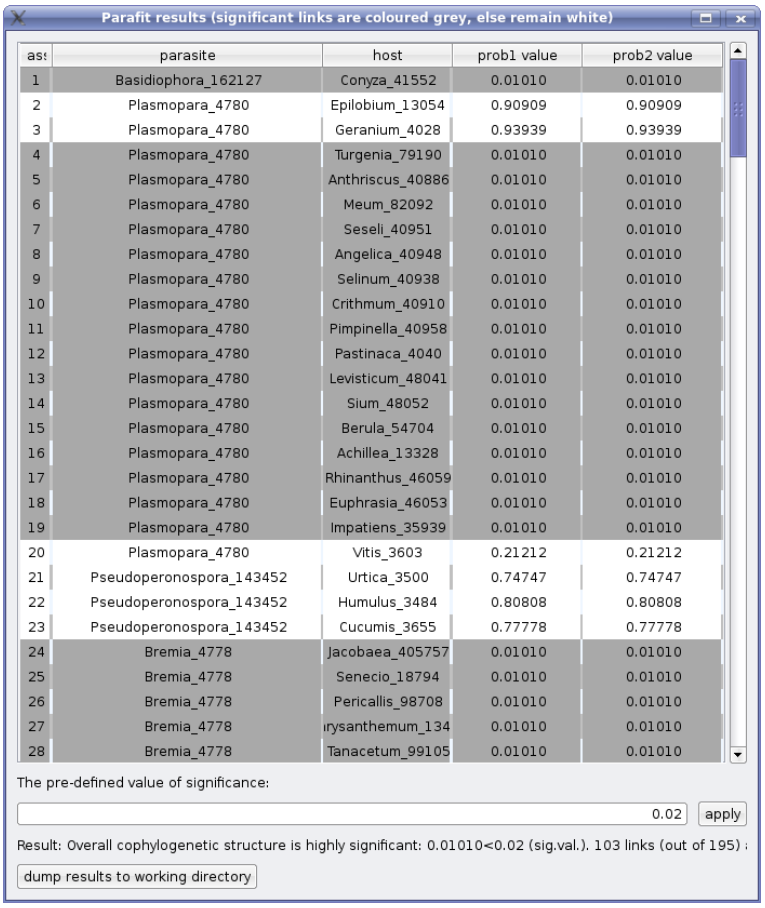


Figure 4.3: Screenshots showing the two configuration tabs of the CopyCat main window.



as:	parasite	host	prob1 value	prob2 value
1	Basidiophora_162127	Conyza_41552	0.01010	0.01010
2	Plasmopara_4780	Epilobium_13054	0.90909	0.90909
3	Plasmopara_4780	Geranium_4028	0.93939	0.93939
4	Plasmopara_4780	Turgenia_79190	0.01010	0.01010
5	Plasmopara_4780	Anthriscus_40886	0.01010	0.01010
6	Plasmopara_4780	Meum_82092	0.01010	0.01010
7	Plasmopara_4780	Seseli_40951	0.01010	0.01010
8	Plasmopara_4780	Angelica_40948	0.01010	0.01010
9	Plasmopara_4780	Selinum_40938	0.01010	0.01010
10	Plasmopara_4780	Crithmum_40910	0.01010	0.01010
11	Plasmopara_4780	Pimpinella_40958	0.01010	0.01010
12	Plasmopara_4780	Pastinaca_4040	0.01010	0.01010
13	Plasmopara_4780	Levisticum_48041	0.01010	0.01010
14	Plasmopara_4780	Sium_48052	0.01010	0.01010
15	Plasmopara_4780	Berula_54704	0.01010	0.01010
16	Plasmopara_4780	Achillea_13328	0.01010	0.01010
17	Plasmopara_4780	Rhinanthus_46059	0.01010	0.01010
18	Plasmopara_4780	Euphrasia_46053	0.01010	0.01010
19	Plasmopara_4780	Impatiens_35939	0.01010	0.01010
20	Plasmopara_4780	Vitis_3603	0.21212	0.21212
21	Pseudoperonospora_143452	Urtica_3500	0.74747	0.74747
22	Pseudoperonospora_143452	Humulus_3484	0.80808	0.80808
23	Pseudoperonospora_143452	Cucumis_3655	0.77778	0.77778
24	Bremia_4778	Jacobaea_405757	0.01010	0.01010
25	Bremia_4778	Senecio_18794	0.01010	0.01010
26	Bremia_4778	Pericallis_98708	0.01010	0.01010
27	Bremia_4778	irysanthemum_134	0.01010	0.01010
28	Bremia_4778	Tanacetum_99105	0.01010	0.01010

Figure 4.4: Screenshot of the `CopyCat` dialog window depicting the results of a Parafit run. Significant links are grey while insignificant links are drawn in white.

ParaFit's and DistPCoA's text mode menu. AxParafit and AxPcoords can be downloaded from the CopyCat homepage together with the current CopyCat version (Meier-Kolthoff et al. 2007). Sources are freely available from http://icwww.epfl.ch/~stamatak/AxParafit.html.

4.2.3 Parallelized AxParafit

While AxPcoords needs less than 24 hours on a single AMD 2.4 GHz Opteron CPU even for large matrices with several thousands of taxa, AxParafit's runtime demands required a parallelization of the most time-consuming steps. Let $A_{\neq 0}$ denote the number of non-zero elements in the association matrix, i.e. the number of associations. Further, let i be the number of iterations AxParafit has to perform (alterable by the user, usually between 100 to $10,000$), and m be the time complexity of AxParafit's main computational step, a dense matrix multiplication (see Legendre et al. 2002; Stamatakis et al. 2007). Considering that the matrix dimensions' upper bound is n with n denoting the number of hosts/parasites, m is roughly $O(n^3)$. Moreover, the time complexity of AxParafit's global test is approximately $O(mi)$, which is also the time complexity of a single test of an individual association (Stamatakis et al. 2007). Since there are $A_{\neq 0}$ associations, the overall time complexity is $O(A_{\neq 0}mi)$.

It can be assumed that $A_{\neq 0}$ will be huge when using large-scale datasets. At least, it will be greater than or equal to n. For this reason, we decided to parallelize the computation of individual tests, which depicts the most time-consuming part of the analysis (Stamatakis et al. 2007). The ParaFit test is an "embarassingly parallel" problem (Stockinger et al. 2009), since calculation of individual tests can be done independently without any communication between concurrent worker tasks. Accordingly, parallelization was accomplished by using a straight-forward master-worker scheme, based on the MPI API (Message Passing Interface, see Gropp et al. 1999).

Ideally, up to $A_{\neq 0}$ single worker tasks can be distributed and simultaneously executed. Together with the fine-grained parallelization of the ACML and MKL libraries, a maximum benefit can be achieved in a Cluster environment consisting of multicore machines.

4.2.4 Grid-enabled CopyCat and AxParafit

Grid Computing

Current advances in physics and life sciences increasingly require computational resources that exceed the capacity of single workstations. For a certain time, this could be handled in an inexpensive way, at least compared to maintaining supercomputers, by assembling Clusters of standard PCs or blade systems. Certainly, this approach remains limited to the resources of a single facility. However, growing demands for computing power

led to the development of several Grid frameworks, which allow to utilize several Clusters from one or more facilities as one large decentralized system. The term "Grid Computing" was coined in analogy to the power grid, which provides easy access to electric power for everyone (Foster et al. 2001). Correspondingly, the ambitious goals of Grid Computing are to provide computing power as easily accessible as the electric power grid. But a mutual influence of both technical areas should be avoided, since no one wants to plug in his coffee machine to the power grid anymore when Mr. General Protection Fault may step out.

Foster et al. (2001) defined the essence of Grid Computing as "coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations". Thus, there are two key components: a software part that enables resource sharing, as well as an umbrella association that bundles the participating institutions.

The software that enables resource sharing and management between the participants of the Grid infrastructure is commonly called "middleware", indicating that it is located between user applications and the operating system. A middleware provides interoperability between different local infrastructures, hardware, and operating systems. It comprises services for resource management, data transfer and user-authentification between different institutions. Several sophisticated middlewares exist to date, the best-known are UNICORE (Forschungszentrum Jülich, see Streit 2009), LCG/gLite (Laure et al. 2006), and the Globus Toolkit (Foster 2005).

Naturally, virtual organizations in the academic field are mainly financed by public funds. In Europe, two such projects are of special interest that are well supported with regard to providing a computing infrastructure for the Life Sciences, Physics (e.g., the CERN Large Hadron Collider project, being the most famous project at this time), and other scientific areas. These are the EGEE ("Enabling Grids for E-sciencE") project, funded by the European Commission, and the german D-Grid initiative harboring several subprojects like the bwGRiD, which, amongst others, operates a Cluster at the University of Tübingen together with the local "Zentrum für Datenverarbeitung".

Let's plug it in

From the user's and software developer's viewpoint, there are some fundamental differences between Clusters and Grids. A single Cluster normally has a shared file system and homogenous hardware infrastructure. Thus, starting an application in a Cluster environment normally consists of collecting and uploading data and applications to a central storage, and either submitting several batch jobs, or submitting a single MPI job, which is then automatically distributed to the Cluster nodes. In contrast, a Grid may comprise many different hardware systems with different processor architec-

tures, and various decentralized storage solutions. Thus, job distribution also requires to transfer a considerable amount of data to the node the job was assigned to. Furthermore, for the middleware to make an optimal scheduling decision, information about runtime and memory requirements at submission time of the jobs is essential. Consequently, "gridifying" applications, i.e., porting them to a Grid environment may be no easy task.

We decided to port `CopyCat` and `AxParafit` to the Grid in a way that hides all these rather technical aspects from the user. Thus, the user-friendly graphical frontend of `CopyCat` remained unchanged, while the underlying engine was adapted to the needs of the Grid infrastructure (see Figure 4.5). `CopyCat` was modified to delegate invocation of `AxPcoords` and `AxParafit` to a Perl script (`AxParafit.pl`), which can easily be modified and adapted to the underlying Grid environment. Status messages of the `AxParafit.pl` script are monitored by `CopyCat` and displayed in its message window, thus keeping the user informed about the progress of the cophylogenetic analysis. After the analysis is performed on the Grid, the results are displayed by the graphical frontend and can be further analyzed with `CopyCat`. Additionally, a command line interface was integrated into `CopyCat` to allow for using the program in batch scripts to speed-up automatable analyses (for an example, see Section 3.2.1, page 63).

The `AxParafit.pl` Perl script first invokes `AxParafit` to calculate the global significance, which also allows to estimate run time and memory consumption of the individual tests. By knowing the amount of parallelizable jobs (i.e., the amount of individual tests) and their resource utilization, the Perl script then can make an appropriate decision about job granularity, i.e., into how many tasks the execution of the test has to be divided in order to limit latencies due to overhead of job submission (for details, see Stockinger et al. 2009). Depending on that, a certain amount of `AxWorker.pl` jobs is submitted to the Grid engine (see Figure 4.5). The middleware then handles transfer of the job parameters, executables and corresponding data to the executing node. `AxParafit.pl` constantly monitors the status of submitted jobs and eventually assembles the final results for `CopyCat`.

The current implementation of `AxParafit.pl` and `AxWorker.pl` was developed for the gLite middleware (Stockinger et al. 2009). But due to the modular design based on these easily modifiable Perl scripts, porting to other middleware systems can be accomplished quickly. Furthermore, adjustments to different usage scenarios will also be possible, like the development of a Web interface for job submission.

4.3 Results

Performance of the original `ParaFit` and `DistPCoA` executables was compared with the accelerated versions of `AxParafit` and `AxPcoords` using a

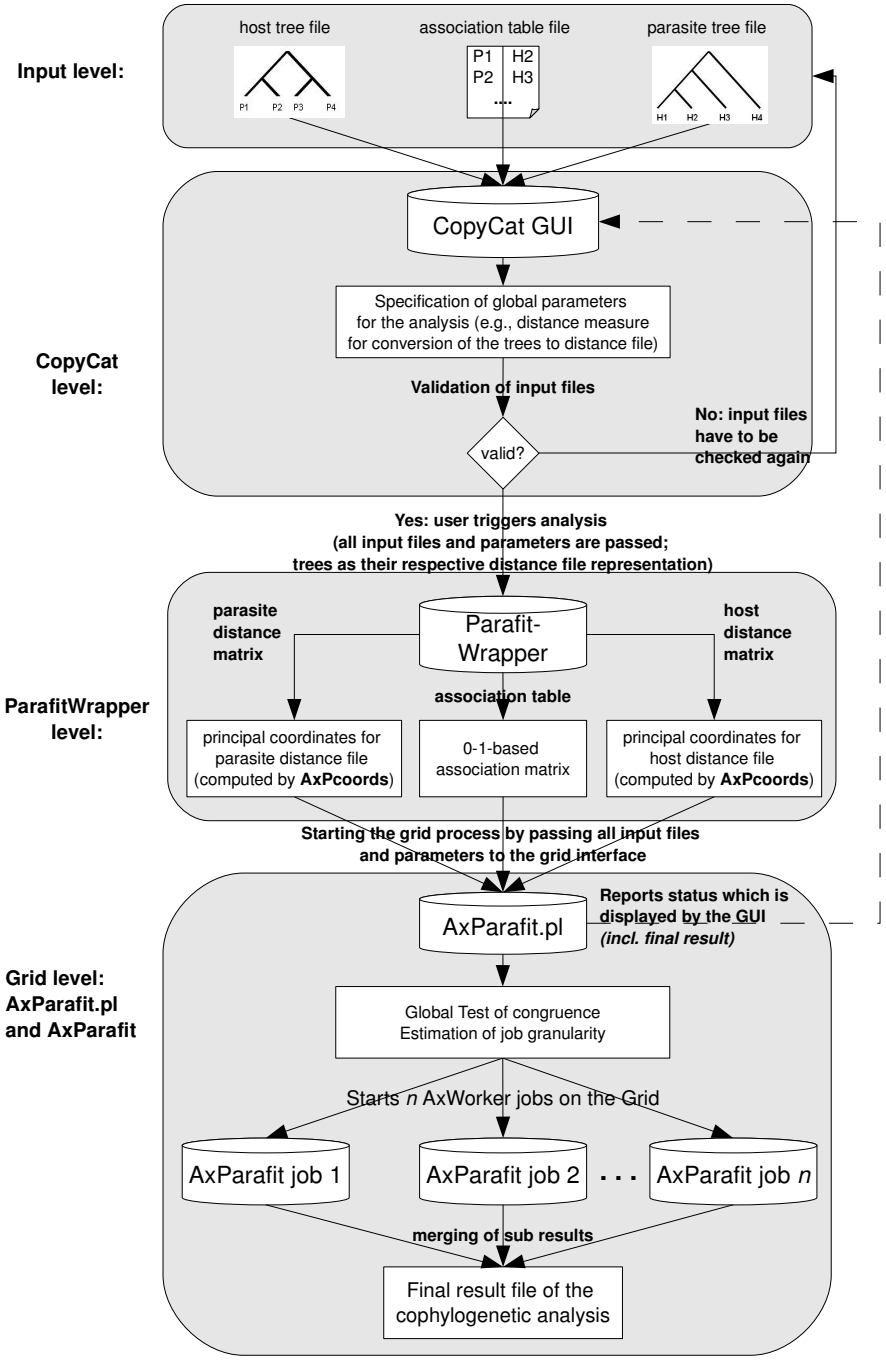


Figure 4.5: Work- and dataflow of a `CopyCat` analysis on the Grid (Stockinger et al. 2009)

system of 36 4-way AMD 2.4 GHz Opteron processors, each node equipped with 8 GB RAM. All executables were compiled with the GNU compiler suite (gcc and g77) having all optimizations enabled (for details, see Stamatakis et al. 2007). Host-parasite association data was derived from a large empirical dataset with more than 30,000 associations that was derived from the EMBL Database using the method described in Meier-Kolthoff (2006).

Figure 4.6 shows the runtime improvement of the sequential version of `AxParafit` in comparison to `ParaFit`. Interestingly, the speedup increases with growing matrix size up to a factor of 61.86, which may be attributed to highly efficient cache-utilization strategies of the BLAS implementation (Stamatakis et al. 2007).

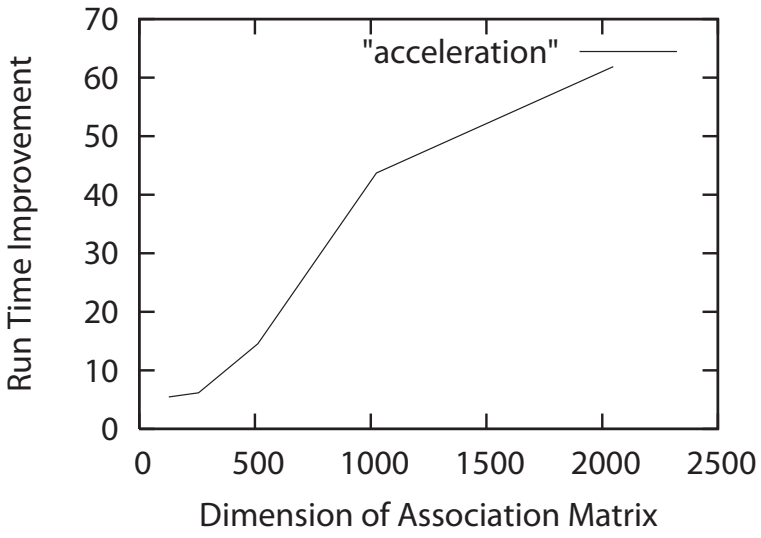


Figure 4.6: Runtime improvement of the sequential version of `AxParafit` over the original `ParaFit` (Stamatakis et al. 2007). Quadratic association matrices of dimensions 128, 256, ... up to 2048 were used.

In Figure 4.7, corresponding results of `AxPcoords` are shown. In comparison to `DistPCoA`, a performance gain up to a factor of 25.74 could be measured. Due to runtime considerations and numerical instability, which was observed using large matrices with `DistPCoA`, the acceleration rate could only be measured up to a matrix size of $4,096$.

Scalability of the MPI-enabled version of `AxParafit` was measured using a Cluster environment with up to 128 nodes. Figure 4.8 shows the speedup depending on the number of allocated CPUs. Due to the embarassingly parallel structure of the `ParaFit` test (Stockinger et al. 2009), the speedup is near-optimal (i.e., linear speedup).

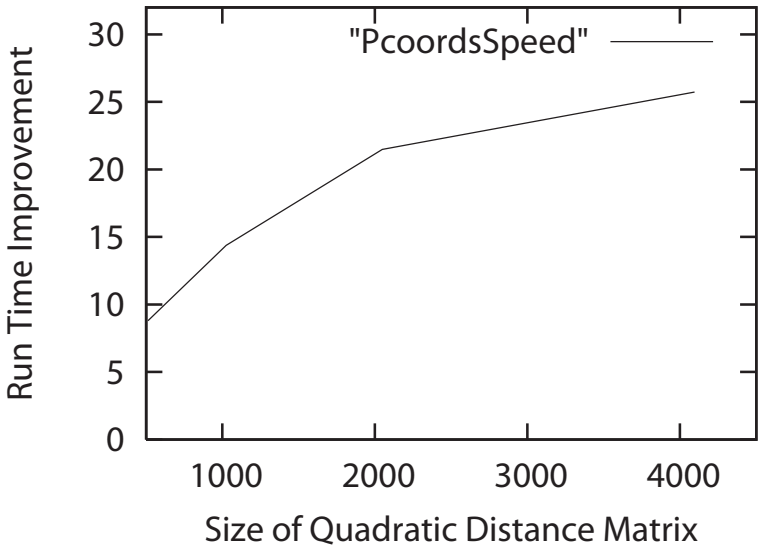


Figure 4.7: Runtime improvement of `AxPcoords` over `DistPCoA` (Stamatakis et al. 2007). Quadratic association matrices of dimensions 128, 256, ... up to 2048 were used.

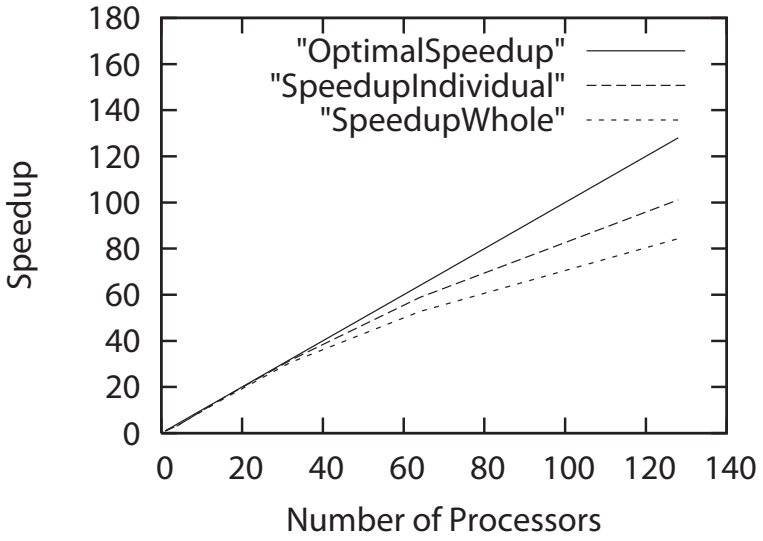


Figure 4.8: Runtime improvement of the parallelized version of `AxParafit` over the sequential version (Stamatakis et al. 2007). "SpeedupIndividual" refers to the speedup of the parallelized part only, whereas "SpeedupWhole" also includes runtime of the sequential part. A quadratic association matrix of size 512 was used on 4, 8, ... up to 128 CPUs.

4.4 Conclusions

So far, the study of deep cophylogenetic relationships was hampered by the absence of a user-friendly and highly efficient toolkit, which ideally, also can easily be used in a Cluster environment. We approached this problem by providing the easy to use graphical frontend `CopyCat` to the scientific community, based on the `ParaFit` statistical test for cophylogeny. `ParaFit`'s statistical properties make this test well-suited for the study of large-scale datasets, while `CopyCat` enables the user to infer host and parasite trees based on taxonomical data without having to collect marker sequences for phylogenetic tree reconstruction.

Additionally, we provided a highly optimized and efficient implementation of the `ParaFit` test showing speedups up to 60, as well as a parallelized implementation based on the MPI API for usage in a Cluster environment. Furthermore, we ported our toolkit to a Grid environment based on the gLite middleware.

Integration of Grid resources into the `CopyCat` graphical frontend depicts a major improvement for non-expert users, who are now able to access Grid resources in a transparent way. The Grid-enabled version of `CopyCat` and `AxParafit` was installed and successfully deployed on the Vital-IT Cluster of the Swiss Institute of Bioinformatics (Stockinger et al. 2009).

To further improve accessibility of the `CopyCat`/`AxParafit` toolkit, we intend to provide a freely accessible web-frontend to these services, including an interface to the `RAxML` web servers (Stamatakis et al. 2008) to allow the user to directly infer host or parasite phylogenies from Multiple Sequence Alignments. Additionally, a CUDA (Compute Unified Device Architecture, see Halfhill 2008; Patterson and Hennessy 2009, Appendix A) port of `AxParafit` is currently developed by co-operation partners.

Chapter 5

Metagenomics

5.1 Introduction

Assessing biodiversity on Earth is a main goal of today's life sciences. While there is a great variety of eukaryotic species (even when focusing on metazoa, consider for example nature's preference for some groups like beetles, which already kept Darwin and Wallace occupied, see Berry 2008), the most dominant group of life-forms certainly are the prokaryotes (Gregory 2005).

But estimating species richness of prokaryotes is severely hampered because many of those (up to 99%) cannot be cultured, and therefore can neither be classified, nor sequenced using traditional sequencing methods. However, recent advances in next-generation sequencing technologies helped cutting the Microbiologists' Gordian knot by providing methods to isolate DNA from an environment and to directly sequence the whole sample in one go. This new methodology was labelled "Metagenomics", denoting "the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms" (Handelsman 2004).

Due to restrictions of conventional sequence technologies, sequencing of single organisms as well as whole communities results in a large amount of short fragments of DNA (called "reads"). Recent sequencing technologies produce several millions of reads in one single run, whereas typical read lengths range from 35 bp to more than 900 bp (Pop and Salzberg 2008). Consequently, there is a great need for computational tools that are able to handle large metagenomic and paleogenomic (i.e., the study of prehistoric genomes, see Birnbaum et al. 2000; Poinar et al. 2006) datasets.

5.2 Methods

5.2.1 Taxonomic binning using MEGAN

An important step of a metagenomic analysis consists of mapping reads obtained from the sequencing process to the corresponding species, thus allowing to assess species richness and abundancy of the metagenomic sample. MEGAN was specifically developed to map large datasets of metagenomic data onto a taxonomic tree using homology search tools. The first step in the taxonomical binning process (see Figure 5.1) consists of using a homology search program like BLAST (Altschul et al. 1990; 1997) or BLASTZ (Schwartz et al. 2003) to look for hits in large sequence databases such as NCBI-NR and NCBI-NT (Wheeler et al. 2008). After blasting the reads against a database, the resulting HSPs (High Scoring Segment Pairs) are mapped to the corresponding taxon (represented by its taxon-ID, see Benson et al. 2008; NCBI 2009c) based on the description line of the database hit sequence. Several filters can be used to reduce the amount of hits of a read sequence to obtain a candidate list for taxonomic binning.

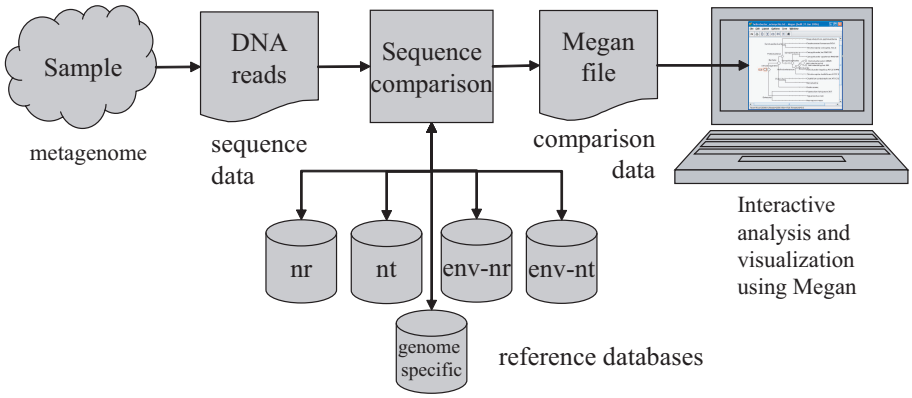


Figure 5.1: Typical workflow of a MEGAN analysis (Huson et al. 2007a).

MEGAN includes a taxonomic tree based on the NCBI taxonomy (Wheeler et al. 2008), which is used to visualize the species distribution of the sample (see Figure 5.2). The taxonomic tree is derived by converting the raw taxonomy data from NCBI (2009c) into a tree representation. In doing so, the hierarchical information is retained, so that tree nodes can be collapsed or expanded in regard to their clade (i.e., superkingdom, phylum, class, order, and family).

When blasting reads against a large database, most reads hit more than one single database sequence. While this poses no problem if all those hits can be mapped to the same taxon, a conflict resolving strategy has to be applied in the other case. Figure 5.3 demonstrates the application of the Lowest Common Ancestor (LCA) assignment algorithm of MEGAN. In this ex-

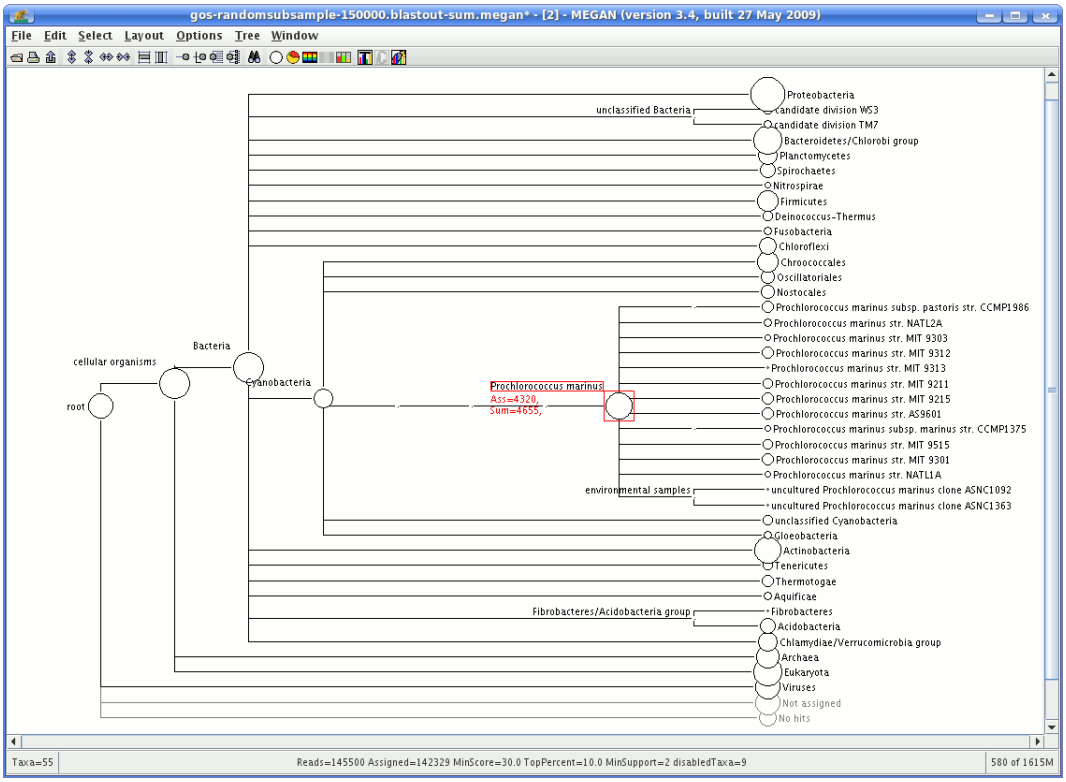


Figure 5.2: Screenshot of the `MEGAN` main window with the taxonomy tree browser. An analysis of a random subsample of the Global Ocean Survey dataset (Rusch et al. 2007) is shown.

ample, three hits could be obtained for the given read, which were mapped to *Campylobacter lari*, *Helicobacter hepaticus*, and *Wolinella succinogenes*. Consequently, the read is assigned to the node depicting the lowest common ancestor of those species in the taxonomy tree, i.e., the common ancestor that is nearest to the leaves (here, the node of the order Campylobacterales). This strategy ensures that the specificity of a hit can directly be observed based on the taxonomic level of its assignment. Accordingly, highly conserved sequences are assigned to high-ranking taxa close to the root, whereas species-specific sequences are placed near to the leaves.

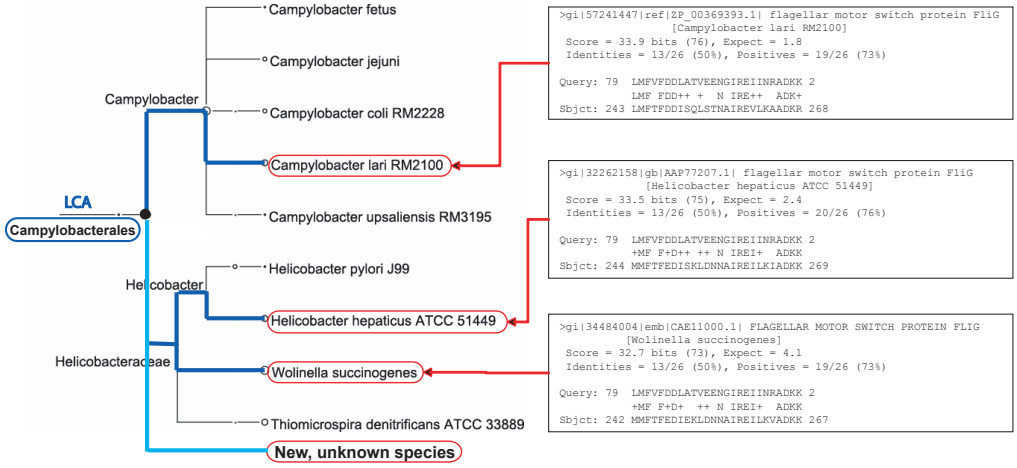


Figure 5.3: LCA assignment algorithm (Huson et al. 2007a).

Species abundancies as measured by the amount of reads mapped to a certain taxon are displayed by the corresponding node sizes. Thus, `MEGAN` allows to visualize the species distribution of a sample in a comprehensible and descriptive way (see Figure 5.2).

`MEGAN` was successfully applied in several metagenomics studies (e.g., Poinar et al. 2006; Urich et al. 2008; Claesson et al. 2009; Huson et al. 2009; Qi et al. 2009; Woyke et al. 2009).

5.2.2 The FragmentAssigner pipeline

Sequence comparison represents the most time-consuming part of a `MEGAN` analysis. Thus, the automation and parallelization of this step is an important factor for each metagenomics project, especially when considering that the amount of sequence data is continously growing due to ongoing improvements in sequencing technology (Mardis 2008).

To analyze metagenomics datasets in a user-friendly way, we developed the `FragAssigner` tool as an in-house solution that provides an easy inter-

face to different local alignment search tools, as well as to different sequence databases.

Basic concept

In the current version, the `FragAssigner` tool provides a command line interface. Databases and local alignment tools can be configured using a configuration file with `keyword=value` pairs (see Figure 5.5). Command line arguments of the different alignment tools are made transparent to the user, so that settings can be kept consistent over a wide range of tools. Such settings are: using low complexity filter and soft masking, seed word size, e-Value threshold and maximum number of HSPs per query sequence (for an explanation, see Korf et al. 2003).

`FragAssigner` supports `BLAST` (Altschul et al. 1990; 1997) in the implementation from Washington University, as well as in the implementation from NCBI. Furthermore, `BLASTZ` (Schwartz et al. 2003) and `BLAT` (Kent 2002) can also be used.

Figure 5.4 shows the layout of the `FragAssigner` pipeline. The tool is invoked by providing a Multi-FASTA file with the reads to be assigned. `FragAssigner` then starts several database queries in parallel (on a multiprocessor system), by using large databases like NCBI NR and NT (Benson et al. 2008). In addition, the archive of completely sequenced prokaryotic genomes from NCBI (2009a) can be included in the local alignment search. Hence, the whole range of genomic DNA can be used to search for homologies.

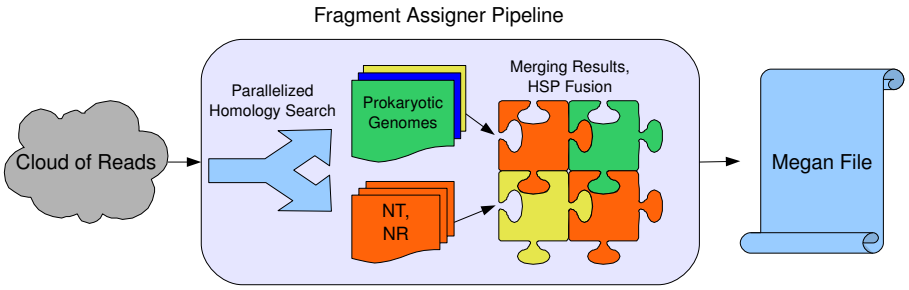


Figure 5.4: Graphical visualization of the `FragAssigner` Pipeline.

The `MEGAN BLAST` import filter assign HSPs to a specific taxon-ID by parsing its database hit description lines. `FragAssigner` however uses a different approach by directly utilizing the taxon-ID information contained in the compiled NR and NT database files. Correspondingly, wrong assignments are avoided, at least if it can be assumed that the assignments provided by NCBI are accurate (an example of such a wrong assignment is discussed in Huson et al. 2007a). To obtain the taxon-ID information, the

tool `fastacmd` distributed with the NCBI BLAST package is used to compile a binary file containing a map between sequence accession IDs (Benson et al. 2008) and the corresponding taxon-ID. This time-consuming process only has to be repeated when a new version of the NT or NR database is installed. In addition, the archive of completely sequenced prokaryotic genomes is automatically mapped to the corresponding taxon-ID based on the species labels.

The `FragAssigner` pipeline was successfully used by Ramona Schmid to evaluate `ReadSim` (Schmid 2006), the predecessor of `MetaSim` (Richter et al. 2008), as well as by Nikita Meyer in his student research project dealing with the comparison of several metagenomics datasets (Meyer 2007).

Improving the specificity of contig assignments

`MEGAN` assigns reads to taxon nodes by filtering the corresponding HSPs, determining the associated species' taxon-IDs, and seeking the lowest common ancestor node of those species (see Figure 5.3). Consequently, a high amount of HSPs mapped to a read may lead to unspecific assignments, i.e., assignments that are close to the root of the taxonomy. On the one hand, this may be due to the fact that the read contains a highly conserved sequence (e.g., a protein domain) shared between a broad range of organisms. On the other hand, when using sequence technologies that yield large read lengths or by assembling reads into contigs (contiguous sequences), the likelihood increases that a sequence (read or contig) may contain more than one open reading frame. Thus, with increasing sequence length, specific assignments become harder to obtain.

In Huson et al. (2007a), we studied the specificity of the LCA assignment process depending on different read lengths from 35 bp to 800 bp. It could be demonstrated that the rate of assigned reads considerably increases with read length, while most reads can be specifically assigned close to the originating species. But subsequent analyses with larger reads close to the length of contigs showed that the assignments move towards more unspecific nodes with growing sequence length (data not shown).

There exist several ways to approach this problem. One possibility may be to refine the LCA assignment process by using improved filtering strategies together with a majority rule for determining the lowest common ancestor. A different approach consists in addressing this problem within the `FragAssigner` pipeline by grouping HSPs together that belong to the same subject sequence. These HSPs can be merged into a large HSP (see Figure 5.6). On the one hand, this reduces the amount of HSPs per read, while on the other hand, the new HSP may be preferred by `MEGAN`'s filtering process over other HSPs that have a lower bitscore.

After the merging of several HSPs into one large HSP, a new sum bitscore and e-Value can be calculated based on Korf et al. (2003, p. 103). Let r

```
# FragAssigner Inifile

localAlignmentTool=NCBIBlastWrapper

NCBIBlastWrapper.appPath=blastall
NCBIBlastWrapper.formatDbPath=formatdb
#NCBIBlastWrapper.wordLength=7

WUBlastWrapper.appPath=wublast
#WUBlastWrapper.wordLength=7

### blast low complexity filter
### 1: filter enabled, 0: filter deactivated
#lowComplexityFilter=1
lowComplexityFilter=0

### soft masking
### default: 0
softMasking=1

### directory containing the prokaryotic genome database
genomesDir=./genomes

NTdbLocation=/path/to/nt-db
NTdbTaxInfoFile=/path/to/nt-db/compiled-taxinfo.txt

NRdbLocation=/path/to/nr-db
NRdbTaxInfoFile=/path/to/nr-db/compiled-taxinfo.txt

taxDataFile=taxondata.txt
blastOutputDir=./blastout
parallelJobCount=8

#eValueThreshold=1E-3
eValueThreshold=10

outputFileType=auch.mgp.MeganFile
#outputFileType=auch.blasthsp.CGVIZFile

### exclude groups from homology search
### groups:
###   bacteriaGenomes
###   ncbiNR
###   ncbiNT
#removeGroups=bacteriaGenomes,ncbiNR,ncbiNT
#removeGroups=ncbiNR,ncbiNT
### uses NT only:
#removeGroups=bacteriaGenomes,ncbiNR
### uses NR only:
removeGroups=bacteriaGenomes,ncbiNT
```

Figure 5.5: Example of a `FragAssigner` configuration file.

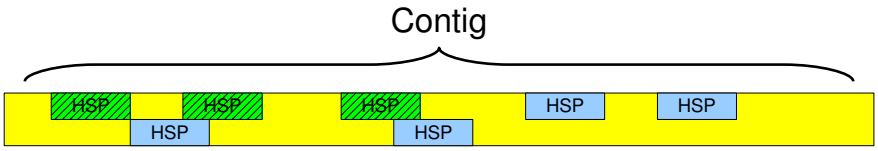


Figure 5.6: A contig (contiguous sequence, may comprise 1000 to more than 100,000 bp) with several HSPs. The green, shaded HSPs belong to the same subject sequence and could be successfully combined into one single HSP with a bitscore that is higher than any of the single HSPs.

denote the number of HSPs to be merged, $s_{\text{bit}}(i)$ the bitscore of the ith HSP, κ the adjustment constant of the Karlin-Altschul statistics (Karlin and Altschul 1990; Korf et al. 2003, p. 65), and m' as well as n' the effective query and database lengths (see Korf et al. 2003, p. 67). Further, let g be the number of gaps between the HSPs. Then, the sum bitscore $S_{\text{sumbit}}(r)$ can be obtained by using the following equation:

$$S_{\text{sumbit}}(r) = \sum_{i=1}^{r} s_{\text{bit}}(i) \; + \frac{r \ln \kappa - r \ln(\kappa m' n')}{\ln 2} \tag{5.1}$$

When combining HSPs that are collinear, a smaller penalty can be used (see Formula 5.2). HSPs are called collinear when their query and subject sequences have the same direction and their intervals do not overlap.

$$S_{\text{sumbit}}(r) = \sum_{i=1}^{r} s_{\text{bit}}(i) + \frac{r \ln \kappa - \ln(\kappa m' n') - (r-1) \cdot (\ln \kappa + 2 \ln g) - \ln(r!)}{\ln 2} \tag{5.2}$$

The HSP merging algorithm works as follows: For each query (read) sequence, HSPs belonging to the same subject sequence are determined. A list of non-overlapping candidates for merging is then built using a greedy algorithm similar to the approach described in Section 2.2.1 (p. 8). Afterwards, the algorithm tries to extend each candidate by successively merging it with its neighbours and testing whether the resulting sum bitscore is larger than the maximum bitscore of the individual HSPs. If such an optimally merged HSP can be found, the original HSPs are replaced by the new combined HSP. This step is repeated until no further improvement can be achieved.

Taxonomy-based HSP fusion

Another source for ambiguous assignments especially affects reads from species that are represented by a plenitude of different strains in the se-

quence databases. In this case, the reads will be scattered over the species node and its subnodes, as shown in Figure 5.8 for *E. coli* reads. To improve `MEGAN`'s assignment results, we decided to approach this problem in `FragAssigner` using an ad-hoc strategy based on the previously discussed HSP fusion.

In contrast to the previous algorithm, candidates for HSP fusion are determined using taxonomic information. First, each HSP is internally assigned to its species node in the taxonomy, thus assignments to different strains of the same species are all remapped to the same taxon. This is made in a transparent way, so that the remapping only affects results if the HSP merging process succeeds. In the next step, all HSPs belonging to the same read and remapped taxon-ID are combined. If the bitscore of the resulting HSP (recalculated based on Formula 5.1) is greater than the maximum of the individual HSPs, the individual HSPs are replaced with the combined HSP, which is assigned to the species node.

The taxonomy-based filter can be deactivated in case that a remapping to species nodes is not desired (e.g., when subspecies/strain assignments are biologically meaningful).

5.2.3 MetaSim, a sequencing simulator for Metagenomics

Developing software, especially in the scientific area, also requires the development of test cases or datasets that allow to evaluate the software's accuracy. Namely, the quality of programs for taxonomic binning of reads can be determined by simulation studies based on the generation of a considerable amount of reads from already sequenced genomes, so that the originating species is known. This enables to evaluate the taxonomic binning software by comparing its results with the true assignments. Naturally, such a process can be done in the wet lab by sequencing organisms whose genome is already known, but this depicts a rather expensive and time-consuming procedure.

`MEGAN` was initially assessed using `ReadSim` (Huson et al. 2007a;b), a read simulation software that extracts reads from a genomic sequence and simulates sequencing errors of different sequencing technologies (Schmid 2006). However, `ReadSim` was not specifically designed for simulating large metagenomics datasets, but rather to simulate sequencing of single genomes. Hence, it did not allow for simulating different species abundancies or for the incorporation of taxonomic information.

Consequently, we decided to develop `MetaSim`, a sequencing simulator for genomics and metagenomics (Richter et al. 2008). Within `MetaSim` (see Figure 5.7), different taxon profiles can be generated and managed. Relative abundancies of selected species can be defined in those profiles, enabling the user to fine-tune the species distribution of the metagenomics sample to be simulated. The current version of `MetaSim` supports error models of Sanger

(Meldrum 2000), Roche's 454 (Margulies et al. 2005), and Illumina (Bentley 2006) sequencing technologies.

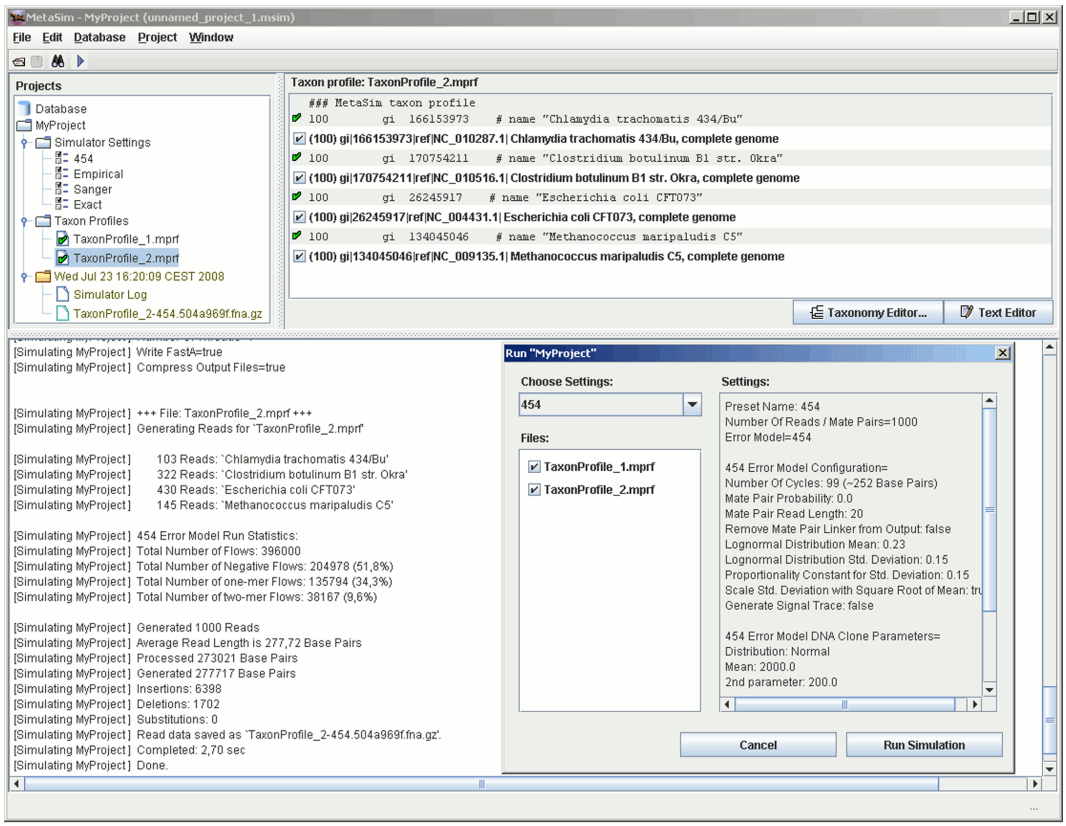


Figure 5.7: Screenshot of `MetaSim`'s main window and the dialog for starting the simulation run (Richter et al. 2008).

Additionally, `MetaSim` contains a population sampler to model the complexity of real-world datasets. The population sampler generates a population of offsprings from a given genome. For that purpose, either a user-generated phylogenetic tree can be loaded into the program, or a random tree is automatically generated by using the Yule-Harding model (Yule 1925; Harding 1971). The user can select a genome, which is then used to generate a population of offsprings. Sequence evolution is simulated along the phylogenetic tree based on the Jukes-Cantor model of DNA evolution (Jukes and Cantor 1969). The transition rate α is used together with the branch length of the phylogenetic tree to estimate a probability of change for each site. The value of α can be adjusted by the user within reasonable limits ($0 < \alpha < 1/3$).

MetaSim is freely available at http://www-ab.informatik.uni-tuebingen.-de/software/metasim. To date, MetaSim has been successfully used in several studies (Haque et al. 2009; Hoff et al. 2009; Zagordi et al. 2009).

5.3 Results

5.3.1 HSP fusion algorithm

To evaluate the performance of the HSP fusion algorithm, we simulated 1000 fragments from the *Escherichia coli* strain K12 genome using ReadSim. Fragment lengths were between 500 and 10,000 bp with a mean length of 2000 bp. The lengths were chosen to simulate typical sizes of small contigs.

Figure 5.8 shows the results based on a MEGAN analysis without utilizing the HSP fusion filter. Most reads are assigned to the family "Enterobacteriaceae". Only a minority of reads is clearly assigned to the *E. coli* taxon or its subnodes. However, applying the HSP fusion algorithm based on collinearity, as well as on taxonomic information leads to a specific assignment of the majority of reads to the *E. coli* taxon (Figure 5.9). This clearly demonstrates the effectivity of the HSP fusion algorithm.

5.3.2 MetaSim and MEGAN

To evaluate the accuracy of MEGAN's taxonomic binning approach, we generated three different species abundance profiles using three different simulated sequencing technologies, altogether resulting in nine different datasets. Species abundance profiles were modeled in analogy to Mavromatis et al. (2007), corresponding to low, medium, and high complexity communities.

Approximately 15 Mbp of reads were generated for each dataset. Afterwards, the reads were blasted against the NR database, and were assigned to taxa using MEGAN.

Results indicate that the amount of assignable reads increases with read length, together with the amount of true positives. But even for small read sizes, the amount of false positives remained low ($< 2\%$), which corroborates the observations made in Huson et al. (2007a). Overall, MEGAN shows high accuracy in assigning reads to corresponding taxa. A thorough analysis of the simulation results can be found in Richter et al. (2008).

5.4 Conclusions

Assigning reads to taxonomic groups (taxonomic binning) represents an important step of each metagenomic analysis (Raes et al. 2007; Kunin et al. 2008). Therefore, we developed MEGAN as a user-friendly graphical application for the visualization of species diversity in metagenomics datasets.

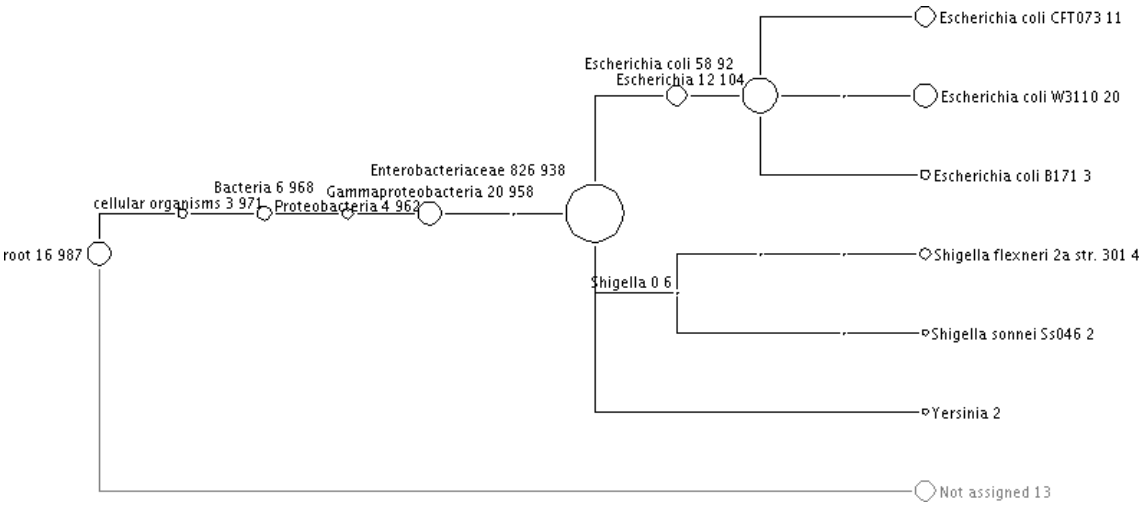


Figure 5.8: Test of assignment accuracy using contigs with `MEGAN`. Here, 1000 reads were generated from *E. coli* strain K12 using `ReadSim` with a read length between 500 and 10,000 bp (mean length: 2000 bp) to simulate typical lengths of small contigs. Most reads were assigned to Enterobacteriaceae.

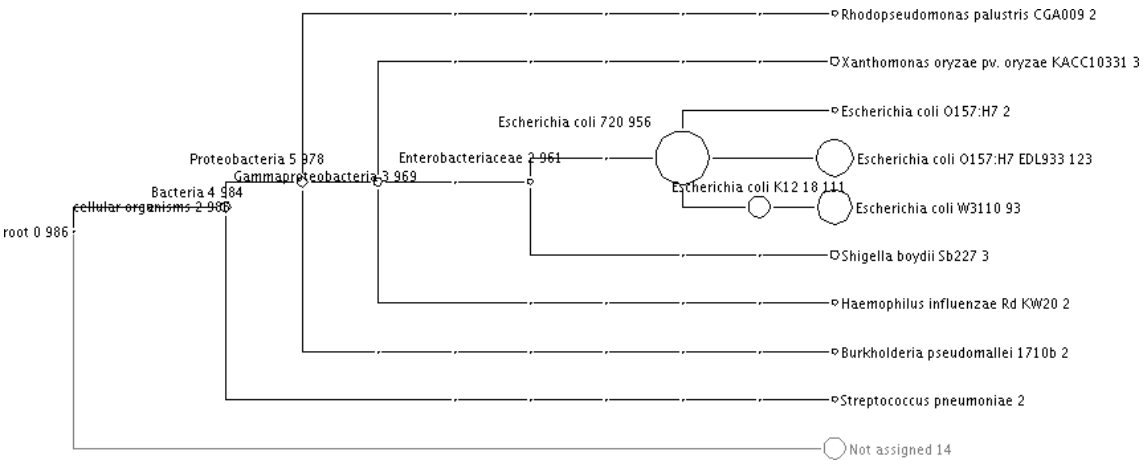


Figure 5.9: Test of assignment accuracy using contigs with `MEGAN`. The same dataset was used as for Figure 5.8. In addition, the HSP fusion strategy of `FragAssigner` was used to pre-filter HSPs (collinearity as well as taxonomy based). Accordingly, most reads could be assigned to the *E. coli* species node, thus showing higher specificity.

MEGAN scales well even on standard hardware and is able to efficiently handle datasets comprising up to several hundred gigabytes of BLAST output.

Additionally, we developed MetaSim, a sequencing simulator specifically designed for the simulation of metagenomics datasets. By including a population sampler, the complexity of real world metagenomic datasets can be reproduced more closely. While MetaSim can be used to test any software that processes reads of a metagenomics sample (like assemblers, taxonomic binning software, etc.), we explicitly tested the accuracy of MEGAN, using nine simulated datasets of different complexity level and sequencing technology. The results clearly indicate that MEGAN has a low rate of false positives, while the amount of assignable reads increases with read length. Overall, specificity of the assignments made by MEGAN's LCA algorithm is high (Huson et al. 2007a; Richter et al. 2008). Anyhow, an improvement of the assignment of contigs may be achieved by applying the HSP fusion algorithm as outlined in Section 5.2.2.

Chapter 6

Conclusions and Outlook

In this thesis, several topics were discussed that are all related to phylogenetics. In Chapter 2, the `GBDP` framework was presented fo inferring whole genome phylogenies based on local alignment search tools (Henz et al. 2005; Auch et al. 2006b;a; 2009a;b). The framework was applied to datasets of prokaryotic genomes, as well as organelle (mitochondria and plastids) genomes of the major eukaryotic groups of plants, fungi and animals. The most interesting observation is that the obtained whole genome phylogenies are largely, but not completely, in agreement with the reference taxonomy provided by NCBI. Since the NCBI taxonomy is primarily based on the phylogeny of single markers like 16S rRNA, one may conclude that vertical inheritance clearly plays the dominant role for most species. We also analyzed a common set of 17 prokaryotic genes and screened them for the occurrence of HGT (horizontal gene transfer, see Chapter 3). Despite the high amount of potential HGT, a noticeable congruence could be observed between the individual gene trees and the corresponding supermatrix tree. Furthermore, the individual trees and the supermatrix tree mostly were in accordance with the NCBI taxonomy.

Thus, there seems to exist an unobscured vertical inheritance signal both in the "averaged" phylogenies based on whole genomes, as well as in the set of individual common genes. For this reason, we conclude that the majority of the genome follows the phylogeny of marker genes (mainly translational genes), with the exception of some groups of genes that are more likely to undergo horizontal gene transfer (e.g., metabolic genes, see Kanhere and Vingron 2009). Bringing all this together, we take sides with the camp of the positivists (the expression was coined by Dagan and Martin 2006), stating that there is a microbial tree of life, at least for the majority of taxa (and genes). This may seem to be impudent, but it is supported by recent findings of other groups (House 2009; Puigbò et al. 2009; Swithers et al. 2009). However, our assumption does not deny the possibility of a era dominated by horizontal inheritance at the beginning of the history

of life on our planet (the progenote era, see Woese and Fox 1977; Woese 2002). Certainly, there also exist conflicting views denying the dominance of a reasonable tree-like evolutionary structure in prokaryotes (Dagan and Martin 2006; Doolittle and Bapteste 2007; Bapteste and Boucher 2009). However, the jury is still out on this question, and it may be more realistic to use network-based phylogenetic methods to allow for uncertainty and incongruence between gene phylogenies (Huson and Bryant 2006).

Moreover, the employment of `GBDP` to species delineation yields promising results. Two corresponding publications describing the usage of `GBDP` in this context (Auch et al. 2009a;b) will be published in the open access journal of the "Genomics Standards Consortium" ("Standards in Genomic Sciences").

In Chapter 4 and 5, we presented user-friendly software packages for biologists. With `CopyCat` and the Grid-enabled versions of `AxParafit` and `AxPcoords`, large-scale cophylogenetic analyses have now become feasible (Stamatakis et al. 2007; Stockinger et al. 2009). Consequently, an empirical verification of the Fahrenholz rule can be approached in the future (see also Begerow et al. 2004; Refrégier et al. 2008; Garamszegi 2009). The Fahrenholz rule states that cospeciation may be the predominant factor in host-parasite evolution (Eichler 1948). We are currently preparing a large-scale dataset comprising several tenths of thousands of host and parasite taxa, which hopefully allows us to shed light on that question. In this context, we look forward to the ongoing CUDA (Halfhill 2008) port of `AxParafit`, which will allow us to use large GPU farms for that purpose.

We assisted the analysis of large metagenomics datasets by providing `MEGAN` to the scientific community. `MEGAN` is a user-friendly software for the efficient analysis and taxonomic classification of metagenomic samples. It was successfully applied in several metagenomic studies (e.g., Poinar et al. 2006; Urich et al. 2008; Claesson et al. 2009; Huson et al. 2009; Qi et al. 2009; Woyke et al. 2009). Furthermore, we developed `MetaSim`, a sequencing simulator for metagenomics. With `MetaSim`, current and future software for metagenomic and genomic analysis can be evaluated. Consequently, we used `MetaSim` to successfully evaluate the performance of `MEGAN` in taxonomic binning (Richter et al. 2008).

Overall, when looking on the growing field of computational biology, one receives the impression of an endless journey to the horizon of knowledge and perception. But it is hard to decide whether this will be more like Alice's Adventures in Wonderland (Carroll 1865), or rather like a Sisyphean challenge (Köhlmeier 2002).

Appendix A

Publications

A.1 Peer-reviewed papers

1. Hendrik N. Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D. E. MacPhee, Bernard Buigues, Alexei Tikhonov, Daniel Huson, Lynn P. Tomsho, Alexander Auch, Markus Rampp, Webb Miller, Stephan C. Schuster. **Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA.** *Science* 311:392-394, 2006.

   We sequenced 28 million base pairs of DNA in a metagenomics approach, using a woolly mammoth (Mammuthus primigenius) sample from Siberia. As a result of exceptional sample preservation and the use of a recently developed emulsion polymerase chain reaction and pyrosequencing technique, 13 million base pairs (45.4%) of the sequencing reads were identified as mammoth DNA. Sequence identity between our data and African elephant (Loxodonta africana) was 98.55%, consistent with a paleontologically based divergence date of 5 to 6 million years. The sample includes a surprisingly small diversity of environmental DNAs. The high percentage of endogenous DNA recoverable from this single mammoth would allow for completion of its genome, unleashing the field of paleogenomics.

2. Alexander F. Auch, Stefan R. Henz, Barbara R. Holland, Markus Göker. **Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences.** *BMC Bioinformatics* 7:350, 2006.

   Background: Phylogenetic methods which do not rely on multiple sequence alignments are important tools in inferring trees directly from completely sequenced genomes. Here,

we extend the recently described Genome BLAST Distance Phylogeny (GBDP) strategy to compute phylogenetic trees from all completely sequenced plastid genomes currently available and from a selection of mitochondrial genomes representing the major eukaryotic lineages.

BLASTN, TBLASTX, or combinations of both are used to locate high-scoring segment pairs (HSPs) between two sequences from which pairwise similarities and distances are computed in different ways resulting in a total of 96 GBDP variants. The suitability of these distance formulae for phylogeny reconstruction is directly estimated by computing a recently described measure of "treelikeness", the so-called δ value, from the respective distance matrices. Additionally, we compare the trees inferred from these matrices using UPGMA, NJ, BIONJ, FastME, or STC, respectively, with the NCBI taxonomy tree of the taxa under study.

Results: Our results indicate that, at this taxonomic level, plastid genomes are much more valuable for inferring phylogenies than are mitochondrial genomes, and that distances based on breakpoints are of little use. Distances based on the proportion of "matched" HSP length to average genome length were best for tree estimation. Additionally we found that using TBLASTX instead of BLASTN and, particularly, combining TBLASTX and BLASTN leads to a small but significant increase in accuracy. Other factors do not significantly affect the phylogenetic outcome. The BIONJ algorithm results in phylogenies most in accordance with the current NCBI taxonomy, with NJ and FastME performing insignificantly worse, and STC performing as well if applied to high quality distance matrices. δ values are found to be a reliable predictor of phylogenetic accuracy.

Conclusion: Using the most treelike distance matrices, as judged by their δ values, distance methods are able to recover all major plant lineages, and are more in accordance with Apicomplexa organelles being derived from "green" plastids than from plastids of the "red" type. GBDP-like methods can be used to reliably infer phylogenies from different kinds of genomic data. A framework is established to further develop and improve such methods. δ values are a topology-independent tool of general use for the development and assessment of distance methods for phylogenetic inference.

3. Daniel H. Huson, Alexander F. Auch, Ji Qi, Stephan C. Schuster. **MEGAN Analysis of Metagenomic Data.** *Genome Research* 17:377-386, 2007.

   Metagenomics is the study of the genomic content of a sample of organisms obtained from a common habitat using targeted or random sequencing. Goals include understanding the extent and role of microbial diversity. The taxonomical content of such a sample is usually estimated by comparison against sequence databases of known sequences. Most published studies use the analysis of paired-end reads, complete sequences of environmental fosmid and BAC clones, or environmental assemblies. Emerging sequencing-by-synthesis technologies with very high throughput are paving the way to low-cost random shotgun approaches. This paper introduces MEGAN, a new computer program that allows laptop analysis of large metagenomic data sets. In a preprocessing step, the set of DNA sequences is compared against databases of known sequences using BLAST or another comparison tool. MEGAN is then used to compute and explore the taxonomical content of the data set, employing the NCBI taxonomy to summarize and order the results. A simple lowest common ancestor algorithm assigns reads to taxa such that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence. The software allows large data sets to be dissected without the need for assembly or the targeting of specific phylogenetic markers. It provides graphical and statistical output for comparing different data sets. The approach is applied to several data sets, including the Sargasso Sea data set, a recently published metagenomic data set sampled from a mammoth bone, and several complete microbial genomes. Also, simulations that evaluate the performance of the approach for different read lengths are presented.

4. Jan P. Meier-Kolthoff, Alexander F. Auch, Daniel H. Huson, Markus Göker. **COPYCAT: Co-phylogenetic Analysis tool.** *Bioinformatics* 23(7):898-900, 2007.

   We have developed the software COPYCAT which provides an easy and fast access to cophylogenetic analyses. It incorporates a wrapper for the program PARAFIT, which conducts a statistical test for the presence of congruence between host and parasite phylogenies. COPYCAT offers

various features, such as the creation of customized hostparasite association data and the computation of phylogenetic host/parasite trees based on the NCBI taxonomy.

5. Daniel H. Huson, Alexander F. Auch, Ji Qi, Stephan C. Schuster. **Metagenome analysis using MEGAN.** In *Proceedings of the 5th Asia-Pacific Bioinformatics Conference*, Volume 5 of Series on Advances in Bioinformatics and Computational Biology. Edited by D. Sankoff, L. Wang and F. Chin, Imperial College Press 2007:7-16.

   In metagenomics, the goal is to analyze the genomic content of a sample of organisms collected from a common habitat. One approach is to apply large-scale random shotgun sequencing techniques to obtain a collection of DNA reads from the sample. This data is then compared against databases of known sequences such as NCBI-nr or NCBI-nt, in an attempt to identify the taxonomical content of the sample. We introduce a new software called MEGAN (Meta Genome ANalyzer) that generates species profiles from such sequencing data by assigning reads to taxa of the NCBI taxonomy using a straight-forward assignment algorithm. The approach is illustrated by application to a number of datasets obtained using both sequencing-by-synthesis and Sanger sequencing technology, including metagenomic data from a mammoth bone, a portion of the Sargasso sea data set, and several complete microbial test genomes used for validation proposes.

6. Alexandros Stamatakis, Alexander F. Auch, Jan P. Meier-Kolthoff, Markus Göker. **AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa.** *BMC Bioinformatics* 8:405, 2007.

   Background: Current tools for Co-phylogenetic analyses are not able to cope with the continuous accumulation of phylogenetic data. The sophisticated statistical test for host-parasite co-phylogenetic analyses implemented in Parafit does not allow it to handle large datasets in reasonable times. The Parafit and DistPCoA programs are the by far most compute-intensive components of the Parafit analysis pipeline. We present AxParafit and AxPcoords (Ax stands for Accelerated) which are highly optimized versions of Parafit and DistPCoA respectively.
   Results: Both programs have been entirely re-written in C. Via optimization of the algorithm and the C code as well

as integration of highly tuned BLAS and LAPACK methods AxParafit runs 561 times faster than Parafit with a lower memory footprint (up to 35% reduction) while the performance benefit increases with growing dataset size. The MPI-based parallel implementation of AxParafit shows good scalability on up to 128 processors, even on medium-sized datasets. The parallel analysis with AxParafit on 128 CPUs for a medium-sized dataset with an 512 by 512 association matrix is more than 1,200/128 times faster per processor than the sequential Parafit run. AxPcoords is 826 times faster than DistPCoA and numerically stable on large datasets. We outline the substantial benefits of using parallel AxParafit by example of a large-scale empirical study on smut fungi and their host plants. To the best of our knowledge, this study represents the largest co-phylogenetic analysis to date.

Conclusion: The highly efficient AxPcoords and AxParafit programs allow for large-scale co-phylogenetic analyses on several thousands of taxa for the first time. In addition, AxParafit and AxPcoords have been integrated into the easy-to-use CopyCat tool.

7. Daniel C. Richter, Felix Ott, Alexander F. Auch, Ramona Schmid, Daniel H. Huson. **MetaSim - A Sequencing Simulator for Genomics and Metagenomics.** *PLoS ONE* 3(10): e3373 doi:10.1371/journal.pone.0003373, 2008.

Background: The new research field of metagenomics is providing exciting insights into various, previously unclassified ecological systems. Next-generation sequencing technologies are producing a rapid increase of environmental data in public databases. There is great need for specialized software solutions and statistical methods for dealing with complex metagenome data sets.

Methodology/Principal Findings: To facilitate the development and improvement of metagenomic tools and the planning of metagenomic projects, we introduce a sequencing simulator called MetaSim. Our software can be used to generate collections of synthetic reads that reflect the diverse taxonomical composition of typical metagenome data sets. Based on a database of given genomes, the program allows the user to design a metagenome by specifying the number of genomes present at different levels of the NCBI taxonomy, and then to collect reads from the metagenome using a sim-

ulation of a number of different sequencing technologies. A population sampler optionally produces evolved sequences based on source genomes and a given evolutionary tree.

Conclusions/Significance: MetaSim allows the user to simulate individual read datasets that can be used as standardized test scenarios for planning sequencing projects or for benchmarking metagenomic software.

8. Daniel H. Huson, Daniel C. Richter, Suparna Mitra, Alexander F. Auch, Stephan C. Schuster. **Methods for Comparative Metagenomics.** *BMC Bioinformatics*, 10(Suppl 1):S12, 2009.

   Background: Metagenomics is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution, in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans. The recent development of ultra-high throughput sequencing technologies, which do not require cloning or PCR amplification, and can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects. Increasingly, there is a need for new ways of comparing multiple metagenomics datasets, and for fast and user-friendly implementations of such approaches.

   Results: This paper introduces a number of new methods for interactively exploring, analyzing and comparing multiple metagenomic datasets, which will be made freely available in a new, comparative version 2.0 of the stand-alone metagenome analysis tool MEGAN.

   Conclusion: There is a great need for powerful and user-friendly tools for comparative analysis of metagenomic data and MEGAN 2.0 will help to fill this gap.

9. Heinz Stockinger, Alexander F. Auch, Markus Göker, Jan Meier-Kolthoff, Alexandros Stamatakis. **Large-Scale Co-Phylogenetic Analysis on the Grid.** *International Journal of Grid and High Performance Computing*, 1(1):39–54. 2009.

   Phylogenetic data analysis represents an extremely compute-intensive area of Bioinformatics and thus requires high-performance technologies. Another compute- and memory-intensive problem is that of hostparasite co-phylogenetic analysis: given two phylogenetic trees, one for the hosts (e.g.,

mammals) and one for their respective parasites (e.g., lice) the question arises whether host and parasite trees are more similar to each other than expected by chance alone. CopyCat is an easy-to-use tool that allows biologists to conduct such co-phylogenetic studies within an elaborate statistical framework based on the highly optimized sequential and parallel AxParafit program. We have developed enhanced versions of these tools that efficiently exploit a Grid environment and therefore facilitate large-scale data analyses. Furthermore, we developed a freely accessible client tool that provides co-phylogenetic analysis capabilities. Since the computational bulk of the problem is embarrassingly parallel, it fits well to a computational Grid and reduces the response time of large scale analyses.

10. Alexander F. Auch, Hans-Peter Klenk, Markus Göker. **Standard operating procedure for calculating genome-to-genome distances based on high-scoring sequence pairs.** To appear in *Standards in Genomic Sciences.*

    DNA-DNA hybridization (DDH) is a widely applied wet-lab technique to obtain an estimate of the overall similarity between the genomes of two organisms. To base the species concept for prokaryotes ultimately on DDH was chosen by microbiologists as a pragmatic approach for deciding about the recognition of novel species but also allowed a relatively high degree of standardization compared to other areas of taxonomy. However, DDH is tedious and error-prone and first and foremost cannot be used to incrementally establish a comparative database. Recent studies have shown that in-silico methods for the comparison of genome sequences can be used to replace DDH. Considering the ongoing rapid technological progress of sequencing methods, genome-based prokaryote taxonomy is coming into reach. However, calculating distances between genomes is dependent on multiple choices for software and program settings. We here provide an overview over the modifications that can be applied to distance methods based on high-scoring sequence pairs (HSPs) or maximally unique matches (MUMs) and that need to be documented. As a reference implementation, we introduce the GGDC web server (http://www.gbdp.org/species).

11. Alexander F. Auch, Mathias von Jan, Hans-Peter Klenk, Markus Göker. **Digital DNA-DNA hybridization for microbial species**

delineation by means of genome-to-genome sequence comparison. To appear in *Standards in Genomic Sciences*.

The pragmatic species concept for Bacteria and Archaea is ultimately based on DNA-DNA hybridization (DDH). While enabling the taxonomist, in principle, to obtain an estimate of the overall similarity between the genomes of two strains, this technique is tedious and error-prone and cannot be used to incrementally built up a comparative database. Recent technological progress in the area of genome sequencing calls for bioinformatics methods to replace the wet-lab DDH by in-silico genome-to-genome comparison. We here investigate state-of-the-art methods for inferring whole-genome distances in their ability to mimic DDH. Algorithms to efficiently determine high-scoring sequences pairs or maximally unique matches perform well as a basis of inferring intergenomic distances. The examined distance functions, which are able to cope with heavily reduced genomes and repetitive sequence regions, outperform previously described ones regarding correlation with and error ratios in emulating DDH. Simulation of incompletely sequenced genomes indicates that some distance formulas are very robust against missing fractions of genomic information. Digitally derived genome-to-genome distances show a better correlation with 16S rRNA gene sequence distances than DDH values. The future perspectives of genome-informed taxonomy are discussed, and the investigated methods are made available as a web service for genome-based species delineation.

A.2 Other Publications

12. Alexander F. Auch, Stefan. R. Henz, Markus Göker. **Phylogenies from whole Genomes - Methodological update within a distance-based framework.** Poster at GCB 2006, and additionally published via *TOBIAS-lib*. URN: urn:nbn:de:bsz:21-opus-34178. 2006.

    Methods which derive pairwise distances directly from complete sequenced genomes are a potentially important and efficient tool within the growing field of phylogenomics. We have shown in two previous studies that the Genome BLAST Distance Phylogeny (GBDP) approach leads to reliable phylogenetic estimates if applied to prokaryotic as well as plastid and mitochondrial genomes. Basically, GBDP first invokes tools such as BLAST to identify high-scoring segment pairs (HSPs) between all pairs of genomes; afterwards, pairwise distances are estimated based on different formulae.

    Here, we examine (1) a new GBDP distance formula, based on a combination of two previously existing ones; (2) use of BLAT instead of BLASTN and TBLASTX HSP search; (3) an alternative measure for the agreement of a distance matrix with a predefined reference topology; (4) alternative topology-independent measures of distance quality per se. All examinations were based on a enlarged dataset compared to that used in our previous study, additionally containing interesting key taxa.

A.3 Submitted Manuscripts

13. Markus Göker, Guido W. Grimm, Alexander F. Auch, Ralf Aurahs, Michal Kučera. **A clustering optimization strategy for molecular taxonomy and its application to planktonic foraminifera SSU rDNA.** Submitted to *BMC Biology*.

    Background: Identifying species is challenging in the case of organisms for which often only molecular data are available. Even if morphological characteristics are well established, molecular taxonomy is often necessary to emend current taxonomic concepts and to analyze environmental DNA sequences. Typically, for this purpose clustering approaches to delineate molecular operational taxonomic units have been applied using arbitrary choices regarding the distance formulae, threshold values and clustering algorithms. Also, calculation of distance matrices has proved difficult in the case of high alignment ambiguity.

    Results: Here, we report on a clustering optimization method to establish a molecular taxonomy of planktonic foraminifera based on highly divergent small subunit rDNA sequences. The method enables one to determine the combination of alignment program, distance function and clustering setting that results in an optimal agreement with non-molecular reference data. Optimization was applied to both alignment-based and alignment-free distance calculation. The latter, which we adapted for use with partly non-homologous sequence fragments caused by distinct primer pairs, outperformed multiple sequence alignment. Resampling and permutation methods indicate that clustering optimization is robust regarding taxon sampling and, importantly, against errors in the reference data.

    Conclusion: Our approach offers new perspectives for barcoding of species diversity and for environmental sequencing, where the carriers of the analysed DNA are unknown. Clustering optimization is a general tool for selecting methods and settings for molecular taxonomy which bridges the gap between traditional and modern taxonomic disciplines by specifically addressing the issue of how to optimally account for both traditional species concepts and genetic divergence.

14. Markus Göker, Alexander F. Auch, Daniel H. Huson. **Phylogenetic Accuracy of Alignment-Based and Alignment-Free Methods:**

Effects of Indel Rate and Violation of Fragment Homology.
Submitted to *Systematic Biology*.

Phylogenetic inference from molecular sequences usually proceeds in two steps: multiple sequence alignment followed by subsequent reconstruction of trees. As an alternative, several types of alignment-free sequence comparison methods have been introduced in recent years, but their performance has not been as thoroughly tested. Here we compare the phylogenetic accuracy of common alignment-based algorithms with a variety of recently introduced alignment-independent inference algorithms. Simulations of nucleotide sequence data containing gaps are conducted to test two main hypotheses: (1) in the case of high rates of gap insertion and deletion, alignment-free methods outperform tree inference from sequence alignments; and (2) the violation of the fragment homology condition has a profound impact on the relative accuracy of alignment-independent approaches. The same inference methods are also applied to real-world nucleotide datasets chosen so as to display considerable sequence diversity. The results of the simulation study confirm both hypotheses. If sequences are trimmed irregularly to cause deviations from fragment homology, some alignment-independent algorithms significantly outperform others. A correction method is introduced which increases the robustness of methods that identify high-scoring segment pairs. Empirical tests indicate that the violation of the fragment homology condition has a severe impact on phylogenetic accuracy in real-world datasets, probably caused by sequencing with distinct primer pairs. If the indel rate in the simulation is increased above a specific turning point, the best alignment-free methods are significantly more accurate than any alignment-based approach. Within limits, relative performance can be predicted by a simple measure of alignment variability. However, under conditions of very high indel rates, all algorithms perform poorly in absolute terms. If applied to some of the real-world datasets, the best alignment-free methods outperform multiple sequence alignment, but only insignificantly so.

Appendix B

Contribution

Here, I want to separate the contributions of others from my work clearly and in detail.

Chapter 2: Whole-Genome Phylogeny

The basic idea of the `GBDP` strategy was developed by Daniel Huson, Stefan Henz and myself (see Henz et al. 2005). Together with Markus Göker and Stefan Henz, we considerably refined the `GBDP` framework in Auch et al. (2006b;a). The implementation of `GBDP` in Java was done by me, as well as most phylogenetic inferrences. Based on his profound knowledge on eukaryotic taxonomy and biological data analysis, Markus conceptualized the empirical studies and the analyses of the results. He also developed the idea behind the variance estimation and the single-locus `GBDP` adaptation.

Design, taxon sampling, and implementation of necessary modifications of the `GBDP` framework for the large prokaryotic dataset was done by me.

Chapter 3: Detection of Horizontal Gene Transfer in Prokaryotes

In 2006, I developed the concept for the HGT survey together with Stefan Henz and Stephan Steigele. Stephan did the initial implementation of the topological method. In 2009, we decided to refine the topological method, and I re-implemented a new algorithmic approach using Java, with considerable contribution by Markus Göker who developed the idea of the clade score. Markus and me conceptualized the optimization strategy and the comparison of the different HGT search methods. I implemented all necessary scripts, carried out the survey and the final data analysis. Also, I am responsible (and maybe, blamable) for the biological interpretation of the detected HGT events.

Chapter 4: Cophylogenetic studies

`CopyCat` was developed by Jan P. Meier-Kolthoff during his diploma thesis (Meier-Kolthoff 2006), which was supervised by Markus Göker and me. The parts of the code dealing with phylogenetic trees were implemented by me, as well as the tree measures and the interface to the NCBI taxonomy. I created the installation archives and the `CopyCat` website. Maintenance of the `CopyCat` application and the website is also performed by me.

In 2007, Alexandros Stamatakis developed `AxParafit` and `AxPcoords`. Integration into `CopyCat` was done by Jan and me. Performance tests of `AxParafit` and `AxPcoords` were conducted by Alexandros. The "gridified" version of both ax-programs was implemented by Alexandros and Heinz Stockinger (Stockinger et al. 2009). The Grid-enabled version of `CopyCat` was developed by Jan and me.

Chapter 5: Metagenomics

`MEGAN` was conceptualized, designed and implemented by Daniel Huson. I provided the code for converting the NCBI taxonomy to a tree representation, and conducted the simulation studies described in Huson et al. (2007a). The simulation studies were conceptualized by Daniel Huson and Stephan Schuster. The `FragAssigner` was designed and implemented by me, as well as the idea to merge HSPs to improve the classification of contigs.

`MetaSim` mainly was developed by Felix Ott during his diploma thesis, which was supervised by Daniel Richter and me. Felix based his work on the sourcecode of `ReadSim`, which was developed by Ramona Schmid. Daniel Richter, Daniel Huson and me drafted the basic ideas for `MetaSim`. I contributed the code for the population sampler, which was integrated into `MetaSim` by Felix. The simulation study (Richter et al. 2008) was designed and conducted together with Daniel Richter.

Appendix C

Supplementary Material

C.1 Schema of the GBDP storage database

```
CREATE TABLE ta_file_stats (
    stats_idx bigint NOT NULL,
    name character varying(256) NOT NULL,
    size bigint,
    modify_time timestamp with time zone
);

ALTER TABLE ONLY ta_file_stats
    ADD CONSTRAINT ta_stats_primarykey PRIMARY KEY (stats_idx);
CREATE UNIQUE INDEX ta_stats_idx_name ON ta_file_stats USING btree (name);
ALTER TABLE ta_file_stats CLUSTER ON ta_stats_idx_name;


CREATE TABLE ta_file_streams (
    stream_idx bigint NOT NULL,
    stype integer DEFAULT 0,
    stream bytea,
    stats_idx bigint NOT NULL
);

ALTER TABLE ONLY ta_file_streams ALTER COLUMN stream SET STORAGE EXTERNAL;
ALTER TABLE ONLY ta_file_streams
    ADD CONSTRAINT ta_file_streams_primarykey PRIMARY KEY (stream_idx);
CREATE INDEX ta_file_streams_stats_idx ON ta_file_streams USING btree (stats_idx);

ALTER TABLE ONLY ta_file_streams
  ADD CONSTRAINT ta_file_stats_fk_stats_idx FOREIGN KEY (stats_idx)
  REFERENCES ta_file_stats(stats_idx) ON UPDATE CASCADE ON DELETE CASCADE;


CREATE FUNCTION update_stats() RETURNS trigger
    AS $$DECLARE
```

```
  bloblen bigint;
  currDate timestamp;
  doUpdate boolean;
  statsIdx bigint;
BEGIN
  doUpdate:=FALSE;

  IF (TG_OP = 'DELETE') THEN
    IF (OLD.stype = 0) THEN
      bloblen:=0;
      statsIdx:=OLD.stats_idx;
      doUpdate:=TRUE;
    END IF;
  ELSE -- update or insert
    IF (NEW.stype = 0) THEN
      bloblen:=length(NEW.stream);
      statsIdx:=NEW.stats_idx;
      doUpdate:=TRUE;
    END IF;
  END IF;

  IF (doUpdate=TRUE) THEN
    currDate:=now();
    update ta_file_stats set size=bloblen, modify_time=currDate
      where stats_idx=statsIdx;
  END IF;

  return NEW;
END;
$$
    LANGUAGE plpgsql;

CREATE TRIGGER streams_update_stats_trig
    AFTER INSERT OR DELETE OR UPDATE ON ta_file_streams
    FOR EACH ROW
    EXECUTE PROCEDURE update_stats();


CREATE SEQUENCE ta_file_stats_stats_idx_seq
    INCREMENT BY 1
    NO MAXVALUE
    NO MINVALUE
    CACHE 1;


CREATE SEQUENCE ta_file_streams_stream_idx_seq
    INCREMENT BY 1
    NO MAXVALUE
    NO MINVALUE
```

```
    CACHE 1;


ALTER TABLE ta_file_stats ALTER COLUMN stats_idx
  SET DEFAULT nextval('ta_file_stats_stats_idx_seq'::regclass);
ALTER TABLE ta_file_streams ALTER COLUMN stream_idx
  SET DEFAULT nextval('ta_file_streams_stream_idx_seq'::regclass);
```

C.2 Archaeal Consensus Networks

The Consensus networks (Holland et al. 2004) were generated using `SplitsTree4` Huson and Bryant (2006). A list of shared genes for the Archaeal lineage was obtained using a revised approach based on Henz et al. (2004).

Figure C.1 shows a network comprising all 146 shared genes. Additionally, we generated a network based on a subset showing the highest amount of correspondence among themselves and to the NCBI taxonomy (Figure C.2).

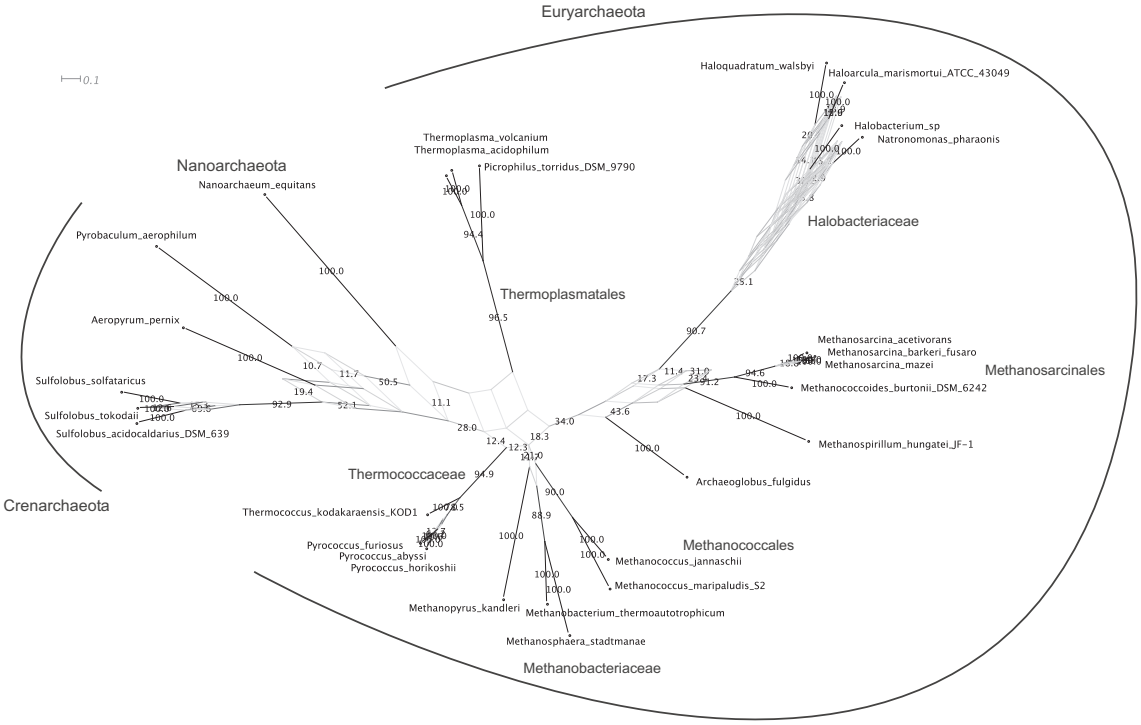


Figure C.1: Consensus network for 28 Archaea based on 146 common genes (each with 100 bootstrap replicates) using a threshold of 10%. The graph contains 47 non-trivial splits and has a c-score of 0.91. The numbers shown represent the bootstrap support for the corresponding edge.

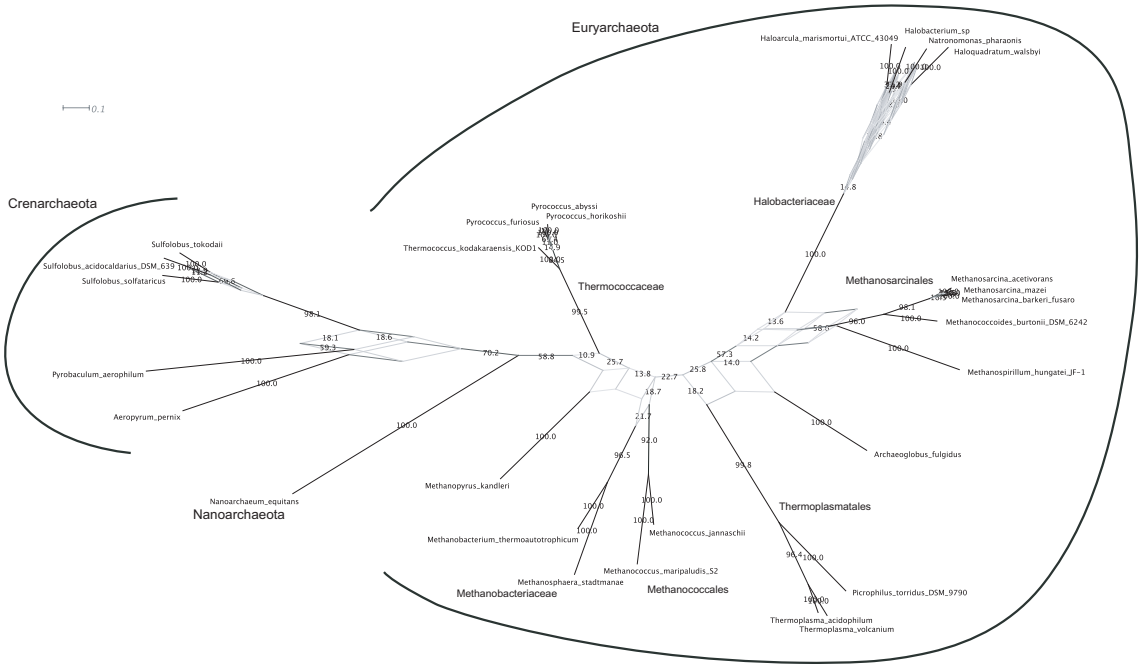


Figure C.2: Consensus network for 28 Archaea based on the bootstrapping replicates of the 25 best genes (threshold of 10%). The graph contains 47 non-trivial splits and has a c-score of 0.96. The numbers represent the bootstrap support for the corresponding edge.

Bibliography

Abascal F, Zardoya R, Posada D, 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics, 21(9):2104–2105.

Abascal F, Zardoya R, Posada D, 2007. ProtTest 1.4 Manual.

ACML, 2007. AMD Core Math Library. http://www.amd.com/acml.

Aguileta G, Marthey S, Chiapello H, Lebrun MH, Rodolphe F, Fournier E, Gendrault-Jacquemard A, Giraud T, 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst Biol, 57(4):613–627.

Altenhoff AM, Dessimoz C, 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol, 5(1):e1000262.

Altschul SF, 1991. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol, 219(3):555–565.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. Journal of Molecular Biology, 215(3):403–410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25(17):3389–3402.

Anisimova M, Gascuel O, 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol, 55(4):539–552.

Apic G, Gough J, Teichmann SA, 2001a. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol, 310(2):311–325.

Apic G, Gough J, Teichmann SA, 2001b. An insight into domain combinations. Bioinformatics, 17 Suppl 1:S83–S89.

Auch AF, Henz SR, Göker M, 2006a. Phylogenies from whole genomes: Methodological update within a distance-based framework. published via TOBIAS-lib. URN: urn:nbn:de:bsz:21-opus-34178.

Auch AF, Henz SR, Holland BR, Göker M, 2006b. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. BMC Bioinformatics, 7:350.

Auch AF, Klenk HP, Göker M, 2009a. Standard operating procedure for calculating genome-to-genome distances based on high-scoring sequence pairs. To appear in Standards in Genomic Sciences.

Auch AF, von Jan M, Klenk HP, Göker M, 2009b. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. To appear in Standards in Genomic Sciences.

Bandelt HJ, Dress AW, 1992a. A canonical Decomposition Theory for Metrics on a Finite Set. Adv. Math., 92:47–105.

Bandelt HJ, Dress AW, 1992b. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol, 1(3):242–252.

Bapteste E, Boucher Y, 2009. Epistemological impacts of horizontal gene transfer on classification in microbiology. Methods Mol Biol, 532:55–72.

Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam J, Doolittle WF, 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. Mol Biol Evol, 25(1):83–91.

Begerow D, Göker M, Lutz M, Stoll M, 2004. On the evolution of smut fungi and their hosts. In Agerer R, Blanz P, Piepenbring M, editors, Frontiers in Basidiomycete Mycology, IHW Press, München, pp. 81–98.

Beiko RG, Harlow TJ, Ragan MA, 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A, 102(40):14332–14337.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL, 2008. GenBank. Nucleic Acids Res, 36(Database issue):D25–D30.

Bentley DR, 2006. Whole-genome re-sequencing. Curr Opin Genet Dev, 16(6):545–552.

Bentley SD, Maiwald M, Murphy LD, et al., 2003. Sequencing and analysis of the genome of the Whipple's disease bacterium Tropheryma whipplei. Lancet, 361(9358):637–644.

Berry A, 2008. Natural Selection & Beyond. The Intellectual Legacy of Alfred Russel Wallace, Oxford University Press, chapter "Ardent Beetle Hunters": Natural History, Collecting, and the Theory of Evolution, pp. 47–65.

Bininda-Emonds ORP, 2005. Supertree construction in the genomic age. Methods Enzymol, 395:745–757.

Birnbaum D, Coulier F, Pbusque MJ, Pontarotti P, 2000. "Paleogenomics": looking in the past to the future. J Exp Zool, 288(1):21–22.

Blanchette, Bourque, Sankoff, 1997. Breakpoint Phylogenies. Genome Inform Ser Workshop Genome Inform, 8:25–34.

BLAS, 2007. Basic Linear Algebra Package Homepage. http://www.netlib.org/blas.

Bordenstein SR, Paraskevopoulos C, Hotopp JCD, Sapountzis P, Lo N, Bandi C, Tettelin H, Werren JH, Bourtzis K, 2009. Parasitism and mutualism in Wolbachia: what the phylogenomic trees can and cannot say. Mol Biol Evol, 26(1):231–241.

Boucher Y, Bapteste E, 2009. Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. Bioessays, 31(5):526–536.

Bové JM, 1993. Molecular features of mollicutes. Clin Infect Dis, 17 Suppl 1:S10–S31.

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H, 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol, 54(5):743–757.

Brooks D, 1981. Hennig's parasitological method: a proposed solution. Systematic Zoology, 30(3):229–249.

Brooks D, 1990. Parsimony analysis in historical biogeography and coevolution: methodological and theoretical update. Systematic Zoology, 39(1):14–30.

Bryant D, 2003. A classification of consensus methods for phylogenetics. In Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS, editors, Bioconsensus. American Mathematical Society, volume 61 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 163–183.

Bryant D, Moulton V, 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol, 21(2):255–265.

Buneman P, 1971. The recovery of trees from measures of dissimilarity. In Hodson FR, Kendall DG, Tautu P, editors, Mathematics in the Archaeological and Historical Sciences, Edinburgh: Edinburgh University Press, pp. 387–395.

Burrows M, Wheeler DJ, 1994. A Block-sorting Lossless Data Compression Algorithm. Technical Report 124, Digital Equipment Corporation.

Caldon CE, Yoong P, March PE, 2001. Evolution of a molecular switch: universal bacterial GTPases regulate ribosome function. Mol Microbiol, 41(2):289–297.

Carroll L, 1865. Alice's Adventures in Wonderland. MacMillian and Co., London.

Castresana J, 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution, 17(4):540–552.

Cavalier-Smith T, 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol, 52(Pt 1):7–76.

Charlebois RL, Doolittle WF, 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res, 14(12):2469–2477.

Charleston MA, 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. Math Biosci, 149(2):191–223.

Charleston MA, Page RDM, 2002. TreeMap Homepage. http://taxonomy.zoology.gla.ac.uk/rod/treemap.html.

Charleston MA, Perkins SL, 2006. Traversing the tangle: algorithms and applications for cophylogenetic studies. J Biomed Inform, 39(1):62–71.

Chen F, Mackey AJ, Vermunt JK, Roos DS, 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE, 2(4):e383.

Choi IG, Kim SH, 2007. Global extent of horizontal gene transfer. Proc Natl Acad Sci U S A, 104(11):4489–4494.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P, 2006. Toward automatic reconstruction of a highly resolved tree of life. Science, 311(5765):1283–1287.

Claesson MJ, O'Sullivan O, Wang Q, Nikkil J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW, 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One, 4(8):e6669.

Cole JR, Wang Q, Cardenas E, et al., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res, 37(Database issue):D141–D145.

Colless DH, 1982. Phylogenetics: The Theory and Practice of Phylogenetic Systematics. Systematic Zoology, 31(1):100–104.

Cook RD, 1979. Influential Observations in Linear Regression. Journal of the American Statistical Association, 74(365):169–174.

Cope J, Oberg M, Tufo HM, Woitaszek M, 2005. Shared Parallel Filesystems in Heterogeneous Linux Multi-Cluster Environments. In Proceedings of the 6th LCI International Conference on Linux Clusters.

Dagan T, Artzy-Randrup Y, Martin W, 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A, 105(29):10039–10044.

Dagan T, Martin W, 2006. The tree of one percent. Genome Biol, 7(10):118.

Dagan T, Martin W, 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A, 104(3):870–875.

Dagan T, Martin W, 2009. Getting a better picture of microbial evolution en route to a network of genomes. Philos Trans R Soc Lond B Biol Sci, 364(1527):2187–2196.

Darwin CR, 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray.

Dayhoff M, Schwarz R, Orcutt B, 1978. A model of evolutionary change in proteins. In Dayhoff M, Schwarz R, Orcutt B, editors, Atlas of protein sequence and structure, Maryland, pp. 345–352.

De Soete G, 1986. Optimal variable weighting for ultrametrix and additive tree clustering. Quality&Quantity, 20:169–180.

de Vienne DM, Giraud T, Martin OC, 2007. A congruence index for testing topological similarity between trees. Bioinformatics, 23(23):3119–3124.

Delgado-Friedrichs O, Dezulian T, Huson DH, 2003. A meta-viewer for biomolecular data. GI Jahrestagung, 1:375–380.

Deloger M, Karoui ME, Petit MA, 2009. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol, 191(1):91–99.

Desper R, Gascuel O, 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol, 9(5):687–705.

Desper R, Gascuel O, 2004. Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting. Systematic Biology, 21(3):587–598.

Di Giulio M, 2006. The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA). J Theor Biol, 240(3):343–352.

Dimmic MW, Rest JS, Mindell DP, Goldstein RA, 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol, 55(1):65–73.

Doolittle WF, 1999a. Lateral genomics. Trends Cell Biol, 9(12):M5–M8.

Doolittle WF, 1999b. Phylogenetic classification and the universal tree. Science, 284(5423):2124–2129.

Doolittle WF, 2000. Uprooting the tree of life. Sci Am, 282(2):90–95.

Doolittle WF, Bapteste E, 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A, 104(7):2043–2049.

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P, 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res, 33(1):e6.

Durbin R, Eddy SR, Krogh A, Mitchison G, 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

EBI, 2005. Homepage. http://www.ebi.ac.uk/genomes/organelle.html.

Edgar R, 2004a. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5):1792–1797.

Edgar RC, 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5:113.

Edgar RC, 2004c. MUSCLE: Low-complexity multiple sequence alignment with T-Coffee accuracy. ISMB/ECCB 2004, http://www.iscb.org/ismbeccb2004/short%20papers/07.pdf.

Efron B, 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1):1–26.

Efron B, Halloran E, Holmes S, 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A, 93(23):13429–13434.

Eichler W, 1948. Some rules in ectoparasitism. The Annals and Magazin of Natural History (Series 12), 1 (Series 12:588–598.

Eisen JA, 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev, 10(6):606–611.

EMBL Database, 2007. http://www.ebi.ac.uk/embl.

Emelyanov VV, 2003a. Common evolutionary origin of mitochondrial and rickettsial respiratory chains. Arch Biochem Biophys, 420(1):130–141.

Emelyanov VV, 2003b. Mitochondrial connection to the origin of the eukaryotic cell. Eur J Biochem, 270(8):1599–1618.

Emelyanov VV, 2003c. Phylogenetic affinity of a Giardia lamblia cysteine desulfurase conforms to canonical pattern of mitochondrial ancestry. FEMS Microbiol Lett, 226(2):257–266.

Enright AJ, Dongen SV, Ouzounis CA, 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 30(7):1575–1584.

Esser C, Ahmadinejad N, Wiegand C, et al., 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol, 21(9):1643–1660.

Farahi K, Pusch GD, Overbeek R, Whitman WB, 2004. Detection of lateral gene transfer events in the prokaryotic tRNA synthetases by the ratios of evolutionary distances method. J Mol Evol, 58(5):615–631.

Faraway JJ, 2002. Practical regression and Anova using R. http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Farris J, 1972. Estimating phylogenetic trees from distance matrices. The American Naturalist, 106:645–668.

Felsenstein J, 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology, 27:401–410.

Felsenstein J, 1984. Distance methods for inferring phylogenies: a justification. Evolution, 38(1):16–24.

Felsenstein J, 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39:783–791.

Felsenstein J, 2004. Inferring phylogenies. Sinauer Associates, Massachusetts.

Felsenstein J, Kishino H, 1993. Is There Something Wrong with the Bootstrap on Phylogenies? A Reply to Hillis and Bull. Systematic Biology, 42(2):193–200.

Fitch WM, 2000. Homology a personal view on some of the problems. Trends Genet, 16(5):227–231.

Forterre P, Philippe H, 1999. Where is the root of the universal tree of life? Bioessays, 21(10):871–879.

Foster I, 2005. Globus Toolkit Version 4: Software for Service-Oriented Systems. In IFIP International Conference on Network and Parallel Computing. Springer-Verlag, pp. 2–13.

Foster I, Kesselman C, Tuecke S, 2001. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications, 15(3):200–222.

Foster PG, Jermiin LS, Hickey DA, 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol, 44(3):282–288.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP, 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science, 323(5915):741–746.

Gadagkar SR, Kumar S, 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol, 22(11):2139–2141.

Gadagkar SR, Rosenberg MS, Kumar S, 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zoolog B Mol Dev Evol, 304(1):64–74.

Galperin MY, Koonin EV, 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res, 32(18):5452–5463.

Galtier N, Daubin V, 2008. Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci, 363(1512):4023–4029.

Garamszegi LZ, 2009. Patterns of co-speciation and host switching in primate malaria parasites. Malar J, 8:110.

Gascuel O, 1997. BIONJ: An improved versions of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution, 14(7):685–695.

Ge F, Wang LS, Kim J, 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol, 3(10):e316.

Glansdorff N, Xu Y, Labedan B, 2008. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biol Direct, 3:29.

Goddard W, Kubicka E, Kubicki G, McMorris FR, 1994. The agreement metric for labeled binary trees. Mathematical Biosciences, 123(2):215–226.

Göker M, 2003. Was ist Koevolution? published via TOBIAS-lib. URN: urn:nbn:de:bsz:21-opus-41589.

Göker M, Voglmayr H, Blázquez GG, Oberwinkler F, 2009. Species delimitation in downy mildews: the case of Hyaloperonospora in the light of nuclear ribosomal ITS and LSU sequences. Mycol Res, 113(Pt 3):308–325.

Gophna U, Charlebois RL, Doolittle WF, 2006. Ancient lateral gene transfer in the evolution of Bdellovibrio bacteriovorus. Trends Microbiol, 14(2):64–69.

Gophna U, Doolittle WF, Charlebois RL, 2005. Weighted genome trees: refinements and applications. J Bacteriol, 187(4):1305–1316.

Goremykin V, Hellwig F, 2005. Evidence for the most basal split in land plants dividing Bryophyte and Tracheophyte lineages. Plant Systematics and Evolution, 254:93–103.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM, 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol, 57(Pt 1):81–91.

Gotthelf A, 1999. Darwin on Aristotle. J Hist Biol, 32(1):3–30.

Gray MW, 1989. The evolutionary origins of organelles. Trends Genet, 5(9):294–299.

Gray MW, Burger G, Lang BF, 1999. Mitochondrial evolution. Science, 283(5407):1476–1481.

Gray MW, Burger G, Lang BF, 2001. The origin and early evolution of mitochondria. Genome Biol, 2(6):REVIEWS1018.

Greene JC, Mayr E, 1992. From Aristotle to Darwin: reflections on Ernst Mayr's interpretation in The Growth of biological thought. J Hist Biol, 25(2):257–284.

Gregory TR, editor, 2005. The Evolution of the Genome. Elsevier Academic Press.

Gribaldo S, Philippe H, 2004. Pitfalls in tree reconstruction and the phylogeny of Eukaryotes. In Hirt RP, Horner DS, editors, Organelles, genomes and Eukaryote phylogeny, CRC Press, Boca Raton/London/New York/Washington, D.C., pp. 133–152.

Gropp W, Lusk E, Skjellum A, 1999. Using MPI: Portable Parallel Programming with the Message Passing Interface. MIT Press, Cambridge, MA, USA, 2nd edition.

GSL, 2007. GNU Scientific Library. http://www.gnu.org/software/gsl.

Guindon S, Gascuel O, 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Molecular Biology and Evolution, 19(4):534–543.

Guindon S, Gascuel O, 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, 52(5):696–704.

Guojie JL, 2005. Professional Java Native Interfaces with SWT/JFace. Wiley.

Haeckel E, 1894. Systematische Phylogenie. Verlag von Georg Reimer, Berlin.

Halfhill T, 2008. Parallel Processing with CUDA. Microprocessor Journal.

Halpern AL, Huson DH, Reinert K, 2002. Algorithms in Bioinformatics, Springer Verlag, New York, chapter Segment Match Refinement and Applications, pp. 126–139.

Hanage WP, Fraser C, Spratt BG, 2006. Sequences, sequence clusters and bacterial species. Philos Trans R Soc Lond B Biol Sci, 361(1475):1917–1927.

Handelsman J, 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev, 68(4):669–685.

Hansen H, Bachmann L, Bakke TA, 2003. Mitochondrial DNA variation of Gyrodactylus spp (Monogenea, Gyrodactylidae) populations infecting Atlantic salmon, grayling, and rainbow trout in Norway and Sweden. Int J Parasitol, 33(13):1471–1478.

Haque MM, Ghosh TS, Komanduri D, Mande SS, 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics, 25(14):1722–1730.

Harding EF, 1971. The Probabilities of Rooted Tree-Shapes Generated by Random Bifurcation. Advances in Applied Probability, 3(1):44–77.

Hasegawa M, Hashimoto T, 1993. Ribosomal RNA trees misleading? Nature, 361(6407):23.

Hashimoto T, Nakamura Y, Kamaishi T, Nakamura F, Adachi J, Okamoto K, Hasegawa M, 1995. Phylogenetic place of mitochondrion-lacking protozoan, Giardia lamblia, inferred from amino acid sequences of elongation factor 2. Mol Biol Evol, 12(5):782–793.

Hatfield D, Diamond A, 1993. UGA: a split personality in the universal genetic code. Trends Genet, 9(3):69–70.

Hejnol A, Obst M, Stamatakis A, et al., 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci, 276(1677):4261–4270.

Henikoff S, Henikoff JG, 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A, 89(22):10915–10919.

Henz SR, Auch AF, Huson DH, Nieselt-Struwe K, Schuster SC, 2003. Whole Genome-based Prokaryotic Phylogeny. ECCB 2003, http://www.inra.fr/eccb2003/posters/pdf/short/S_huson.ps.

Henz SR, Huson DH, Auch AF, Moulton V, Schuster SC, 2004. A consensus network of prokaryotic gene trees. In Gramada A, Bourne P, editors, Proc. of RECOMB "Currents in Computational Molecular Biology 2004".

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC, 2005. Whole-genome prokaryotic phylogeny. Bioinformatics, 21(10):2329–2335.

Herbeck JT, Degnan PH, Wernegreen JJ, 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). Mol Biol Evol, 22(3):520–532.

Hess PN, De Moraes Russo CA, 2007. An empirical test of the midpoint rooting method. Biological Journal of the Linnean Society, 92(4):669–674.

Hillis D, Bull J, 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology, 42(2):182–192.

Hoff KJ, Lingner T, Meinicke P, Tech M, 2009. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res, 37(Web Server issue):W101–W105.

Holland BR, Huber KT, Dress A, Moulton V, 2002. δ plots: a tool for analyzing phylogenetic distance data. Mol Biol Evol, 19(12):2051–2059.

Holland BR, Huber KT, Moulton V, Lockhart PJ, 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol, 21(7):1459–1461.

House CH, 2009. The tree of life viewed through the contents of genomes. Methods Mol Biol, 532:141–161.

Huelsenbeck J, Rannala B, Yang Z, 1997. Statistical tests of host-parasite cospeciation. Evolution, 51(2):410–419.

Huelsenbeck JP, 1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evol, 12(5):843–849.

Huelsenbeck JP, Rannala B, Larget B, 2000. A Bayesian framework for the analysis of cospeciation. Evolution, 54(2):352–364.

Huson DH, 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics, 14(1):68–73.

Huson DH, 2003. What if I don't have a tree? Split decomposition and related models. Curr Protoc Bioinformatics, Chapter 6:Unit 6.7.

Huson DH, Auch AF, Qi J, Schuster SC, 2007a. MEGAN analysis of metagenomic data. Genome Res, 17(3):377–386.

Huson DH, Auch AF, Qi J, Schuster SC, 2007b. Metagenome analysis using MEGAN. In D Sankoff LW, Chin F, editors, Proceedings of the 5th Asia-Pacific Bioinformatics Conference. Imperial College Press, volume 5 of *Series on Advances in Bioinformatics and Computational Biology*, pp. 7–16.

Huson DH, Bryant D, 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol, 23(2):254–267.

Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC, 2009. Methods for comparative metagenomics. BMC Bioinformatics, 10(Suppl 1):S12.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R, 2007c. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics, 8:460.

Inagaki Y, Susko E, Roger AJ, 2006. Recombination between elongation factor 1alpha genes from distantly related archaeal lineages. Proc Natl Acad Sci U S A, 103(12):4528–4533.

Jain R, Rivera MC, Lake JA, 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A, 96(7):3801–3806.

Janzen D, 1985. Coevolution as process. What parasites of animals and plants do not have in common. In Kim K, editor, Coevolution of parasitic arthropods and mammals, New York: Wiley.

Jones D, Taylor W, Thornton J, 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences, 8(3):275–282.

Jukes T, Cantor C, 1969. Evolution of protein molecules. In Munro H, editor, Mammalian protein metabolism, New York: Academic Press, pp. 21–132.

Kanhere A, Vingron M, 2009. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. BMC Evol Biol, 9(1):9.

Karlin S, Altschul SF, 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A, 87(6):2264–2268.

Katoh K, Misawa K, Kuma Ki, Miyata T, 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14):3059–3066.

Kent WJ, 2002. BLAT–the BLAST-like alignment tool. Genome Res, 12(4):656–664.

Kimura M, Ohta T, 1972. On the stochastic model for estimation of mutational distance between homologous proteins. Journal of Molecular Evolution, 2(1):87–90.

Köhler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos DS, 1997. A plastid of probable green algal origin in Apicomplexan parasites. Science, 275(5305):1485–1489.

Köhlmeier M, 2002. Das große Sagenbuch des klassischen Altertums. Piper Verlag, München.

Kolaczkowski B, Thornton JW, 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature, 431(7011):980–984.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV, 2002. Selection in the evolution of gene duplications. Genome Biol, 3(2):RESEARCH0008.

Konstantinidis KT, Tiedje JM, 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A, 102(7):2567–2572.

Koonin EV, 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet, 39:309–338.

Koonin EV, Wolf YI, 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res, 36(21):6688–6719.

Korf I, Yandell M, Bedell J, 2003. BLAST - An Essential Guide to the Basic Local Alignment Search Tool. O'Reilly & Associates, Inc., first edition.

Koski LB, Golding GB, 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol, 52(6):540–542.

Krause K, 2008. From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. Curr Genet, 54(3):111–121.

Kuhner MK, Felsenstein J, 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol, 11(3):459–468.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P, 2008. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev, 72(4):557–578.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA, 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res, 15(7):954–959.

Kurland CG, Canback B, Berg OG, 2003. Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A, 100(17):9658–9662.

Kurland CG, Collins LJ, Penny D, 2006. Genomics and the irreducible nature of eukaryote cells. Science, 312(5776):1011–1014.

Lang BF, Gray MW, Burger G, 1999a. Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet, 33:351–397.

Lang BF, Seif E, Gray MW, O'Kelly CJ, Burger G, 1999b. A comparative genomics approach to the evolution of eukaryotes and their mitochondria. J Eukaryot Microbiol, 46(4):320–326.

LAPACK, 2007. Linear Algebra PACKage. http://www.netlib.org/lapack.

Laure E, Fisher SM, Frohner A, et al., 2006. Programming the Grid with gLite. Computational Methods in Science and Technology, 12(1):33–45.

Lawrence JG, Ochman H, 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol, 10(1):1–4.

Lecointre G, Philippe H, L HLV, Guyader HL, 1993. Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol, 2(3):205–224.

Lee M, 2001. Unalignable sequences and molecular evolution. Trends in Ecology and Evolution, 16(12):681–685.

Lefkovitch L, 1993. Optimal set covering for biological classification, Agriculture, Canada, p. 173.

Legendre P, Anderson MJ, 1998. Program DistPCoA. Département de sciences biologiques, Université de Montréal, p. 10.

Legendre P, Desdevises Y, Bazin E, 2002. A statistical test for host-parasite coevolution. Systematic Biology, 51(2):217–234.

Legendre P, Lapointe FJ, 2004. Assessing the congruence among distance matrices: single malt Scotch whiskies revisited. Australian and New Zealand Journal of Statistics, 46:615–629.

Legendre P, Legendre L, 1998. Numerical Ecology. Elsevier, Amsterdam, 2nd edition.

Leipe DD, Aravind L, Koonin EV, 1999. Did DNA replication evolve twice independently? Nucleic Acids Res, 27(17):3389–3401.

Lerat E, Daubin V, Ochman H, Moran NA, 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol, 3(5):e130.

Lewin B, 2004. Genes VIII. Pearson Prentice Hall, international edition.

Li L, Stoeckert CJ, Roos DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res, 13(9):2178–2189.

Lim PO, Sears BB, 1992. Evolutionary relationships of a plant-pathogenic mycoplasmalike organism and Acholeplasma laidlawii deduced from two ribosomal protein gene sequences. J Bacteriol, 174(8):2606–2611.

Linnaeus C, 1758. Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Laurentii Salvii, Holmiae, 10th edition.

Liu SV, Saunders NJ, Jeffries A, Rest RF, 2002. Genome analysis and strain comparison of correia repeats and correia repeat-enclosed elements in pathogenic Neisseria. J Bacteriol, 184(22):6163–6173.

Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T, 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol, 23(1):40–45.

Lovett PS, Ambulos NP, Mulbry W, Noguchi N, Rogers EJ, 1991. UGA can be decoded as tryptophan at low efficiency in Bacillus subtilis. J Bacteriol, 173(5):1810–1812.

Lu Y, Sze SH, 2009. Improving accuracy of multiple sequence alignment algorithms based on alignment of neighboring residues. Nucleic Acids Res, 37(2):463–472.

Luque I, Riera-Alberola ML, Andújar A, Ochoa de Alda JAG, 2008. Intraphylum diversity and complex evolution of cyanobacterial aminoacyl-tRNA synthetases. Mol Biol Evol, 25(11):2369–2389.

Makarenkov V, Legendre P, 2001. Optimal variable weighting for ultrametric and additive trees and K-means partitioning: methods and software. Journal of classification, 18(2):245–271.

Mantel N, 1967. The detection of disease clustering and a generalized regression approach. Cancer Res, 27(2):209–220.

Mantel N, Valand RS, 1970. A technique of nonparametric multivariate analysis. Biometrics, 26(3):547–558.

Mardis ER, 2008. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 9:387–402.

Margulies M, Egholm M, Altman WE, et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057):376–380.

Marri PR, Hao W, Golding GB, 2007. The role of laterally transferred genes in adaptive evolution. BMC Evol Biol, 7 Suppl 1:S8.

Martens M, Dawyndt P, Coopman R, Gillis M, Vos PD, Willems A, 2008. Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus Ensifer (including former Sinorhizobium). Int J Syst Evol Microbiol, 58(Pt 1):200–214.

Matsugi J, Murao K, Ishikura H, 1998. Effect of B. subtilis TRNA(Trp) on readthrough rate at an opal UGA codon. J Biochem, 123(5):853–858.

Mavromatis K, Ivanova N, Barry K, et al., 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods, 4(6):495–500.

McInerney JO, Cotton JA, Pisani D, 2008. The prokaryotic tree of life: past, present... and future? Trends Ecol Evol, 23(5):276–281.

McKenzie A, Steel M, 2000. Distributions of cherries for two models of trees. Math Biosci, 164(1):81–92.

Meier-Kolthoff JP, 2006. Large-scale cophylogenetic analysis. Diploma thesis, University of Tübingen, Wilhelm Schickard Institute.

Meier-Kolthoff JP, Auch AF, Huson DH, Göker M, 2007. COPYCAT: cophylogenetic analysis tool. Bioinformatics, 23(7):898–900.

Meinilä M, Kuusela J, Zietara MS, Lumme J, 2004. Initial steps of speciation by geographic isolation and host switch in salmonid pathogen Gyrodactylus salaris (Monogenea: Gyrodactylidae). Int J Parasitol, 34(4):515–526.

Meldrum D, 2000. Automation for genomics, part two: sequencers, microarrays, and future trends. Genome Res, 10(9):1288–1303.

Merkle D, Middendorf M, 2005. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. Theory Biosci, 123(4):277–299.

Meyer N, 2007. Vergleich metagenomischer Datensätze mit MEGAN. Studienarbeit, University of Tübingen, Wilhelm Schickard Institute.

Mirkin B, Muchnik I, Smith TF, 1995. A biologically consistent model for comparing molecular phylogenies. J Comput Biol, 2(4):493–507.

MKL, 2007. Intel Math Kernel Library. http://software.intel.com/en-us/intel-mkl/.

Moran NA, Mira A, 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol, 2(12):RESEARCH0054.

Nakamura Y, Itoh T, Matsuda H, Gojobori T, 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet, 36(7):760–766.

NCBI, 2005. Homepage. http://www.ncbi.nlm.nih.gov/.

NCBI, 2006. Complete Microbial Genomes. http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi.

NCBI, 2008. The Genetic Codes. http://www.ncbi.nlm.nih.gov/Taxonomy/-Utils/wprintgc.cgi.

NCBI, 2009a. Archive of completely sequenced prokaryotic genomes. ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz.

NCBI, 2009b. Clusters of Orthologous Groups database. http://www.ncbi.nlm.nih.gov/COG/new/.

NCBI, 2009c. Taxonomy Database. ftp://ftp.ncbi.nih.gov/pub/taxonomy.

Notredame C, Higgins D, Heringa J, 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology, 302:205–217.

O'Donoghue P, Luthey-Schulten Z, 2003. On the evolution of structure in aminoacyl-tRNA synthetases. Microbiol Mol Biol Rev, 67(4):550–573.

openclipart, 2009. Open Clip Art Library. http://www.openclipart.org/.

Ott M, Zola J, Stamatakis A, Aluru S, 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In SC '07: Proceedings of the 2007 ACM/IEEE conference on Supercomputing. New York, NY, USA: ACM, pp. 1–11.

Otu H, Sayood K, 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics, 19(16):2122–2130.

Page R, 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol, 43(1):58–77.

Page R, 2003. Tangled Trees. Phylogeny, Cospeciation and Coevolution. Chicago University Press.

Page R, Charleston M, 1998. Trees within trees: Phylogeny and historical associations. Trends Ecol Evol, 13:356–359.

Page RDM, 1990. Component Analysis: A valiant failure? Cladistics, 6(2):119–136.

Parkhill J, Achtman M, James KD, et al., 2000. Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. Nature, 404(6777):502–506.

Patterson DA, Hennessy JL, 2009. Computer Organization and Design: The Hardware/Software Interface. Morgan Kaufmann Publishers Inc., 4th edition.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F, 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol, 5:50.

Pisani D, Cotton JA, McInerney JO, 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol, 24(8):1752–1760.

Podell S, Gaasterland T, 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol, 8(2):R16.

Podell S, Gaasterland T, Allen EE, 2008. A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. BMC Bioinformatics, 9:419.

Poe S, 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst Biol, 47(1):18–31.

Poinar HN, Schwarz C, Qi J, et al., 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science, 311(5759):392–394.

Pop M, Salzberg SL, 2008. Bioinformatics challenges of new sequencing technology. Trends Genet, 24(3):142–149.

Poptsova MS, Gogarten JP, 2007. The power of phylogenetic approaches to detect horizontally transferred genes. BMC Evol Biol, 7:45.

Posada D, Crandall KA, 2001. Selecting the best-fit model of nucleotide substitution. Syst Biol, 50(4):580–601.

PostgreSQL, 2008. PostgreSQL homepage. http://www.postgresql.org/.

Puigbò P, Wolf YI, Koonin EV, 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol, 8(6):59.

Qi J, Wang B, Hao BI, 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol, 58(1):1–11.

Qi W, Nong G, Preston JF, Ben-Ami F, Ebert D, 2009. Comparative metagenomics of Daphnia symbionts. BMC Genomics, 10:172.

R, 2008. The R project for Statistical Computing. http://www.r-project.org.

Raes J, Foerstner KU, Bork P, 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol, 10(5):490–498.

Ragan MA, 2001. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett, 201(2):187–191.

Ragan MA, Harlow TJ, Beiko RG, 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends Microbiol, 14(1):4–8.

Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM, 2003. Tropheryma whipplei Twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res, 13(8):1800–1809.

Rappé MS, Connon SA, Vergin KL, Giovannoni SJ, 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature, 418(6898):630–633.

Rattei T, Tischler P, Arnold R, Hamberger F, Krebs J, Krumsiek J, Wachinger B, Stümpflen V, Mewes W, 2008. SIMAP–structuring the network of protein similarities. Nucleic Acids Res, 36(Database issue):D289–D292.

Refrégier G, Le Gac M, Jabbour F, Widmer A, Shykoff JA, Yockteng R, Hood ME, Giraud T, 2008. Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. BMC Evol Biol, 8:100.

Remm M, Storm CE, Sonnhammer EL, 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol, 314(5):1041–1052.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH, 2008. MetaSim: a sequencing simulator for genomics and metagenomics. PLoS ONE, 3(10):e3373.

Ricklefs RE, Fallon SM, Bermingham E, 2004. Evolutionary relationships, cospeciation, and host switching in avian malaria parasites. Syst Biol, 53(1):111–119.

Rivera MC, 2007. Genomic analyses and the origin of the eukaryotes. Chem Biodivers, 4(11):2631–2638.

Rivera MC, Lake JA, 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature, 431(7005):152–155.

Robinson D, Foulds L, 1981. Comparison of phylogenetic trees. Mathematical Biosciences, 53:131–147.

Rokas A, Williams BL, King N, Carroll SB, 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature, 425(6960):798–804.

Ronquist, 2001. Tree Fitter Homepage. http://www.ebc.uu.se/systzoo/research/treefitter/treefitter.html.

Rosselló-Mora R, 2006. Molecular Identification, Systematics and Population Structure of Prokaryotes, Springer, Berlin, chapter DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation, pp. 23–50.

Rusch DB, Halpern AL, Sutton G, et al., 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol, 5(3):e77.

Saccone C, Pesole G, 2003. Handbook of Comparative Genomics: Principles and Methodology. Wiley-Liss, first edition.

Saitou N, Nei M, 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4:406–425.

Sanjuán R, Wróbel B, 2005. Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. Syst Biol, 54(2):218–229.

Sankoff D, Blanchette M, 1997. The median problem for breakpoints in comparative genomics. In Jiang T, Lee DT, editors, Computing and Combinatorics, Proc. COCOON'97. Lecture Notes in Computer Science. volume 1276. Springer Verlag, New York.

Sankoff D, Bryant D, Deneault M, Lang BF, Burger G, 2000. Early eukaryote evolution based on mitochondrial gene order breakpoints. J Comput Biol, 7(3-4):521–535.

Saruhashi S, Hamada K, Miyata D, Horiike T, Shinozawa T, 2008. Comprehensive analysis of the origin of eukaryotic genomes. Genes Genet Syst, 83(4):285–291.

Schmid R, 2006. Metagenomics and 454 sequencing. Diploma thesis, University of Tübingen, Wilhelm Schickard Institute.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W, 2003. Human-mouse alignments with BLASTZ. Genome Res, 13(1):103–107.

Schwarz G, 1978. Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461–464.

Seward J, 2008. BZIP Homepage. http://www.bzip.org/.

Shimodaira H, 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol, 51(3):492–508.

Shimodaira H, Hasegawa M, 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics, 17(12):1246–1247.

Simonson AB, Servin JA, Skophammer RG, Herbold CW, Rivera MC, Lake JA, 2005. Decoding the genomic tree of life. Proc Natl Acad Sci U S A, 102 Suppl 1:6608–6613.

Singer GA, Hickey DA, 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol, 17(11):1581–1588.

Sirand-Pugnet P, Citti C, Barré A, Blanchard A, 2007a. Evolution of mollicutes: down a bumpy road with twists and turns. Res Microbiol, 158(10):754–766.

Sirand-Pugnet P, Lartigue C, Marenda M, et al., 2007b. Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. PLoS Genet, 3(5):e75.

Smith SA, Beaulieu JM, Donoghue MJ, 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol Biol, 9:37.

Snel B, Bork P, Huynen MA, 1999. Genome phylogeny based on gene content. Nature Genetics, 21(1):108–110.

Sokal RR, Michener CD, 1958. A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin, 28:1409–1438.

Sokal RR, Sneath PH, 1963. Principles of numerical taxonomy. San Francisco: Freeman.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM, 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science, 318(5855):1449–1452.

Soria-Carrasco V, Castresana J, 2008. Estimation of phylogenetic inconsistencies in the three domains of life. Mol Biol Evol, 25(11):2319–2329.

Spencer M, Susko E, Roger AJ, 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol, 22(5):1161–1164.

Stamatakis A, 2006a. Phylogenetic models of rate heterogeneity: a high performance computing perspective. In 20th International Parallel and Distributed Processing Symposium (IPDPS 2006), Proceedings, 25-29 April 2006, Rhodes Island, Greece. IEEE.

Stamatakis A, 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21):2688–2690.

Stamatakis A, 2009. RAxML Software homepage. http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm.

Stamatakis A, Auch AF, Meier-Kolthoff J, Göker M, 2007. AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. BMC Bioinformatics, 8:405.

Stamatakis A, Hoover P, Rougemont J, 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol, 57(5):758–771.

Steel M, 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet, 21(6):307–309.

Stevens J, 2004. Computational aspects of host-parasite phylogenies. Brief Bioinform, 5(4):339–349.

Stockinger H, 2006. Grid Computing in Physics and Life Sciences. Proceedings of World Academy of Science, Engineering and Technology, 18:1–6.

Stockinger H, 2007. Defining the grid: a snapshot on the current view. The Journal of Supercomputing, 42(1):3–17.

Stockinger H, Auch AF, Göker M, Meier-Kolthoff J, Stamatakis A, 2009. Large-Scale Co-Phylogenetic Analysis on the Grid. International Journal of Grid and High Performance Computing, 1(1):39–54.

Streit A, 2009. UNICORE: Getting to the heart of Grid technologies. eStrategies / Projects, 9th edition, British Publishers Ltd, pp. 8–9.

Studier JA, Keppler KJ, 1988. A note on the neighbour-joining algorithm of Saitou and Nei. Mol Biol Evol, 5:729–731.

Swithers KS, Gogarten JP, Fournier GP, 2009. Trees in the web of life. J Biol, 8(6):54.

Swofford D, 1991. When are phylogeny estimates from molecular and morphological data incongruent? In Phylogenetic analysis of DNA sequences, New York,Oxford: Oxford University Press, pp. 295–333.

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS, 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol, 50(4):525–539.

Talavera G, Castresana J, 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology, 56(4):564–577.

Tang J, Moret BME, 2003. Scaling up accurate phylogenetic reconstruction from gene-order data. Bioinformatics, 19 Suppl 1:i305–i312.

Tatusov RL, Fedorova ND, Jackson JD, et al., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4:41.

Tatusov RL, Galperin MY, Natale DA, Koonin EV, 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res, 28(1):33–36.

Tatusov RL, Koonin EV, Lipman DJ, 1997. A genomic perspective on protein families. Science, 278(5338):631–637.

Taylor DJ, Piel WH, 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol, 21(8):1534–1537.

Thompson J, Higgins D, Gibson C, 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22(22):4673–4680.

Thorley JL, Page RD, 2000. RadCon: phylogenetic tree comparison and consensus. Bioinformatics, 16(5):486–487.

Thornton JW, Kolaczkowski B, 2005. No magic pill for phylogenetic error. Trends Genet, 21(6):310–311.

Ulitsky I, Burstein D, Tuller T, Chor B, 2006. The average common substring approach to phylogenomic reconstruction. Journal of Computational Biology, 13(2):336–350.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC, 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One, 3(6):e2527.

Uzzell T, Corbin KW, 1971. Fitting discrete probability distributions to evolutionary events. Science, 172(988):1089–1096.

Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D, 2007. Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev, 71(3):495–548.

Vinh LS, von Haeseler A, 2005. Shortest triplet clustering: reconstructing large phylogenies using representative sets. BMC Bioinformatics, 6:92.

Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF, 2004. Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. Mol Phylogenet Evol, 33(3):615–625.

Vital-IT, 2009. http://www.vital-it.ch/.

Wallace AR, 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. III. On the tendency of varieties to depart indefinitely from the original type. J. Proc. Linn. Soc. London, 3:53–62.

Wang LS, Jansen RK, Moret BME, Raubeson LA, Warnow T, 2003. Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. http://www.smi.stanford.edu/projects/helix/psb02/wang.pdf.

Wayne LG, Brenner DJ, Colwell RR, et al., 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Bacteriol, 37(4):463–464.

Whaley RC, Petitet A, 2005. Minimizing Development and Maintenance Costs in Supporting Persistently Optimized BLAS. Software: Practice & Experience, 35(2):101–121.

Wheeler DL, Barrett T, Benson DA, et al., 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 36(Database issue):D13–D21.

Whelan S, Lio P, Goldman N, 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends in Genetics, 17(5):262–272.

Wiezer A, Merkl R, 2005. A comparative categorization of gene flux in diverse microbial species. Genomics, 86(4):462–475.

Woese CR, 1987. Bacterial evolution. Microbiol Rev, 51(2):221–271.

Woese CR, 2002. On the evolution of cells. Proc Natl Acad Sci U S A, 99(13):8742–8747.

Woese CR, Fox GE, 1977. The concept of cellular evolution. J Mol Evol, 10(1):1–6.

Woese CR, Kandler O, Wheelis ML, 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A, 87(12):4576–4579.

Woese CR, Olsen GJ, Ibba M, Söll D, 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev, 64(1):202–236.

Wolfe KH, Morden CW, Palmer JD, 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci U S A, 89(22):10648–10652.

Woyke T, Xie G, Copeland A, et al., 2009. Assembling the marine metagenome, one cell at a time. PLoS One, 4(4):e5299.

Wróbel B, 2008. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. J Appl Genet, 49(1):49–67.

Yamao F, Muto A, Kawauchi Y, Iwami M, Iwagami S, Azumi Y, Osawa S, 1985. UGA is read as tryptophan in Mycoplasma capricolum. Proc Natl Acad Sci U S A, 82(8):2306–2309.

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR, 1985. Mitochondrial origins. Proc Natl Acad Sci U S A, 82(13):4443–4447.

Yang Z, 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol, 10(6):1396–1401.

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

Yap WH, Zhang Z, Wang Y, 1999. Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol, 181(17):5201–5209.

Yule GU, 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. Philosophical Transactions of the Royal Society of London Ser. B, Biol. Sci., 213:21–87.

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N, 2009. Research in Computational Molecular Biology, Springer Berlin, chapter Deep Sequencing of a Genetically Heterogeneous Sample: Local Haplotype Reconstruction and Read Error Correction, pp. 271–284.

ZDV, 2009. Zentrum für Datenverarbeitung, High-Performance Computing resources. http://www.zdv.uni-tuebingen.de/dienste/computing/.

Zharkikh A, Li WH, 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol Phylogenet Evol, 4(1):44–63.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT, 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res, 16(9):1099–1108.

Zimmer C, 2009. Origins. On the origin of eukaryotes. Science, 325(5941):666–668.