Hybrid methods inspired by the mutual dependency of sequence alignment and phylogeny reconstruction

Thesis submitted for the degree "Doctor of Philosophy"

by

Eyal Privman

Submitted to the Senate of Tel-Aviv University

June, 2010

To my dearest Era and to my mother, for their love, support and endurance.

And to Mrs. Caren Silver of Bialik College, for teaching me the art, science and joy of English writing.

Acknowledgements

None of the research described in this manuscript was done in solitary study. I was fortunate to have several influential collaborators and advisors. First and foremost, I would like to thank my advisor Prof. Tal Pupko, from whom I learned about modeling of molecular evolution and much more. Indeed, Tal's extensive C++ code library of maximum likelihood algorithms served as the solid, fertile grounds from which sprung all of the algorithmic developments presented here. Tal has been a considerate and supportive supervisor. Most of all, I am grateful for the academic freedom to explore and to err. Tal and I shared several instances of scientific inspiration, a truly remarkable experience. We explored and studied in depth our intuitive hypotheses – some were disproved but a couple lead to valuable results, accompanied with the sense of true ownership for an original idea. I am also grateful for the wonderful research environment that Tal fosters in the lab. In this respect I would also acknowledge the extraordinary collection of the most friendly, helpful and supportive lab-mates in the past, present, and future of the Pupko group. Osnat Penn and Haim Ashkenazy were direct collaborators in the research described here, but many others have worked closely with me on related and unrelated research, which I chose not to include in this manuscript. Their assistance, contributions and advice were highly valuable. Another important friend and collaborator is Uri Shachar, who made an substantial contribution to Chapter 4 in coding the infrastructure for the iterative alignment and phylogeny reconstruction algorithm.

Dr. Matan Ninio and Prof. Nir Friedman from the Hebrew University of Jerusalem were of substantial guidance in the first chapter of my studies, which was a tight collaboration and direct continuation of Matan's graduate studies. I am in their dept for a glimpse at the deepest secrets of Bayesian modeling and inference, of the most cutting- (or rather bleeding-) edge techniques of Unix-based data processing, and of scientific rhetoric. Further significant contribution and inspiration in this chapter came from the collaboration with Dr. Julien Dutheil from University of Montpellier II in the study of bacterial phylogeny.

During my graduate studies I had the opportunity to collaborate with experimental biologists and to get a real feel for the value of bioinformatics in follow-up experimental work and the molecular biological assays that verify hypotheses in the living organisms themselves. I chose not to include this work in the present manuscript, but should nevertheless acknowledge these dear friends and companions. Adi Barzel, Dr. Uri Gophna, Prof. Gil Segal, and Prof. Martin Kupiec from the neighboring Department of Molecular Microbiology and Biotechnology were a great source of influnce, knowledge, and inspiration, and with whom I enjoyed collaborations that allowed me to find meaning for my computations in the real world. I am grateful for that experience. It has also led me to seek a position in an applied-biology laboratory for my post-doctoral studies, in the field of social insect evolution.

Finally, I would like to thank my dearest wife Era for her love and support, and especially for her endurance of some of the more trying times in these years. In the years before my graduate studies, this was the role of my parents and for that I am likewise grateful.

Eyal Privman

This work was carried out under the supervision of

Prof. Tal Pupko

Abstract

Repeated revolutions in sequencing technologies have facilitated the accumulation of large collections of homologous DNA sequences. A major endeavor in current molecular biological research is to exploit these data by comparative analysis, in order to gain insights into the function of these biological sequences. A wide range of comparative sequence analyses, from molecular phylogenetics to protein three-dimensional structure prediction, depend on multiple sequence alignment (MSA) and phylogenetic tree reconstruction as the fundamental data structures for comparative analysis. Sophisticated algorithms have been developed for both tasks, but in practice, extensive portions of the MSA and of the tree are often unreliable. The independent difficulties in each of the two challenges are exacerbated because they are inherently intertwined – MSA algorithms use a tree to guide progressive sequence alignment, while tree reconstruction algorithms rely on an MSA. This mutual dependency unavoidably leads to propagation of errors between the two stages of sequence analysis.

The studies compiled in this thesis strive to address the challenges in reducing errors in the reconstruction of both alignment and phylogeny, and explore the bidirectional passing of errors between the two. Specifically, in Chapter 3 hybrid phylogeny reconstruction methods are developed to take advantage of accurate evolutionary modeling in a Bayesian probabilistic approach in combination with efficient distance-based algorithms. Significant contribution to accuracy is achieved using models of evolutionary rate variation, and more advanced covarion-like models of site-specific rate variation are also implemented. The application of these

methods to two specific phylogenetic case studies is discussed. Next, Chapter 4 uses an iterative scheme to investigate the contribution of improved guide trees to the accuracy of different progressive alignment algorithms.

While these efforts strive to reduce errors, it is also imperative to understand and characterize the various sources of error that remain. The investigation of mutual dependency demonstrates that uncertainties in the guide trees used by progressive alignment methods are a major source of alignment uncertainty. This insight is used in Chapter 5 to develop a novel method for quantifying the robustness of each column of the alignment to uncertainty in the guide tree. Evaluation using benchmark data shows that this confidence measure accurately identifies unreliable alignment regions and allows filtering or masking of residues where errors are predicted. Chapter 6 describes an implementation of the new measure in the GUIDANCE web server, which offers powerful predictors to identify alignment errors together with the tools to deal with such errors. Thereby, researchers are provided with warning signs and preventive measures to protect downstream MSA-based analyses from the detrimental effects of alignment errors.

Throughout these studies, the proposed algorithmic improvements are evaluated using widely accepted benchmark databases including both simulated and real sequences. The methods developed allow utilization of advanced probabilistic models of sequence evolution together with the leading algorithms for phylogeny and alignment. Nevertheless, special attention is given to highly efficient algorithmic choices that permit analysis of the quantities of data

generated by the rapidly advancing sequencing technologies. Thereby, accurate analysis is now feasible for large datasets of many thousands of sequences, which previously could only have been studied by simplistic methods. Efforts were made to distribute these methods to the wider scientific community, and to formulate them as modular tools that can be merge with parallel methodological improvements. The methodologies developed herein lie at the foundation of comparative sequence analysis and are expected to contribute to the accuracy and reliability of subsequent studies of molecular biology.

Table of Contents

1 Background

The study of genetics has progressed from the early discoveries of single genes, through the characterization of specific DNA sequences and the amino acid sequences which they code, to the modern sequencing of whole genomes. Such progress would not have been possible without a continuous chain of dramatic revolutions in biotechnology, providing order-of-magnitude enhancements in sequencing yield every few years. Today, the Genome Analyzer platform (Illumina Inc.) is leading the so-called "next generation" sequencing market with a capacity to generate 1.5 giga base in a single two-day run (Ansorge 2009; Karow 2009). Sequence databanks are growing rapidly, Genbank currently holding 114 giga bases (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt), and researchers struggle with analysis of the data.

1.1 Comparative sequence analysis

To translate mass sequencing into biologically meaningful insights, advanced tools in bioinformatics have been developed for searching sequence databases and for comparative sequence analysis. A wide range of computational approaches rely on comparison of homologous sequences to arrive at predictions regarding the structure and function of macromolecules. The immediate first step after sequencing a new gene is a BLAST search that may instantly provide clues to the function of the gene, via known functions of its homologs. Many more advanced comparative analyses rely on this power of guilt by association. Typically, the basis for such comparisons is a pairwise or multiple sequence alignment (MSA).

Homologous sequences of nucleotides or amino acids are arranged one on top of the other, such that similar residues are aligned in columns (see example in Figure 1.1). It is important to make a distinction between two approaches to alignment that differ in the type of similarity they identify. Similarity among residues aligned in the same column may be due to true homology, i.e., residues descended from a common ancestral residue, either unchanged or through substitutions, while the gaps in a sequence as representing either a deletion or insertion mutation. Alternatively, similarity of aligned residues may result from analogy through convergent evolution, for example, if similar residues were independently inserted in the same position of the sequence in two separate lineages. In this case the aligned residues may carry the same structural or biochemical role, but they cannot be considered as a genuine homology.

```
ARLMEMIQEEVA------DPIVKGG
ARLMSMIQEEVS------EPLVKGG
AQLMSMIHEETS------DLFVKG-
ARLMAQIKEEPAGSKGSADPTVRGG
ARLMPLLQEETEV-------GVQGG
ARLMPLLQEEVEV-------GVQGG
```

Figure 1.1: A multiple sequence alignment. Amino-acid sequences of homologous proteins are aligned so that similar residues are arranged in columns.

An MSA may be the most fundamental data structure in bioinformatics, upon which many types of analysis are conducted. Homology search tools such as BLAST use sequence alignment to identify homologs in the database, and more sophisticated approaches such as PSI-BLAST

and Pfam use an MSA as the search query (Altschul et al. 1997; Sonnhammer, Eddy, and Durbin 1997). Many computational methodologies use the MSA in conjunction with a phylogenetic tree, which together describe the evolutionary dynamics of a sequence in terms of speciation and divergence by substitutions, insertions and deletions. Virtually all molecular evolutionary studies rely on MSAs for phylogeny reconstruction (e.g., Stamatakis, Ludwig, and Meier 2005), for inference of selection forces acting on genes (e.g., Nielsen and Yang 1998; Kim and Nielsen 2004), or for detection of recombination and horizontal gene transfer (e.g., Husmeier and Wright 2001). Usage of MSAs extends to many diverse applications from protein 3D structural modeling to definition of protein domains and sequence motifs; from population genetics to RNA secondary structure prediction. In all such studies, the reconstruction of an MSA and a tree is the necessary first step allowing downstream analyses to derive insights and predictions regarding the biological nature and function of sequences.

1.2 Sequence alignment

In the early days of molecular sequence analysis, when precious few sequences were acquired through much labor, sequences were aligned manually. Nowadays, when researchers wish to analyze large datasets ranging from dozens to thousands of sequences, such a human endeavor is not feasible. In addition, manual alignment is subjective and irreproducible. Automated alignment algorithms have been developed for decades. The first efficient dynamic algorithms for pairwise alignment were developed in the 1970's (Needleman and Wunsch 1970; Smith and Waterman 1981). These algorithms were the basis for practical solutions for multiple

alignment developed in the late 1980's (Feng and Doolittle 1987; Higgins and Sharp 1988; Lipman, Altschul, and Kececioglu 1989). All through the following two decades more and more sophisticated MSA algorithms and novel approaches were proposed, demonstrating that the problem of alignment is still not satisfactorily resolved (representative publications: Thompson, Higgins, and Gibson 1994; Morgenstern et al. 1998; Notredame, Higgins, and Heringa 2000; Katoh et al. 2002; Edgar 2004; Do et al. 2005; Lassmann and Sonnhammer 2005a; Loytynoja and Goldman 2005; Redelings and Suchard 2005; Drummond and Rambaut 2007; Bradley et al. 2009; Liu et al. 2009). These algorithms are typically used to align sequences of single genes or proteins – hundreds or thousands of characters in length. Such studies will be the focus of this dissertation. Following the genomic revolution in the last decade MSA algorithms were specially adapted to whole genome alignment, which is reviewed in Batzoglou (2005) and will not be discussed here.

Reconstructing an MSA proves to be challenging on several levels. Naturally, the more diverged the sequences the greater the challenge. In the extreme, homologous sequences that have undergone multiple substitutions per position can be considered "un-alignable". Less diverged sequences may still contain regions where the correct alignment can be difficult to discern. See for example two common causes for uncertainty in Figure 1.2. Automation of the process as a computer program adds additional challenges. Even if we could define a scoring function to determine the better MSA of any two possible solutions, it is infeasible to search for the optimal scoring alignment because the number of possible solutions grows exponentially

with the number of sequences. Formally, the problem has been shown to be NP-hard (Wang and Jiang 1994; Just 2001). Therefore, alignment programs employ heuristic approaches to generate a high-scoring MSA, although they cannot guarantee finding the optimal solution.

```
(a) ACCGGTAT        ACCGGTAT
    ACC--TAT   ?    ACC--TAT
    ACC--TAT        ACC--TAT
    ACC-ATAT        ACCA-TAT

(b) ACCGGGTAT       ACC---GGGTAT
    ACC---TAT   ?   ACC------TAT
    ACC---TAT       ACC------TAT
    ACCCGCTAT       ACCCGC---TAT
```

Figure 1.2: Examples where the correct alignment cannot be confidently decided. (a) Residues in one sequence are equally similar to several candidate homologous positions in the other. **(b)** Dissimilar stretches can either be aligned as mismatches or interpreted as independent insertion or deletion events.

1.3 Phylogeny reconstruction

Many of the bioinformatics applications that make use of an MSA do so in the light of a phylogenetic tree. Furthermore, from the time of Charles Darwin, taxonomists and phylogeneticists strived to reconstruct the tree of life, describing the evolutionary relationship among species in term of the order of their divergence (Figure 1.3).

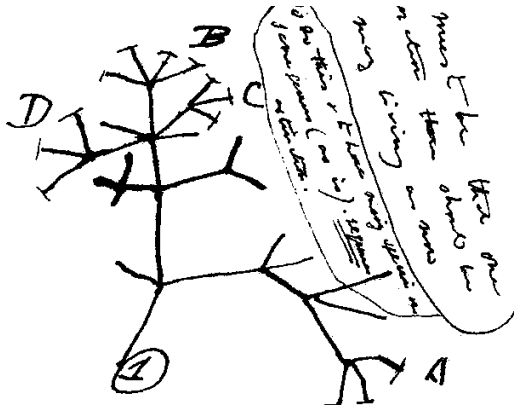

Figure 1.3: Charles Darwin's first sketch of an evolutionary tree from his ''First Notebook on Transmutation of Species'' (1837; this image is in the public domain due to the expiration of copyright)

For Darwin and his followers, and until the molecular revolution of the life sciences, phylogenies were mainly reconstructed based on morphological similarities among organisms. This approach suffers from the difficulty in comparing phenotypes that may reflect complex interactions between many genes and different environments. With the advent of amino acid sequencing in the 1970's (e.g., Goodman et al. 1974) sufficient molecular data became available to allow reconstruction of phylogeny directly from observed genotypic differences. In the next three decades, alignments of nucleotide and amino acid sequences were used as the principle raw material for phylogenetics, accompanied by the development of methodologies to reconstruct phylogenetic trees based on the comparative analysis of aligned sequences. Ideally, one may attempt to simultaneously estimate the phylogeny and MSA, although thus far such attempts proved unbearably computationally intensive (discussed in Section 1.6 below).

Therefore, all widely used approaches begin with building an alignment and then feed it as a fix input to the tree building algorithm.

Three major paradigms dominated the field of molecular phylogeny reconstruction: distance based methods, maximum parsimony, and probabilistic methods (maximum likelihood and Bayesian analysis). The first two led the earlier period. Nowadays, probabilistic methods are considered the most accurate. Distance based methods are still widely used because of their computational efficiency.

Distance based methods are the fastest way to reconstruct a tree from molecular sequences because they reduce the information of long sequences into the evolutionary distances between all pairs of sequences. The pairwise distances are then used to build the tree that best fits them. In the two most widely used distance based methods – unweighted pair group method with arithmetic mean (Sokal and Michener 1958; Sneath and Sokal 1973) and neighbor joining (Saitou and Nei 1987) – the approach is to gradually cluster sequences starting from closely related sequences, or "neighbors", and then move on to the more distantly related until the full tree is obtained. By far, the most widely used method is neighbor joining (NJ), with computational complexity of $O(n^3)$ for a set of n sequences. Moreover, NJ has been proved to be statistically consistent, that is, as the distance estimation approaches the true distances so does the estimated tree approaches the true tree (Atteson 1997). Therefore, it is the preferred choice for quick tree reconstruction. Furthermore, it is the only feasible choice for analyzing thousands of sequences, because more accurate methods suffer from a stronger dependency

on the number of sequences. Various enhancements of the NJ algorithm were proposed (e.g., Gascuel 1997; Bruno, Socci, and Halpern 2000; Howe, Bateman, and Durbin 2002; Mailund et al. 2006; Sheneman, Evans, and Foster 2006). These publications offer improvements in terms of both accuracy and efficiency. Mailund et al. (2006) achieved the most efficient NJ-like algorithm with a reduced computation time of $O(n^2)$.

Maximum parsimony (MP) was the methodology of choice until the 1980's (e.g., Holmquist et al. 1976; Goodman and Pechere 1977; Baba et al. 1981). This approach searches for the tree topology that requires the least evolutionary events (e.g., number of substitutions) to explain the observed variability in the sequences (Eck and Dayhoff 1966). Efficient algorithms were developed for finding the minimum number of events for a given tree (Kluge and Farris 1969; Fitch 1971; Sankoff 1975) and heuristics were developed for searching the solution space for the MP tree (e.g., Kumar, Tamura, and Nei 1994). Furthermore, one of the first attempts to simultaneously reconstruct the tree and the MSA, a concept that I will further discuss below (Section 1.6), was based on the MP principle (Sankoff, Morel, and Cedergren 1973; Wheeler and Gladstein 1994).

During the 1970's and 1980's several considerable shortcomings of MP became evident. Felsenstein (1978) demonstrated that MP is not statistically consistent. That is, some evolutionary trees will be incorrectly reconstructed by MP even if unlimited genetic data were available (the sequence length tends to infinity). Further studies emphasized the implications of this problem to a wider range of scenarios, beyond the classical "long branch attraction"

scenario (e.g., Kim 1996). Nevertheless, some publications still claim that MP enjoys advantages over other approaches, such as the robustness to heterogeneous evolution, i.e., when the evolutionary model and rates may change during the sequences' evolution (Kolaczkowski and Thornton 2004).

Probabilistic methods grew in popularity during the last two decades and replaced maximum parsimony as the leading paradigm. This shift followed considerable methodological research demonstrating their superiority (e.g., Saitou and Imanishi 1989; Hasegawa, Kishino, and Saitou 1991; Hasegawa and Fujiwara 1993; Tateno, Takezaki, and Nei 1994; Huelsenbeck 1995). The maximum likelihood (ML) approach is based on probabilistic models of sequence evolutions and algorithms for efficient computation of the likelihood function *L* – the conditional probability of the sequence data *D* given a tree *T* (Felsenstein 1981):

$$L(T) = P(D|T) \qquad 1\text{-}1$$

This opened the way for the developing heuristics (as in MP methods) to search the solution space for the ML tree (Felsenstein 1989; Strimmer and Von Haeseler 1996; Lewis 1998; Huelsenbeck and Ronquist 2001; Guindon and Gascuel 2003; Stamatakis, Ludwig, and Meier 2005). Typically, these heuristics start with some initial guess of the tree, which is usually acquired via a quick distance based method, and then iteratively try to make small modifications of the tree (e.g., "nearest neighbor interchange" or "subtree pruning and regrafting" that were previously used for MP tree search) and look for trees with higher

likelihood scores until no such improvements can be found (e.g., Guindon and Gascuel 2003; Stamatakis, Ludwig, and Meier 2005). Other approaches include gradual addition of taxa to the tree (Felsenstein 1995) and integrating the likelihood scores of all possible quartets of taxa (Strimmer and Von Haeseler 1996).

One of the strengths of probabilistic models of sequence evolution is that they can explicitly account for biological phenomena. For example, models were developed to relieve the unrealistic assumption of homogonous evolutionary rate (Yang 1993, elaborated on in Chapter 3) and to account for heterogeneities in the amino-acid replacement process (Lartillot and Philippe 2004).

A **Bayesian approach** to the search for the most probable tree is an alternative probabilistic paradigm. Bayesian methodology assumes a prior distribution of certain parameters of the probabilistic model. In the example of modeling among-site variation it is common to use a gamma distribution as the prior on the site-specific rates (Yang 1993). For a recent review see Pupko and Mayrose (2010). For a given rate *r*, the likelihood of the tree can be calculated as above. However, since the rate is not known, we integrate over the prior distribution to calculate the posterior probability of a tree:

$$L(T) = \int_r P(D|Tr)P(r) \qquad 1\text{-}2$$

Although a prior distribution over the rates is assumed, its parameters are commonly estimated by maximizing the likelihood function. These methods thus deviate from the classical Bayesian approach and are thus called "empirical Bayesian". Such algorithms yielded a dramatic improvement in the fit of the model to the sequence data (Yang 1994a; Yang, Goldman, and Friday 1994).

In a fully Bayesian approach the prior distribution of model parameters is sampled as well as the space of all possible trees (Rannala and Yang 1996). The most widely used sampling technique is Markov Chain Monte Carlo (MCMC), which efficiently samples the posterior distribution of trees so that high scoring solutions are sampled more than others (Yang and Rannala 1997; Larget and Simon 1999; Huelsenbeck and Ronquist 2001; Huelsenbeck et al. 2001). For a review see Holder and Lewis (2003). This approach integrates over the many degrees of freedom in complex, parameter-rich models that the ML approach does not tackle well, such as the abovementioned example of heterogeneities in the replacement process.

1.4 Estimation of confidence in phylogeny

The inferred tree can be viewed as a statistical parameter that is inferred from the data. As in other statistical inference methods, it is generally required to assess the reliability of the inferred tree. Most commonly confidence scores are assigned to each branch of the tree, which corresponds to splitting the tree into two clades. By far the most widely used approach is bootstrap sampling (Felsenstein 1985), which can be applied to any MSA-based phylogeny reconstruction method. The bootstrap is a random sample of columns from the MSA that is fed

to the same phylogeny program. For each sample of columns this procedure results in a perturbed tree, where the branching may be somewhat different from the "base" tree obtained without bootstrapping. Typically, 100 such samples are made, producing 100 trees. The bootstrap confidence score for each branch in the base tree is computed by counting the proportion of bootstrap trees that contain that branch unchanged. Thus, high scoring branches are considered robust to noise in the sequence data, whereas low scoring branches represents parts of the tree that cannot be reliably reconstructed based on the phylogenetic signal in the data. Bootstrap sampling of trees is the foundation for the alignment confidence measure developed in Chapter 5.

Within the Bayesian paradigm, an alternative confidence measure is available, based on calculations of posterior probabilities (Rannala and Yang 1996). Some studies claimed that the Bayesian confidence scores are more statistically justifiable (e.g., Alfaro, Zoller, and Lutzoni 2003), but others have shown that bootstrap is an unbiased estimate for the Bayesian posterior probability (Efron, Halloran, and Holmes 1996), and still others claimed that posterior probabilities are too liberal estimates of confidence while bootstrap scores are slightly conservative (Suzuki, Glazko, and Nei 2002).

One difficulty in the Bayesian approach is that the probability space must be evenly sampled by methods such as MCMC. It is difficult to be sure that the sampling is thorough enough, and so the maximum posterior solution may be over-confident if some other regions of the probability space are under-sampled. Furthermore, bootstrap sampling enjoys several practical

advantages, including its relative efficiency, which can be used with fast polynomial complexity algorithms such as NJ, and the easy in which it may be parallelized on multi-core high performance computer clusters.

This thesis does not deal with the comparison between the bootstrap and Bayesian approaches. In general, it may be said that the Bayesian approach offers certain advantages, but nevertheless, the bootstrap approach remains a valuable methodology especially when full Bayesian estimation is not possible due to its high requirements of computation time. See further discussion in Section 1.75 and Chapter 5.

1.5 Progressive alignment: mutual dependency of alignment and phylogeny

Knowledge of the phylogeny proves highly valuable to the reconstruction of MSAs. So much so, that virtually all the widely used, state-of-the-art alignment algorithms (e.g., Thompson, Higgins, and Gibson 1994; Notredame, Higgins, and Heringa 2000; Katoh et al. 2002; Edgar 2004; Do et al. 2005; Loytynoja and Goldman 2005) begin with reconstructing a “guide tree”, which is used to determine the order of construction of the MSA in a process known as “progressive sequence alignment”. This technique (Waterman and Perlwitz 1984; Feng and Doolittle 1987) is a heuristic that aligns pairs of sequences and then pairs of alignments, according to the branching order of the guide tree, gradually building up to the full MSA (Figure 1.4). Each pairwise alignment is computed using a dynamic programming algorithm that efficiently finds the highest scoring solution (Needleman and Wunsch 1970). Theoretically, the dynamic programming solution for a pair of sequences can be extended to any number of

sequences, but the time complexity of this algorithm is $O(l^n)$ for n sequences of length l, making it infeasible for more than a handful of sequences. The progressive heuristic is a compromise for reasonable accuracy in affordable computation time, but it is a greedy heuristic that cannot be expected to find the best scoring solution. The branching order of the guide tree is used to choose the order of pairwise alignments, starting by aligning the closest relatives (neighboring leaves) and adding the more distant and diverged sequences last (deep branches).

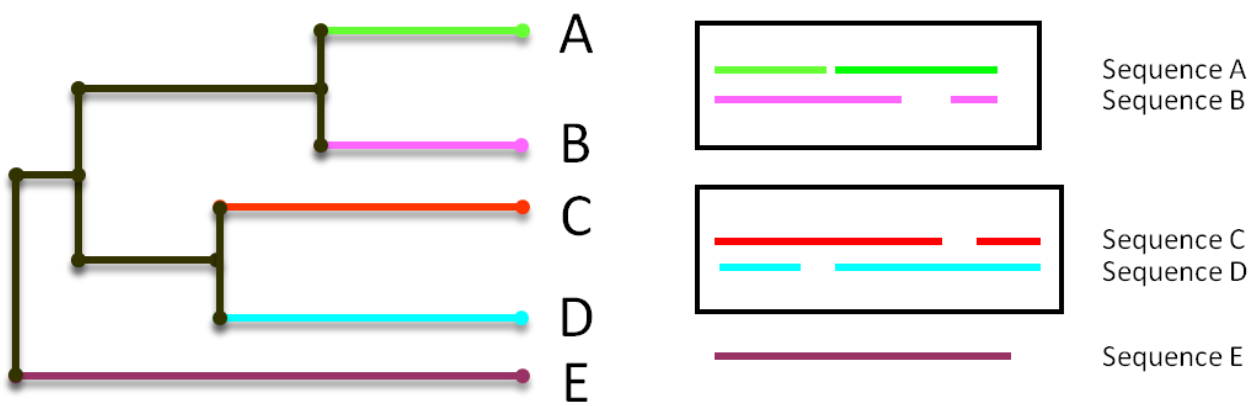

Figure 1.4: An example of progressive sequence alignment. A guide tree is used to determine the order of pairwise sequence alignments. Initially, neighboring leaves of the tree are aligned (A&B, C&D), next pairs of groups of sequences are aligned (AB&CD), and finally the four sequences are aligned to the last one (ABCD&E).

Comparative methodological evaluations consistently demonstrated that algorithms based on the progressive approach lead as the most accurate of the computationally feasible approaches (e.g., Thompson, Plewniak, and Poch 1999; Raghava et al. 2003; Gardner, Wilm, and Washietl

2005; Nuin, Wang, and Tillier 2006). See section 1.6 below for a description of these evaluation assays. These studies also highlight the considerable error rate in MSAs produced by the best algorithms. For example, Nuin, Wang, and Tillier (2006) show that typical sequence datasets often lead to an excess of 20% badly aligned residues in the MSA. Such high error rates motivated a series of improvements on the basic progressive alignment methodology, since its conception in 1987. Thompson, Higgins, and Gibson (1994) suggested several improvements including weighting sequences according to their divergence, varying the amino acid substitution matrix during the progressive alignment, and varying the gap penalties in hydrophilic regions that are probable loops in the protein structure. The significant improvement in accuracy made CLUSTALW the most popular MSA algorithm during the following decade.

Perhaps the most common improvement on the basic progressive scheme is iteration of tree building and progressive alignment (Figure 1.5, e.g., Corpet 1988; Katoh et al. 2002; Edgar 2004; Loytynoja and Goldman 2005). The rationale behind iterations stems from the mutual dependency of progressive alignment and phylogeny reconstruction – that an alignment of improved accuracy can be used to construct a tree of improved accuracy and vice versa. Of course, one must start from a tree built without an MSA or an MSA built without a tree. The common solution to this circular problem is to build a NJ or UPGMA tree based on distances calculated from pairwise alignments. Subsequent iterations recalculate the distances based on an MSA. This approach is further investigated in Chapter 4.

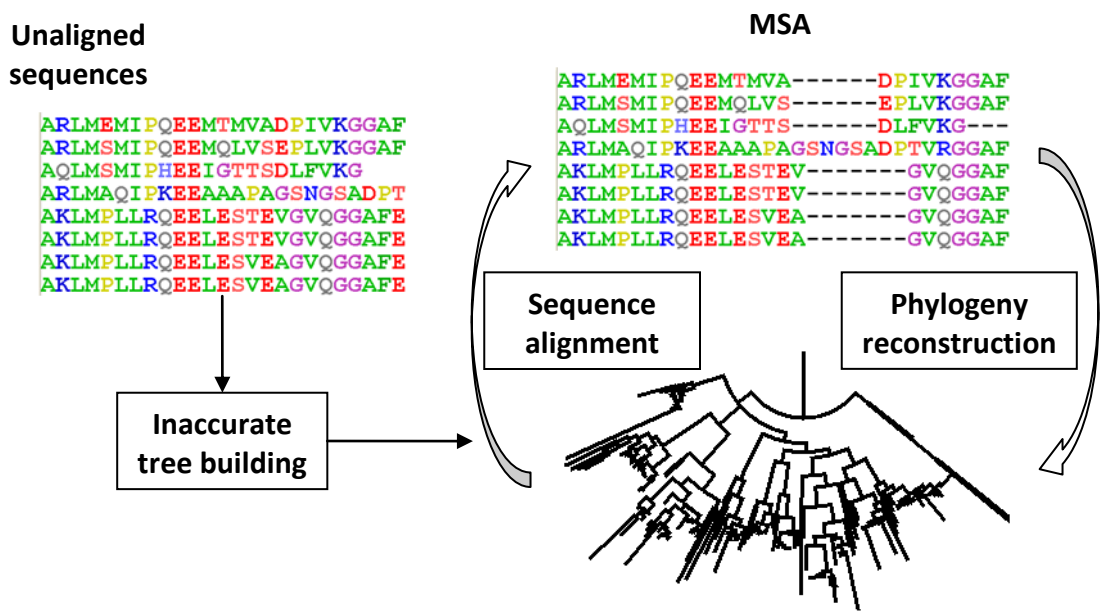

Figure 1.5: Circular dependency of alignment and phylogeny. The mutual dependency of phylogeny reconstruction MSA is commonly addressed by an iterative scheme. The first tree is built based on pairwise alignments, which are less accurate then subsequent MSAs.

As a greedy heuristic, the progressive approach suffers from a major pitfall – that early mistakes in pairwise alignments cannot be rectified with the addition of information from other sequences in latter stages. Therefore, a common practice is post-processing of the MSA known as "iterative refinement" (Gotoh 1996). The alignment is iteratively divided to two groups of sequences corresponding to two subtrees of the phylogeny, which are realigned without changing the alignment within each group. Others suggested limiting errors in each greedy step using consistency scores of the progressively build MSA with a preprocessed library of pairwise alignments (Notredame, Higgins, and Heringa 2000).

A profound correction for the widely used implementation of the progressive scheme was proposed by Loytynoja and Goldman (2005) and implemented in their algorithm PRANK. Contrary to other progressive algorithms, PRANK distinguishes between insertion and deletion events during the process of climbing down the tree from the leaves towards the root. All other algorithms treat a gap in the alignment as an "indel" (insertion or deletion) without making this distinction. The distinction is important for the handling of a gap-containing column in the deeper branches of the tree: If a gap represents a character that was inserted in a certain lineage then it should not be aligned to any other character that was inserted in an independent lineage. On the other hand, if the gap represents a deletion then the same position may be aligned to characters in sister lineages and may be independently deleted elsewhere in the phylogeny. Although at a considerable computational cost, this correction eliminates a dramatic bias towards excess deletions over insertions and the alignment of non-homologous residues in the traditional implementations of progressive alignment (Loytynoja and Goldman 2008). This is an inherent and fundamental flaw in the broadly used implementation of the progressive alignment technique. PRANK corrects this flaw in an approach entitled "phylogeny-aware gap placement" by Loytynoja and Goldman in their later Science paper (2008) where they demonstrate dramatic reduction of such errors compared to all other state-of-the-art progressive algorithms.

1.6 Performance evaluation

Despite all the above, many of the typical alignment problems continue to challenge all state-of-the-art algorithms. This is apparent from evaluation studies of alignment accuracy (e.g., Thompson, Plewniak, and Poch 1999; Raghava et al. 2003; Gardner, Wilm, and Washietl 2005; Nuin, Wang, and Tillier 2006). Such studies attempt to estimate the proportion of aligned residues that are correctly aligned – a critical quality assessment tool in the field of alignment algorithms. There is no straightforward way to do this because one needs to compare the reconstructed alignment to the "true" one, and we can never ascertain with absolute confidence the true alignment of divergent biological sequences. To do so implies full knowledge of the evolutionary history of substitutions, insertions, and deletions. Nevertheless, there are two commonly used "oracles" for obtaining knowledge of the "true" alignment, which are also used for performance evaluation in Chapter 5 of this disertation.

Simulations: Simulated evolution allows full knowledge of the true evolutionary history, including the tree and each substitution, insertion, and deletion event. Therefore, the true MSA and the tree can be compared to the reconstructed ones for evaluation of both alignment and phylogeny reconstruction algorithms. The weakness in this approach is that the simulations are based on simplified models of evolution that may not adequately mimic the evolution of real biological sequences. Thus, efforts have been made to develop richer and more realistic simulators (Stoye, Evers, and Meyer 1998; Nuin, Wang, and Tillier 2006; Fletcher and Yang 2009).

Structural homology: Benchmarks of homologous proteins with solved tertiary structures can be used to generate sequence alignments that are based on structural alignments (Thompson, Plewniak, and Poch 1999; Raghava et al. 2003; Thompson et al. 2005; Van Walle, Lasters, and Wyns 2005). A similar approach was also applied for structural RNA sequences (Gardner, Wilm, and Washietl 2005). Such alignments are considered more accurate than pure sequence-based alignment because homologous structures tend to remain similar for significantly diverged sequences. For example, completely different sequences occupying the same alpha helix in two homologous protein structures will be aligned although no similarity remains at the sequence level. An important distinction between these alignments and simulated alignments is that a pair of aligned residues is not necessarily homologous in the sense that they have evolved from an ancestral residue strictly through substitutions. They may have arisen through a series of insertions and deletions, eventually occupying the same position in the two homologous structures. Furthermore, a weakness of the structural alignment "oracle" is that usually only core secondary structure elements can be reliably aligned, while loop regions are often un-alignable.

Both of the above approaches generate a reference "true" MSA that is re-aligned using the tested alignment algorithm. The most commonly used measures for agreement of the reconstructed MSA with the reference are the column score (CS), which is the percentage of alignment columns in the reference alignment that were accurately reconstructed, and the sum-of-pairs score (SP), which is the percentage of pairs of aligned residues in the reference

MSA that are similarly aligned in the reconstructed MSA (Carrillo and Lipman 1988; Thompson, Plewniak, and Poch 1999). Note however that these scores measure type I errors – a pair of residues that should be aligned to each other are not aligned (they may be aligned to other residues or not aligned to any at all). They do not measure type II errors – residues that should not be aligned. That is, over-alignment is not penalized. This point is significant for assessing the correction by Loytynoja and Goldman (2008) described above, because they correct a bias for under-estimation of insertions that leads to over-alignment.

Simulated sequence benchmarks also make available a reference "true" phylogeny, which allows assessing the accuracy of phylogeny reconstruction (used in Chapter 3). The reconstructed phylogeny is commonly compared to the reference using the Robinson-Foulds distance (Robinson and Foulds 1979). Every branch of a tree defines a partition or "split" of the leaf labels into two subgroups. Every possible topology defines a set of partitions. Thus the Robinson-Foulds distance between two possible topologies is the proportion of common partitions (which correspond to similar branches) between the two topologies. This distance is commonly used to measure the accuracy of phylogeny reconstruction by comparing the reconstructed topology to the reference "true" topology (e.g., Stamatakis, Ludwig, and Meier 2005). For real sequence data, where the true topology can never be know with certainty, it is common to use the likelihood score of the reconstructed tree (see Section 1.3 above) to compare several algorithms.

1.7 Simultaneous estimation of alignment and phylogeny

An important alternative to the above methodologies of the progressive alignment paradigm is a Bayesian estimation of alignment, analogous to the Bayesian approach to phylogeny reconstruction (Section 1.3). An attractive advantage of this approach is the possibility to combine alignment and phylogeny in a unified estimation scheme under a joint probabilistic model of sequence evolution. Joint Bayesian estimation of alignment and phylogeny is done using the MCMC sampling described above, using a probabilistic model of sequence evolution describing substitution, deletion, and insertion events (Thorne, Kishino, and Felsenstein 1991; Thorne, Kishino, and Felsenstein 1992; Hein 2001; Lunter et al. 2005; Redelings and Suchard 2005).

This thesis will not deal with such fully Bayesian methods, however, this alternative should be considered here because it offers important advantages over the greedy progressive method. The Bayesian approach is a rigorous statistical approach to sampling the solution space. Therefore, it is expected to yield more accurate reconstruction than greedy algorithms. In fact, the small number of studies that make use of full Bayesian reconstruction clearly demonstrate its accuracy advantage. The sole reason to continue using progressive algorithms is their advantage in terms of computational efficiency. In most practical alignment problems, the Bayesian approach is infeasible. For example, the README page of the Bayesian alignment algorithm BAli-Phy (Suchard and Redelings 2006) recommends "using 12 or fewer taxa in order to limit the time required…" (http://www.biomath.ucla.edu/msuchard/bali-

phy/README.html). Even for datasets of few taxa, when genome-wide analyses are concerned, the computational burden of Bayesian algorithms may not be affordable. Therefore, most of the current comparative sequence analysis studies cannot afford the computation time required for a full Bayesian analysis. At least in the near future, it is unlikely that the Bayesian approach will be used in more than a small fraction of comparative genetic research. Undoubtedly, the continued exponential growth in computer power will raise the threshold on feasible dataset size and make Bayesian methods a more relevant alternative.

2 Research outline: investigating the mutual dependency of alignment and phylogeny

In my studies I have investigated the existing methodologies for MSA and phylogeny reconstruction, and the mutual dependency between the two problems. This investigation led to the development of novel methodologies addressing the notorious error-proneness of both tasks. These methods were published in three leading peer-reviewed journals for the field of computational biology. This work that started from the development of improved phylogenetic tree reconstruction, followed by an investigation of the mutual dependency of phylogeny and alignment, which finally led to the development of a novel measure for alignment confidence based on the effect of the guide-tree in progressive sequence alignment.

2.1 Iterative phylogeny reconstruction

Chapter 3 (presented in *ECCB* 2006 and published in *Bioinformatics*) describes a hybrid methodology integrating advance probabilistic evolutionary models into distance-based methods for phylogeny reconstruction. The hybrid combines the accuracy of probabilistic approach with the efficiency of distance-based methods. Distance-based methods rely on evolutionary distance estimation and are sensitive to errors in such estimations. In this study, an advanced evolutionary model that accounts for among-site rate variation, which was proven of being superior for the purpose of phylogeny reconstruction (Tateno, Takezaki, and Nei 1994), was integrated into the estimation of evolutionary distances. Rate variation is estimated within a Bayesian framework by extracting information from the entire dataset of sequences,

unlike classical distance-based methods that can only use one pair of sequences at a time. The accuracy of a cascade of distance estimation methods was evaluated, starting from commonly used methods and moving towards the more sophisticated novel methods. A significant improvement in the accuracy of distance estimation by the novel method over the commonly used ones was demonstrated using both real and simulated protein sequence alignments. An implementation of this method is freely distributed as part of the open-source SEMPHY package.

The strengths of the hybrid method facilitated two applied collaborative projects, in bacterial phylogeny and the protein phosphatase 2C superfamily. The first was an investigation into the positioning of thermophilic bacteria at the root of the tree of life (Section 3.4) which I presented in the *Annual Meeting of the Society of Molecular Biology and Evolution* (*SMBE*). The second was the investigation of the phylogeny of the protein phosphatase 2C superfamily (Appendix A), which was published in the *Journal of Molecular Evolution.*

2.2 Iterative phylogeny and alignment

Having an improved phylogenetic accuracy led me to investigate the relationship between the accuracies of tree and alignment. I investigated the hypothesis that improved accuracy of the guide tree will improve the accuracy of the subsequent MSA. I was successful in demonstrating a significant effect of guide tree accuracy in certain scenarios. Previous studies have concluded that the guide tree bears no statistically significant effect on the accuracy of MSAs (e.g., Nelesen et al. 2008). However, these studies were limited to alignments of up to 100

sequences. Since the complexity of the MSA and the potential for error increases with the number of sequences, I hypothesized that the guide tree becomes more important for larger alignments of hundreds of sequences. Indeed, this effect was significant in protein datasets from 400 to 1,000 sequences. Unfortunately, this result was published at the time by Liu *et al.* (2009) before my manuscript was completed. My unpublished results on this effect are described in Chapter 4.

2.3 Relay of uncertainty

An intriguing observation from this study led to the next two chapters of this thesis (recently published in *Molecular Biology and Evolution* and *Nucleic Acid Research*). Despite failure to improve the accuracy score of the alignment in small datasets, I observed that changes in the guide tree gave rise to dramatic changes in the MSA, which were not reflected in significant changes in the accuracy score. I hypothesized that these changes reflect uncertainty in the alignment rather than accuracy improvement. Chapter 45 demonstrates that this is indeed the case – that uncertainties in the guide tree are a major source of uncertainty in the alignment. I developed this insight into a scoring method that assigns confidence scores to each position of the MSA. Simulated and real protein benchmarks were used to show that these scores accurately identify the large majority of alignment errors.

The ability to identify the badly aligned parts of an MSA is a valuable tool for a wide range of comparative sequence studies, wherein such errors may lead to artifacts in the downstream analysis. Thus, Chapter 6 describes the implementation of this method in GUIDANCE, a user-

friendly web server, which is easily accessible for the wide molecular biology community. The utility of GUIDANCE was demonstrated by re-analyzing the heavily studied Vpu protein of HIV-1, an example for a fast evolving sequence that is challenging for alignment. GUIDANCE enables identification of unreliable alignment regions, filtering of un-alignable sequences, and subsequent dramatic reduction in alignment uncertainty. Together, these last two chapters present a novel methodology, previously unavailable for the scientific community, that is expected to empower better usage of alignments in a wide range of comparative sequence studies.

3 Phylogeny reconstruction: increasing the accuracy of pairwise distance estimation using Bayesian inference of evolutionary rates

This chapter is based on a published manuscript:
Ninio, M.*, Privman, E.*, Pupko, T., and Friedman, N. 2007. *Bioinformatics* 23: e136-e141.
***** Equal contribution

3.1 Introduction

Open questions in phylogenetics include reconstruction of early speciation events in the mammalian species tree (e.g., Steppan, Storz, and Hoffmann 2004; Murphy et al. 2007) or more ancient events as far back as the root of the tree of life (e.g., Brown and Doolittle 1995; Forterre and Philippe 1999). Other phylogenies describe series of duplications of gene families, e.g., the hemoglobin family (Hardison and Miller 1993) or the kinase super-family that includes more than 500 putative kinases in the human genome (Manning et al. 2002). Reconstruction of such phylogenies and resolving the many duplication events may shed light on the functional specialization in sub-families of genes.

The size of the analyzed dataset may vary significantly, both in the length of the sequences and in their number. Species trees may be based on the sequencing of a single protein of several hundreds of amino acids, or on several genes, or even on entire genomes. A dataset that includes paralogous sequences may reach large numbers of sequences. For example, a study of "kinome" evolution in fully sequenced mammalian genomes will encompass thousands of sequences. This range of practical situations should be considered when phylogeny reconstruction methods are discussed.

Different approaches to phylogeny have their own strengths and weaknesses, as reviewed in Section 1.3 above. ML and Bayesian methods have been argued to be superior in terms of accuracy and statistical justification but they become computationally infeasible when dealing with large datasets because the tree search space, i.e., the number of possible trees, grows exponentially with the number of sequences (Felsenstein 2004). This problem becomes increasingly aggravating with the rapid accumulation of molecular sequence data. In many molecular studies the subject gene or gene family of interest may lead to hundreds and even thousands of homologous sequences in the databases. Concomitantly, the field of molecular evolution has produced increasingly sophisticated methods for phylogenetic analysis, which are more computationally intensive. These combined advances challenge computational feasibility of contemporary studies of molecular evolution.

Contrary to probabilistic methods, the efficiency of distance-based methods is polynomial in terms of the number of sequences (see Section 1.3). This advantage in computation time makes them essential for dealing with large datasets. The importance of distance methods is not only as a faster, less accurate alternative to ML methods, but also in providing a good starting point of a heuristic search for the ML tree (e.g., Friedman et al. 2002; Guindon and Gascuel 2003). Clearly, if the distance method could be improved then the ML search could be faster, and give better results.

Distance-based methods are made up of two steps:

1) Pairwise distance estimation between all possible pairs of sequences in the dataset
2) Tree reconstruction based on the distances (this stage is blind to the original sequences)

These are two modular stages - any method for distance estimation can be used with any distance-based method for tree reconstruction. While several distance-based tree reconstruction methods have been developed, the initial step of distance estimation received scant attention. Indeed, the simplistic Jukes-Cantor (JC) method (1969) is still common practice for distance estimation, in spite of its oversimplifying assumption that all types of substitutions have equal probabilities. Great efforts have been invested in improved modeling of sequence evolution for use with ML methods. These advanced probabilistic models should also be used for distance estimation. As the following study demonstrates, previously published distance methods are still inadequate in terms of both error and bias.

We present a novel approach to distance estimation with increased accuracy, thereby improving phylogeny reconstruction. Our method is an adaptation of advanced probabilistic models from the ML paradigm to a distance-based approach, making their application to many thousands of sequences computationally feasible. Thus, the analysis of large datasets may now benefit from the improved accuracy of these refined models. The key idea underlying our hybrid methodology is to estimate model parameters from the entire dataset, and to use them

in each pairwise distance estimation. We show that the novel method significantly improves the accuracy of phylogenetic tree reconstruction.

3.2 Methods for maximum likelihood distance estimation

The evolutionary distance *d* between a pair of sequences is defined as the average number of substitutions per sequence site. This measure is related to the time that passed and the rate of substitutions. ML methodology may be used for distance estimation in a similar fashion to finding the ML tree: The maximum likelihood estimate (MLE) for the distance ($\hat{d}$) between two sequences (*A* and *B*) is the distance that maximizes the likelihood of the distance, which is the conditional probability of the sequence data, given a distance *d* and a model of sequence evolution *M* (Zharkikh 1994):

$$\hat{d} = \mathrm{argmax_d}\, \mathrm{L}(d) = \mathrm{argmax_d}(A, B|d, M) \qquad 3\text{-}1$$

The model *M* is a probabilistic model that describes the evolution of biological sequences. All the distance estimation methods discussed in this chapter will be presented as ML estimates under a specific evolutionary model. These models range from the simple JC model to complex models that strive to capture the nature of evolution of protein-coding genes as accurately as possible. This section describes the cascade of ML methods of increasing complexity, culminating in the novel methods we propose.

3.2.1 Probabilistic models of sequence evolution

Generally, ML methods in phylogeny use continuous time Markov models (Karlin and Taylor 1975) that define for any pair of aligned characters *a* and *b* the probability $p_{a\to b}(d)$ of the substitution from *a* to *b* in an evolutionary distance *d*. Additionally, they define the initial character probabilities π_a. Under the simplifying assumption that sites evolve independently and for models that satisfy reversibility $\left(\pi_a p_{a\to b}(d) = \pi_b p_{b\to a}(d)\right)$ the likelihood of the distance can be written as: (Felsenstein 2004)

$$\mathrm{L}(d) = \mathrm{P}(A, B|d) = \prod_{i=1}^{S} \pi_a p_{a_i\to b_i}(d) \qquad 3\text{-}2$$

Where a_i and b_i are the *i*-th pair of aligned residues out of a total *S* positions in the sequence alignment.

The simplest model possible is the JC model that assigns equal probabilities for all types of substitutions (Jukes and Cantor 1969). This over-simplifying assumption was subsequently relieved by models that allow variable $p_{a\to b}(d)$ probabilities. Such models define a rate matrix *Q* where $[Q]_{(a,b)}$ is the rate of replacements from character *a* to *b*. This matrix is used to calculate the $p_{a\to b}(d)$ probabilities: (Durbin et al. 1998)

$$p_{a\to b}(d) = [e^{dQ}]_{(a,b)} \qquad 3\text{-}3$$

The exponent of the matrix *dQ* is usually computed using the eigen vector decomposition of *Q*. This likelihood function is maximized according to Equation 3-1 in order to find the ML distance.

Such models have been initially designed for nucleotides (e.g., Kimura 1980; Yang 1994b). For amino acids, the larger alphabet size (20 instead of 4) requires a significantly larger number of parameters in the model. Therefore, empirical replacement matrices were calculated using large protein datasets. In this work we concentrate on amino acid sequences, for which the computational challenge is greater, although our novel methods can be equally applied to DNA sequences. Specifically, we use the JTT matrix (Jones, Taylor, and Thornton 1992).

The most significant oversight of this model, which is used by current distance-based phylogeny methods, is the assumption of equal replacement rates at all sequence sites. In this chapter, we shall refer to the method that uses this model the *homogeneous rates* method. However, evolutionary rates vary considerably among sites, due to non-uniform selection forces (Yang 1996).

3.2.2 Among site rate variation

Models that explicitly take into account among-site rate variation (ASRV) were shown to be statistically superior to the homogeneous models (Yang 1994a) in ML phylogeny reconstruction. ASRV is modeled by assuming that each site *i* in the sequence has a different rate, r_i, relative to the average rate over all sites. Thus, a site of rate 2 evolves twice as fast as

the average. This is equivalent to multiplying the distance by the rate in the likelihood calculation for each site:

$$\mathrm{L}(d) = \mathrm{P}(A, B|r, d) = \prod_{i=1}^{S} \pi_a p_{a_i \to b_i}(d \cdot r_i) \qquad 3\text{-}4$$

This equation assumes that rates are known. Since this is not the case, two approaches can be taken. One approach is to estimate the rate r_i similarly to the distance, under the ML principle. However, a Bayesian approach gives better performance (Mayrose et al. 2004). A prior distribution of rates $R(r)$ is assumed. The likelihood is then computed by averaging over all possible rates:

$$\mathrm{L}(d) = \mathrm{P}(A, B|R, d) = \prod_{i=1}^{S} \int_{r=0}^{\infty} R(r) \pi_a p_{a_i \to b_i}(d \cdot r) dr \qquad 3\text{-}5$$

The most common choice for $R(r)$ is the gamma distribution with the mean set to one (Yang 1993). Such gamma density function has one free parameter α that allows for different distribution shapes. The distance and the α parameter can be estimated simultaneously for each pair of sequences, using ML. We shall refer to this method as the *pairwise α* method. For lack of analytic solution, a discrete approximation of the gamma distribution is commonly used (Yang 1994a). Here we use 32 discrete, equal-probability bins, whose means are calculated based on the value of α.

3.2.3 Iterative inference of model parameters

The *pairwise α* method estimates the α parameter for each pair of sequences independently. However, the variability of rates in a protein is generally common to all sequences across a given MSA. Thus, there is no reason to estimate the rate parameters for each pair of sequences. Moreover, such estimation of many parameters from scant data is likely to result in high errors (Figure 3.1a). It would be preferable to use all sequences in order to estimate the rate parameters globally. However, such estimation requires knowledge of the phylogenetic tree, which we have not yet reconstructed. This kind of circular situation calls for an iterative process of optimization. Sullivan et al. (2005) studied iterative parameter optimization in the context of ML tree search. Here we suggest a similar approach for distance-based tree reconstruction. The iterative scheme allows for an "ML-NJ" hybrid, consisting of an ML optimization stage and an NJ tree reconstruction stage (or any other distance-based method). In each iteration, global ASRV information is extracted from the entire MSA using the tree reconstructed in the previous iteration. This "global information" is then used to re-estimate the pairwise distances more accurately, and then re-build the tree (Figure 3.1b).

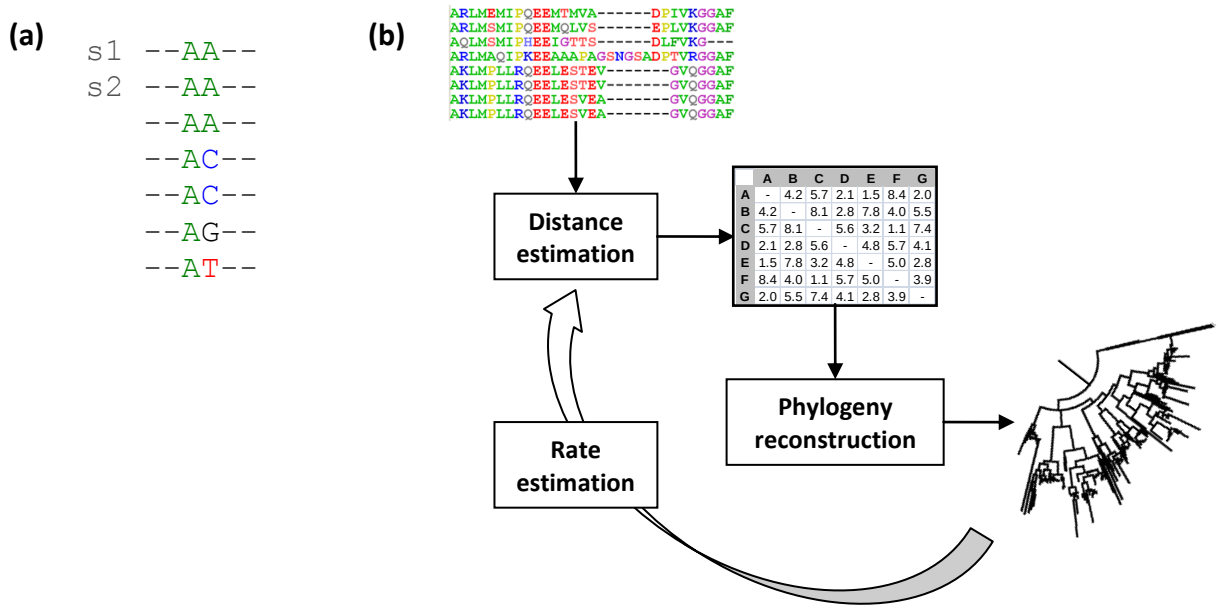

Figure 3.1: Utilizing the entire MSA to estimate the variation of rates among sites. (a) When looking only at the first two sequences of this simple example, both sites of the alignment are identical and there is no reason to think that they evolve with different rates. However, when all seven sequences are used, we can deduce that the rate of the second site is larger than that of the first one. **(b)** The proposed iterative approach that utilizes rate information from all sequences to improve distance estimation.

We propose three alternatives for the iterative estimation of ASRV parameters:

- *Iterative α*: Initial pairwise distances are estimated using the *homogeneous rates* method, and a tree is reconstructed. This tree is used to infer α, which is then used to improve the estimation of the pairwise distances. These iterations are repeated until the likelihood converges.

- *Iterative rates*: This method uses the evolutionary rate at each position as the "global information". The MLEs of these rates are iteratively estimated, and then used to recalculate the distances by maximizing Equation 3-4. This method captures more information about the ASRV than the *iterative α* method.
- *Iterative posterior*: A posterior rates distribution is estimated for each site rather than relying on a single estimate of the ML rate. This distribution is then used in Equation 3-5 instead of the prior distribution $R(r)$. In the discrete approximation that is used here the posterior probability of each rate category is calculated for each site.

3.3 Evaluation of the distance estimation methods

The performance of the different methods was evaluated in three comparative studies. The results presented here are for the five methods summarized in

Table **3.1**.

Table 3.1: The distance estimation methods used in the evaluation studies

Name	Evolutionary model
Homogeneous rates	No rate variation
Iterative α	Independent estimation of α for each sequence pair
Iterative α	Global estimation of α
Iterative rates	Global estimation of the ML rate at each site
Iterative posterior	Global estimation of the posterior distribution of the rate at each site

3.3.1 Reconstructing trees from protein sequence alignments

The ultimate goal of improving distance estimation is to increase the accuracy of the reconstructed tree topology. Therefore, the accuracy of reconstruction using the novel methods was evaluated both for real and simulated protein sequences. We used the NJ method for tree reconstruction (Saitou and Nei 1987), which is the most popular distance-based method (see Section 1.3), although our novel distance estimation methods can be equally used with any distance-base method.

Table 3.2: Tree reconstruction using different distance estimation methods

	Pairwise α	Iterative rates	Iterative α	Iterative posterior
ΔLL per position†	-0.0655	+0.0151	+0.0077	+0.0177
Improved topology‡	7%	31%	32%	44%

† The average difference in the log-likelihood per position scores compared to the *homogeneous rates* method.
‡ The proportion of trees for which there was a difference in the topology and an improved likelihood compared to homogeneous rates.

We used a dataset of 84 protein MSAs that was composed by Aloy et al. (2001). For each MSA, trees were reconstructed by the hybrid ML-NJ, using each of the five different distance methods. We compared the trees in terms of their log-likelihood scores under the gamma ASRV model. Such comparison might be affected by differences in branch length estimation under the different models. Therefore, branch lengths and α optimization was performed on the fixed tree topologies that were constructed by NJ. Each log-likelihood score was divided by the length of the MSA to produce the average log-likelihood score per position.

Table 3.2 lists the differences between the score of each method and the score of the *homogeneous rates* method, which is used as a reference. The second line indicates the percentage of MSAs for which there was a difference in the tree topology that resulted in improved likelihood, compared to the *homogeneous rates* method. Compared with this reference, the *pairwise α* method produces trees of lower likelihood. On the other hand, all three iterative methods improve the average likelihood scores. The *iterative posterior* method achieved the best results, with an average improvement of 0.0177 log-likelihood points per position and an improved topology for 44% of the MSAs. We used simulation studies to further investigate this pattern.

3.3.2 Reconstructing trees from simulated multiple sequence alignments

Accuracy of tree reconstruction from real protein sequences can only be compared in terms of the likelihood of the trees, since the true phylogeny is not known. For this reason we applied the different methods to protein MSAs that were simulated according to a known tree, and evaluated their accuracy by comparing the reconstructed tree to the original "true" tree. We used ten trees that were reconstructed by the homogeneous rates method in the previous section as the basis for the simulated MSAs. Thus, these simulations represent several tree topologies of real phylogenies. We chose MSAs with a number of sequences around 50.

The gamma-ASRV model was used to simulate sequence evolution according to those tree topologies. The simulations were repeated for ten values of α: 0.1 (highly variable rates), 0.2, 0.5, 0.7, 1.0, 1.3, 1.6, 2.0, 2.5 (relatively homogeneous rates). For each α, a vector of 1000

rates was sampled from the gamma distribution. Each of the ten trees was used with each of the ten rate vectors to simulate an MSA of 1000 columns. This procedure was repeated ten times, resulting in ten MSAs for each tree and for each α value, a total of 1,000 simulated MSAs. Each distance method was used (in the context of the ML-NJ hybrid) to reconstruct a tree from each MSA and the resulting trees were compared to the original "true" tree that was used to simulate the MSA.

The performance of the five methods was evaluated in terms of log-likelihood scores (as above) and in terms of the topological distance between the inferred and the original tree. The later is measured by the percentage of splits or branches that both trees agree on, known as the Robinson-Foulds distance (Robinson and Foulds 1979, see Section 1.6).

Figure 3.2 plots these two accuracy measures as a function of the α value that was used in the simulations. Both scoring measures agreed on the ranking of the five methods: *iterative posterior* > *iterative rates* > *iterative α* > *homogeneous rates* > *pairwise α*. Paired *t*-tests indicate that these differences are highly significant (p-value < 10^{-5} for all comparisons).

The results for the simulated MSAs agree with the pattern that was observed for the real protein sequences. The differences in the log-likelihood per position are also comparable. An interesting observation is that *pairwise α* performs especially badly for simulations with extreme values of α, and is therefore worse than *homogeneous rates*. This is probably the result of large errors in the α estimates, which are based on two sequences only.

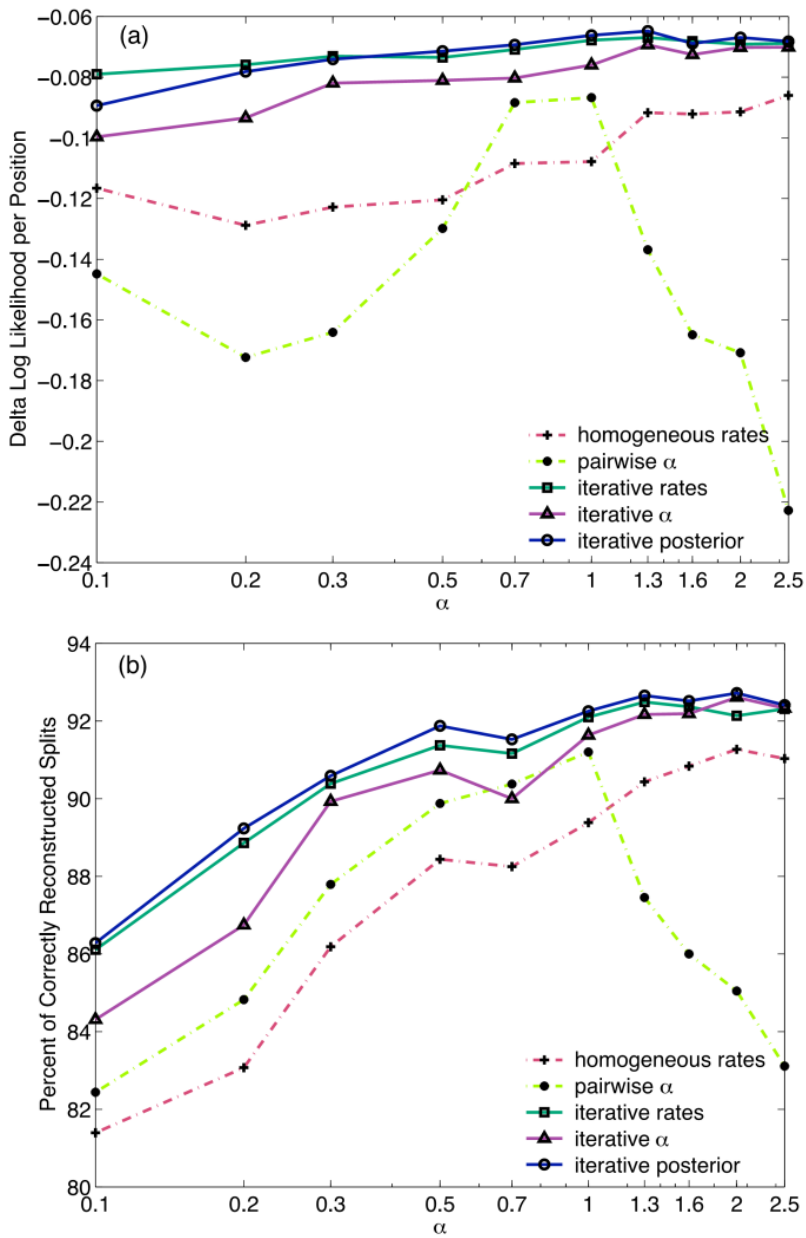

Figure 3.2: Accuracy of tree reconstruction using different distance estimation methods. Accuracy is plotted vs. the α value that was used in the simulations (in log-scale). **(a)** The difference in the log-likelihood per position of the reconstructed tree, compared to the true tree. **(b)** The percentage of split agreement with the true tree.

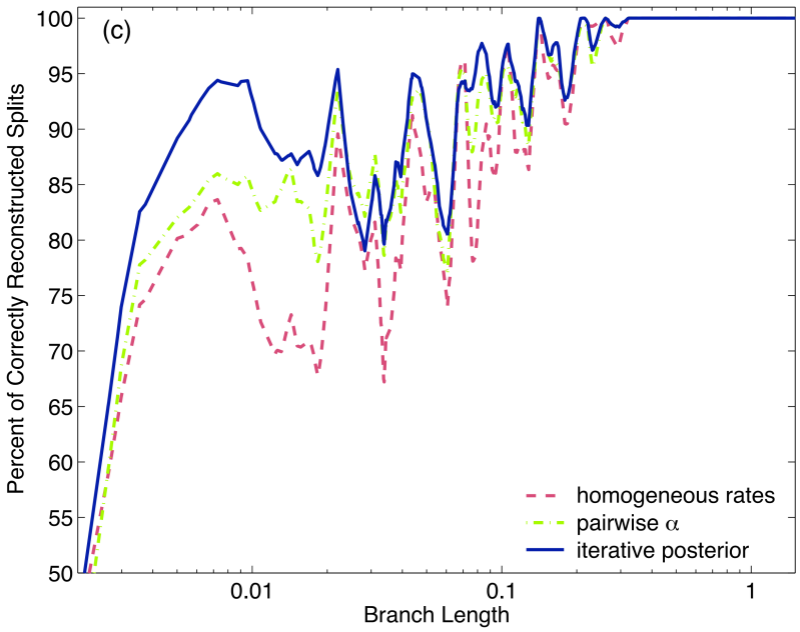

Figure 3.3: Percentage of correctly reconstructed splits vs. the corresponding branch length. The curves were created using the LOWESS function (locally weighted scatter plot smooth) in MATLAB.

Compared to the commonly used *homogeneous rates* method, the *iterative posterior* method improves the log-likelihood score by 0.02-0.05 points per position, depending on α. In terms of the topological accuracy of the tree, the percentage of correctly reconstructed splits is improved by 2-6%, depending on α. A larger improvement is evident for α values less than 1. Not surprisingly, this result shows that the novel method will be especially significant for proteins with large rate heterogeneity. The improvement in correct split reconstruction is usually very valuable, as we observed that many of the longer branches are easily reconstructed with any distance estimation method, and a relatively small number of short

branches is commonly the more challenging part of the phylogeny. This pattern is plotted in Figure 3.3. The largest impact is on branch lengths around 0.01, where the proportion of correctly reconstructed splits is improved by 20%.

3.3.3 Evaluation of the accuracy of distance estimation on pairs of sequences

The evaluation of tree reconstruction above clearly shows the superiority of the iterative methods. However, it is interesting to understand how is the improvement in the accuracy affected by different factors, such as the pairwise distances and the α parameter. For example, improvements in the accuracy for relatively distant pairs of sequences might be more significant than for close pairs. In addition, the different methods may vary in the extent of their bias in distance estimation. Therefore, we used simulations of pairs of sequences to study the effects of these factors. We investigated the error and the bias by comparing the estimated distance with the original distance that was used in the simulation.

The same protocol that was used to simulate MSAs was adapted to simulate pairs of sequences 1,000 amino acids long. One thousand pairs were simulated for each combination of the ten different α values and ten different evolutionary distances between 0.01 and 1.5. In total, 100,000 pairs of sequences were simulated. For the iterative methods we used the previously simulated MSAs in order to estimate the required "global information", i.e., the global α parameter, and for each site - the ML rate and the posterior distribution of the rate. For each pair we used an MSA that was simulated with the same rates vector, so that the new sequence pair can be treated as though it belongs to the same dataset.

The accuracy of the five distance estimation methods was evaluated on these simulations. In addition, a sixth method (labeled *true rates*) was added as a frame of reference. This method is similar to the *iterative rates* method, except it was given the true rates that were used to simulate the sequences instead of the MLEs of the rates. This information is obviously not available for real proteins. It is used here in order to demonstrate the limit of the accuracy of this class of ML methods, when given the most accurate "global information" possible.

The results were analyzed in terms of the error and the bias in distance estimation. The relative mean square error (RMSE) and the relative mean error (RME) were used to measure the error and the bias respectively:

$$RMSE = Avg\left(\left(\frac{\hat{d}-d_{true}}{d_{true}}\right)^2\right) \qquad RME = Avg\left(\frac{\hat{d}-d_{true}}{d_{true}}\right) \qquad 3\text{-}6$$

3.3.3.1 Accuracy as a function of the evolutionary distance

In Figure 3.4 the RMSE and RME of each method are plotted as a function of the true distance by which the sequence pairs were simulated. The results are shown for simulations with an α value of 0.7.

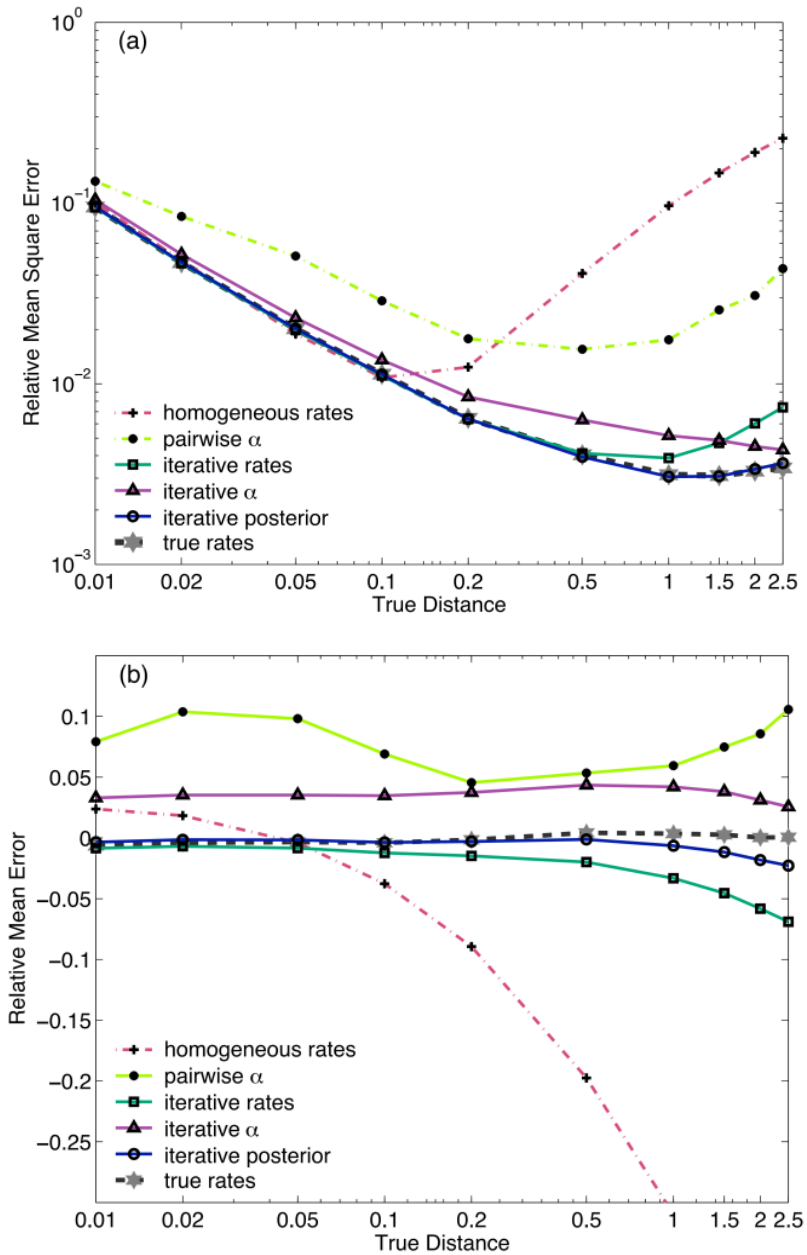

Figure 3.4: Error and bias of different distance estimation methods as a function of the true distance. Sequences were simulated with α=0.7. Each data point is an average based on 1,000 independent sequence pairs. **(a)** RMSE as a measure of the error. **(b)** RME as a measure of the bias.

The improved accuracy of the novel iterative methods is evident from Figure 3.4a, mainly for large distances. It seems that only for large distances, where many sites undergo multiple replacements, there is a significant advantage to the more refined models. For small distances most methods produce very similar errors. For distances larger than 0.2 all the ASRV methods are significantly more accurate than the *homogeneous rates* method. The major contributing factor to the inaccuracy of *homogeneous rates* is probably its considerable bias for underestimation (Figure 3.4b), which increases dramatically with the distance.

Among the ASRV methods, the iterative methods that use "global information" are significantly more accurate than the *pairwise α* method that does not. Again, we attribute this result to the insufficiency of the information in two sequences for accurate estimation of ASRV parameters. Interestingly, there is a noticeable bias for overestimation (over 10 percent) in the *pairwise α* method, for both small and very large distances. The iterative methods, on the other hand, do not display a significant bias. The *iterative posterior* method seems to be especially unbiased.

The accuracy of all methods never exceeds that of the *true rates* method, as expected for the optimal "global information". Surprisingly, even for very large distances, the three iterative methods produce RMSE values that are no more than 1.5 times larger than those of the *true rates* reference. In general, the *iterative posterior* method is more accurate than the other two methods. Its advantage is especially noticeable for large distances, where its error is almost equal to the gold standard set by *true rates*.

It is worthwhile to note the effect of the improved accuracy of pairwise distances on the successful reconstruction of tree topology. The significant improvement in distance accuracy was for distant pairs (distances larger than 0.2), while the improved reconstruction was mainly in the shortest branches of the trees (of length around 0.01, Figure 3.3). Evidently, the accurate estimation of large pairwise distances is essential for resolving difficult splits that correspond to short branches. This effect is probably due to distant pairs of sequences connected by a path in the tree that includes very short branches. The large pairwise distances are used by NJ to resolve those internal branches.

3.3.3.2 Accuracy as a function of α

When ASRV models are applied to protein sequences the estimated α values typically range between 0.5 and 3.0. In order to test the effect of the degree of rate variation on the accuracy of the distance estimation methods we plotted the error and the bias against α. Figure 3.5 presents the results for a distance of 1.0, which is large but not uncommon. At most of the biologically relevant α values the three iterative methods are clearly more accurate than the simpler methods. However, at α values of 0.5 and smaller the *iterative posterior* and the *iterative rates* methods become less accurate, while the *iterative α* method remains nearly as accurate as the *true rates* reference.

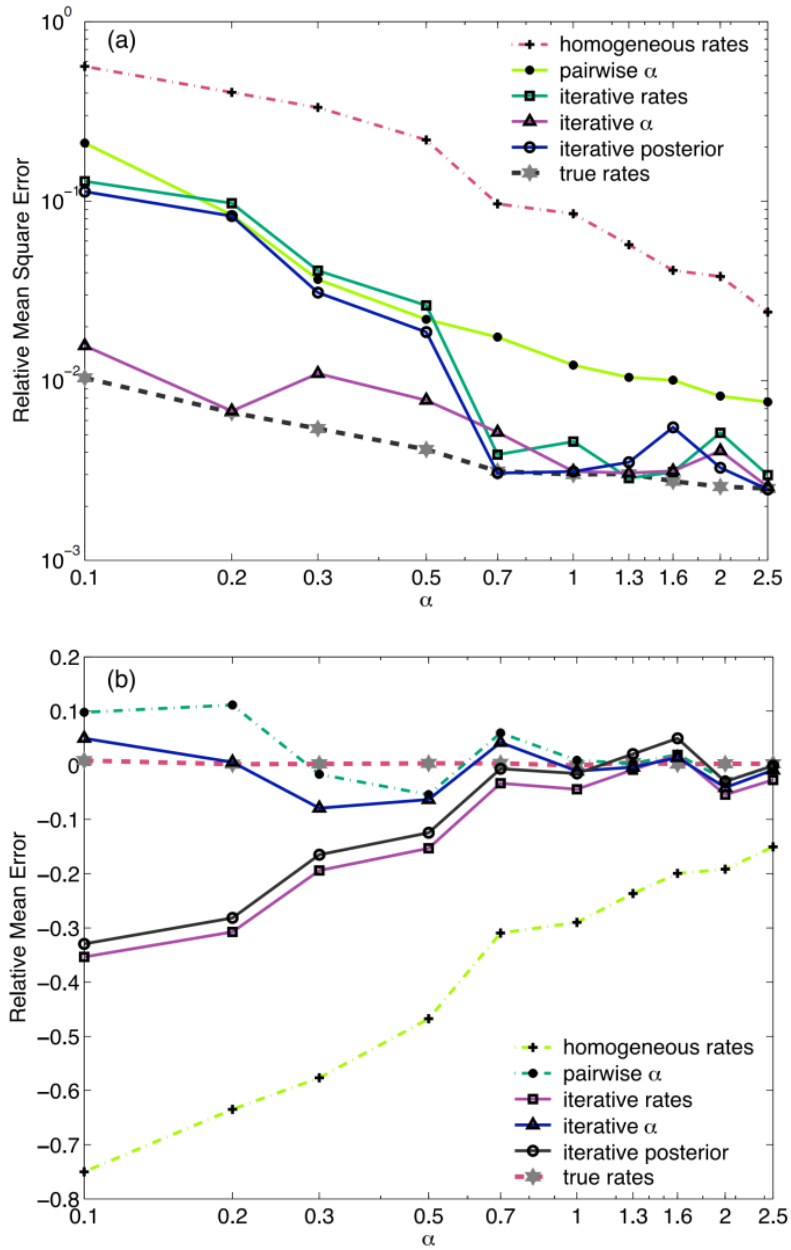

Figure 3.5: Error and bias of different distance estimation methods as a function of α. Sequences were simulated with a pairwise distance of 1.0. Each data point is an average based on 1,000 independent sequence pairs. **(a)** RMSE as a measure of the error. **(b)** RME as a measure of the bias.

This increased error is correlated with a bias for underestimation (Figure 3.5b). We investigated the cause of this bias, finding that it was preceded by underestimation in the branch lengths of the trees that were reconstructed from the simulated MSAs. The bias of the ML estimation of the branch lengths at small α values was never reported before. This is an interesting and important result in itself, which merits further investigation, as it surely affects any other evolutionary analysis that makes use of the branch lengths of trees. In our analysis, the shortening of the branch lengths resulted in overestimation of the rate at each site, which caused underestimation of distances by *iterative rates* and *iterative posterior*. Nevertheless, the novel methods we present here produce high accuracy in all evolutionary scenarios except for the very extreme end of the rate variability in biological protein sequences.

3.4 Modularity of the hybrid approach – application to bacterial phylogeny

This section is a collaboration that was presented as a contributed talk in:
Privman, E., Dutheil, J., Ninio, M., Friedman, N., Galtier, N., and Pupko, T. *The Annual Meeting of the Society for Molecular Biology and Evolution (SMBE)* 2007. Halifax, Canada.

A valuable property of the hybrid ML-NJ is modularity. Two modules in the algorithm can be freely substituted with other counterparts: Here, standard NJ is used as the distance-based tree reconstruction module, but it can be replaced by any of the other NJ variants and other algorithms that build a tree based on a distance matrix (see Section 1.3). The second interchangeable part is the evolutionary model that is used in ML estimation of distances. The above results show that the ASRV model improves distance estimation accuracy. Similarly, other enhancements of the model realism may be integrated into the ML-NJ hybrid. Any

model parameters may be estimated in the ML stage of the iteration, where the rate parameters are estimated, and then subsequently used in the distance estimation stage (Figure 3.1b above). To demonstrate this potential of the hybrid scheme, this section describes the reconstruction of the bacterial phylogeny using a more advanced evolutionary model. I will first present the phylogenetic question at hand and then describe how the ML-NJ hybrid can help to address it.

3.4.1 Was the first living cell a thermophile?

The origin of life and the last universal common ancestor (LUCA) are topics of debate that remain continuously active since the days of Charles Darwin. It has been hypothesized that the first living cells developed in high temperature settings, near volcanic activity. These niches harbor thermophilic species of bacteria and archaea. This hypothesis requires that thermophilic species should be found at the root of the tree of life, as was the result of early reconstructions of the tree based on ribosomal RNA (rRNA) sequences (e.g., Woese 1987). However, more recent studies using a range of phylogenetic approaches, from supertree methods to estimation of the GC content of ancestral sequences (Galtier, Tourasse, and Gouy 1999; Daubin, Gouy, and Perriere 2001; Galtier 2001; Brochier and Philippe 2002; Daubin, Gouy, and Perriere 2002) found evidence that the LUCA was a mesophile, living in moderate temperatures. Several of these authors (including Galtier, Daubin, and their colleagues) argue that their phylogenetic innovations overcome biases of less sophisticated methods, which erroneously place the thermophiles at the root of the tree. This placing may be an artifact

resulting from the "long branch attraction" phenomena (Felsenstein 1978), which is expected for the highly diverged rRNA sequences (and the rest of the genome) of the thermophiles.

Sophisticated, more realistic models of evolution were suggested for improved accuracy of the tree of life. One such model enhancement allows the rate of a specific site to vary along the tree (in addition to allowing the rate to vary among sites). These are known as covarion-like models or models of site-specific rate variation (SSRV). SSRV models were suggested for the purposes of deep phylogeny reconstruction (Germot and Philippe 1999; Lopez, Forterre, and Philippe 1999; Philippe et al. 2000) because the dramatic changes in the rRNA sequences of mesophile vs. thermophile species appear to involve many shifts in the evolutionary rate of sequence sites. Although these models have been applied to ancient phylogenies (e.g. Galtier 2001), they were used for estimating the GC content of LUCA and not for phylogeny reconstruction. The SSRV model was considered computationally complex, and Galtier (2001) limited himself to less than 40 rRNA sequences. Therefore the model was not used for an ML tree search that requires evaluations of many topologies. The challenge is then to utilize the SSRV model for efficient search for the ML tree of life, and to include as many sequences as possible in order to maximize the evolutionary information available.

3.4.2 Reconstruction of the bacterial phylogeny from rRNA sequences

We integrated the SSRV model into the hybrid ML-NJ method. The efficiency of the distance-based approach allows the analysis of a large sequence dataset with the SSRV model. A dataset of 861 bacterial sequences of the small subunit (SSU) rRNA was retrieved from the European

Ribosomal RNA Database (Wuyts, Perriere, and Van De Peer 2004). We included one representative from each bacterial genus that was sequenced to date. A high-quality structure-based alignment of these sequences was downloaded from the above database. Two archaeal sequences were included as an outgroup – *Aphrodite sulfophila* and *Acidolobus aceticus*.

The SSRV model with four discrete rate categories was used in combination with the Tamura (1992) substitution model. Each iteration of the ML-NJ hybrid (Figure 3.1b) took approximately 9 days on a 2.4GHz, 64bit Opteron processor, for the 861 rRNA sequences. The likelihood score converged quickly after two such iterations and the resulting phylogeny is presented in Figure 3.6. To produce bootstrap confidence scores (Section 1.4) bootstrap trees were reconstructed using a single iteration given the "global information" from the second iteration of the iterative run. Bootstrap runs were parallelized on a Linux cluster, but due to computational resource limitations only 31 bootstrap repeats were run.

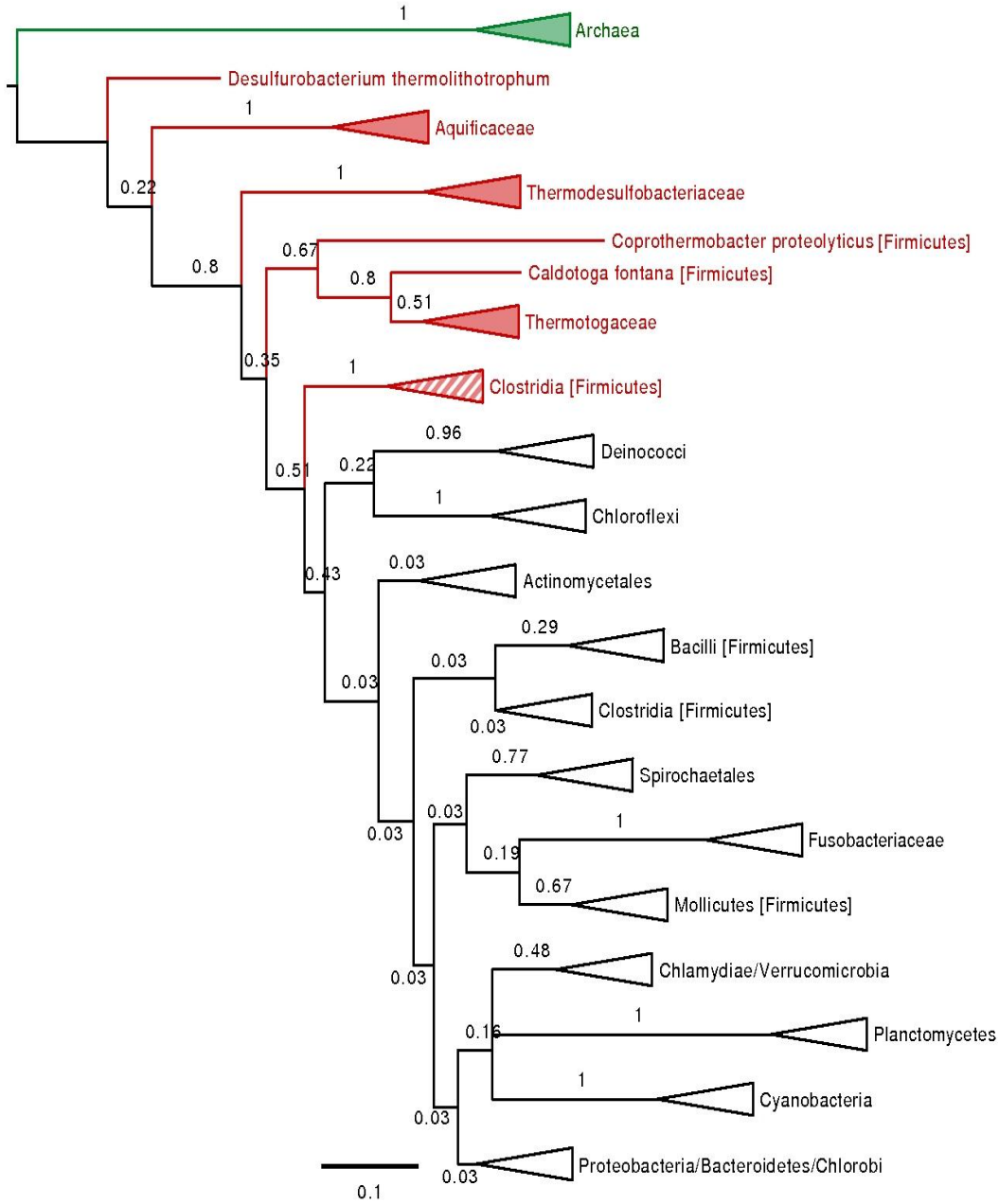

Figure 3.6: Phylogeny based on bacterial rRNA sequences, reconstructed using the SSRV model. Phyla are collapsed and shown as triangles. Thermophiles are colored in red. *Clostridia* are marked in red stripes because only some the genera are thermophilic. An outgroup of two archaeal sequences was used (colored in green). The *Firmicutes* phylum is not a monophyletic group in this tree and therefore its classes are shown separately. Bootstrap scores presented are based on 31 bootstrap replicates.

The resulting tree is similar to a standard NJ tree (not shown) with respect to the positioning of the thermophiles at the root. Thus, the analysis using an SSRV model supports the hypothesis that the ancestor of the bacteria was a thermophile. The comparison between the NJ and SSRV-NJ trees reveals some differences in the positioning of other phyla, however, these branches have a low bootstrap support.

3.5 Summary

Current state-of-the-art distance-based phylogeny reconstruction methods neglect to take ASRV into consideration. Thus, such methods suffer from high errors and bias, as we show in our simulation studies. Our results also demonstrate that an attempt to estimate ASRV parameters for each pair of sequences independently will inevitably suffer from large errors. Therefore, we propose an iterative ML-NJ hybrid algorithm to extract more refined "global" ASRV information from the entire dataset, using the tree that was estimated in the previous iteration. While all previously suggested distance-based methods consider each pair of sequences separately, the iterative method makes use of all available sequences, allowing more accurate parameter estimation for the gamma-ASRV model. We use the "global information" for a novel Bayesian distance estimation method that integrates the posterior distribution of the rate at each site into the estimation of the distance.

We demonstrate the improved accuracy of the hybrid method through a comparative study of distance estimation methods and their use in NJ. The *iterative posterior* method produces trees of significantly improved likelihood for both real and simulated protein MSAs. The

simulations also show that this novel method correctly reconstructs a larger percentage of the branches of the true tree. Using simulations of sequence pairs we show that the "global information" that is available to the iterative method reduces errors and bias in distance estimation. Our simulations demonstrate that these advantages are considerable in almost all scenarios, and are increasingly significant for large evolutionary distances and for proteins of high rate variability.

Finally, the integration of the SSRV model demonstrates the modularity of the hybrid ML-NJ algorithm and its potential for reconstructing large phylogenies using realistic evolutionary models. The analysis of the 861 bacterial genera is the first application of the SSRV model for tree reconstruction based on a large bacterial dataset.

4 Iterative use of improved phylogeny reduces alignment errors

4.1 Introduction

The improved accuracy of tree reconstruction (Chapter 3 above) led me to investigate alignment accuracy. An accurate multiple sequence alignment (MSA) typically relies on a phylogenetic guide tree in the process of "progressive sequence alignment" (see Section 1.5 above). Accurate reconstruction of phylogenies, in turn, requires an MSA. This circular dependence is usually solved by iterative, alternating phylogeny reconstruction and sequence alignment. Most algorithms use a distance-based method to reconstruct an initial tree based on pairwise sequence alignments. The tree is used as a guide tree for progressive sequences alignment, producing the first MSA. The process is then repeated - the MSA is used to reconstruct the tree, and a second MSA is produced. Most implementations do only two such iterations (Section 1.5. E.g. Thompson, Higgins, and Gibson 1994; Notredame, Higgins, and Heringa 2000; Edgar 2004; Katoh et al. 2005).

The first tree is based on pairwise distances derived from pairwise alignments. The second tree is based on pairwise distances derived from the entire MSA, and is hence expected to be more accurate. Similarly, the first MSA may also suffer from errors introduced during the progressive alignment, because the erroneous tree is used to determine the order of addition of sequences to the MSA. Generally, the second tree is more accurate because it is based on a more accurate MSA, and the second MSA is more accurate because it is based on a more accurate guide tree. The accuracy may be further improved in subsequent iterations.

It can be expected that errors in the guide tree lead to errors in the MSA. However, previous investigations concluded that alignment accuracy is rather robust to errors in the tree (e.g., Nelesen et al. 2008). In agreement with that view, progressive alignment programs use the simplest and fastest available tree reconstruction methods, such as NJ or UPGMA, rather than more accurate and elaborate methods (see Section 1.3).

However, here we investigate the hypothesis that as the number of sequences increases, so does the impact of errors in phylogeny on the accuracy of the final MSA. The rationale behind this hypothesis is that larger trees inevitably harbor more errors, which may cause alignment errors and lead to cascading failures during the longer process progressive alignment of many sequences. This issue has never been tested when hundreds or thousands of sequences are analyzed. In smaller datasets (e.g., Nelesen et al. (2008) analyzed 25 to 100 sequences) the standard guide tree may be accurate enough, and any errors that may be introduced during the progressive alignment do not propagate as much as in larger datasets.

To test our hypothesis we analyzed alignment accuracy using either a simple NJ guide tree or a tree reconstructed by the hybrid distance-likelihood method described in Chapter 3, which significantly increases the guide tree accuracy. We used an iterative scheme of alternating phylogeny reconstruction and sequence alignment. We included the popular alignment algorithm CLUSTALW, as well as PRANK – a recent "phylogeny-aware" improvement of the classical progressive alignment algorithm, which makes better use of the guide tree in the inference of insertions and deletions. We used simulation studies to evaluate these methods.

Our results demonstrate that for datasets of hundreds of sequences or more, the use of more accurate phylogeny significantly improves alignment accuracy. This conclusion holds both for the conventional CLUSTALW and for the "phylogeny-aware" PRANK.

4.2 Methods

We chose relatively accurate alignment and phylogeny reconstruction methods that are still capable of analyzing large datasets efficiently (in this study, up to 1,000 sequences), which we integrated into an iterative scheme.

The phylogeny reconstruction method described in Chapter 3 above achieves high accuracy while retaining the ability to process thousands of sequences by combining the benefits of distance-based and probabilistic methods. In the present work we compare this hybrid ML-NJ method to simple NJ with respect to the accuracy of the resulting MSA. We integrated the hybrid method with progressive sequence alignment algorithms to test whether the improved guide trees results in improved alignment accuracy.

We chose two MSA algorithms: CLUSTALW (Thompson, Higgins, and Gibson 1994) and PRANK (Loytynoja and Goldman 2005). See Section 1.5 above for a description of these algorithms. Both implementations provide the option of an input guide tree, which is necessary for our iterative scheme. CLUSTALW was included because it is a well-established, widely-used program. PRANK, the "phylogeny-aware" alignment algorithm, was included because we

hypothesized that improved guide trees can be of greater assistance for an algorithm that makes more advanced use of the tree compared to previous progressive alignment methods.

4.2.1 Iterative alignment and phylogeny

Our iterative algorithm is based on the standard scheme illustrated in Figure 1.5, except that we use the ML-NJ hybrid tree reconstruction method instead of simple NJ. It comprises of two modular building blocks: one of the two alignment algorithms and the hybrid phylogeny algorithm. Since the tree reconstruction module can only work on an MSA, the first iteration uses a standard NJ, which is based on pairwise alignments, to build the first MSA. Therefore, the beginning of the iterative flow is equivalent to a standard run of the MSA program.

Following this, each iteration begins with tree reconstruction using the hybrid method, which is then used as the guide tree for the sequence alignment stage. The process can be run for a pre-determined number of iterations, or for as long as the alignment score improves (as is given by CLUSTALW). For our simulation benchmarks we ran four iterations, producing four MSAs in addition to the initial MSA that is based on the simple NJ guide tree.

4.2.2 Alignment benchmark data

All phylogeny and alignment methods were tested on a benchmark of simulated MSAs. We used SIMPROT (Pang et al. 2005) which simulates the evolution of protein sequences along a given phylogenetic tree, including amino-acid substitution events, insertions and deletions,

according to empirically determined statistical distributions. Simulations started with a sequence length of 300 amino acids.

In order to achieve the most realistic simulated data possible we used a tree that was reconstructed by PHYML (Guindon and Gascuel 2003) from a large alignment of 400 thymidylate synthase sequences. We trimmed this tree to produce datasets of variable sizes by randomly removing leaves to generate different size categories: 100, 200, 300, or 400 sequences. To produce even larger datasets, we duplicated the tree in three copies, connected them to a new root with a branch length of 0.1. We then repeated the random trimming process to produce datasets of 500, 600, 700, 800, 900, and 1,000 sequences. The random process of tree-trimming and SIMPROT sequence simulations was repeated 40 times for each size category (100 – 1000) to generate 40 independent datasets.

We measured the alignment accuracy for each MSA compared to the reference "true" MSA. We used the sum-of-pairs (SP) score that measures the percentage of pairs of aligned residues in the reconstructed MSA that agree with the reference MSA (see Section 1.6) as implemented in the bali_score program (http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/).

4.3 Results

4.3.1 CLUSTALW

We ran four iterations of the algorithm on simulated datasets ranging from 100 sequences to 1,000. Figure 4.1a plots the alignment accuracy of the MSA produced in each iteration for five

individual datasets of 400 sequences. Although some cases show reduced errors and some elevated errors, the general trend is improved accuracy. In Figure 4.1b we plot as a function of the number of sequences the average improvement of the SP score of the last MSA from the fourth iteration compared to the initial MSA (which is equivalent to a standard run of CLUSTALW). A significant improvement was achieved for all dataset sizes. Interestingly, these results do not show a positive correlation between the number of sequences and the accuracy improvement.

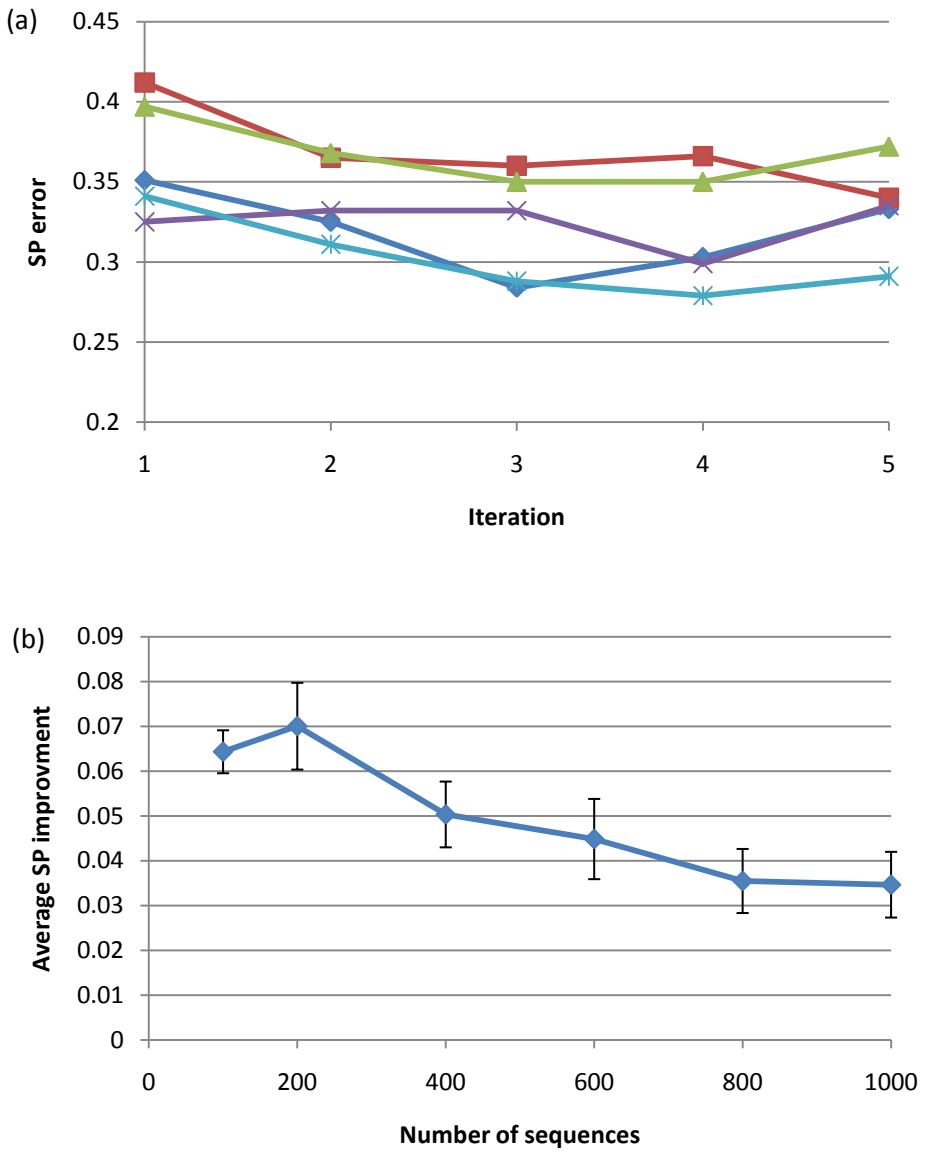

Figure 4.1: Reduced alignment errors following iterations with CLUSTALW. (a) Alignment error rates measured by the SP score of the MSA produced at the end of each iteration, for 10 independent datasets of 400 sequences. **(b)** The average improvement in SP score of the last MSA relative to the first one, as a function of the number of sequences. Each data point represents 40 independent datasets. The standard deviation is indicated by vertical bars.

4.3.2 PRANK

Next, we turned to test our iterative scheme with PRANK as the MSA algorithm. With the abovementioned datasets, no statistically significant improvement was achieved by the iterative algorithm compared to a standard PRANK run (data not shown). A possible reason for this failure is the fundamental difference in the "phylogeny aware gap placement" methodology of PRANK compared to classical implementations of progressive alignment, such as CLASTALW. The PRANK algorithm is aimed at aligning sequence positions that are genuinely homologous, i.e. a residue that existed in the common ancestor was propagated to present sequences via substitution mutations alone (Ari Loytynoja, personal communication). For example, two homologous sequences may contain an arginine in the same position, but these arginine residues were inserted in two independent events in the parallel lineages. Such a pair is analogous rather than homologous, and should not be aligned in a PRANK alignment. On the other hand, classical algorithms such as CLASTALW will tend to align such pairs.

PRANK was originally designed for aligning genomic DNA sequences, which are not very diverged. Conversely, distantly related protein sequences have undergone multiple consecutive insertion and deletion events and many pairs of similar amino-acid may have arisen through analogy, making the inference of true site-specific homology very difficult. Therefore Loytynoja himself argues that PRANK is not suitable for such datasets (personal communication and the PRANK website http://www.ebi.ac.uk/goldman-srv/prank/).

For this reason we were concerned that our simulations produced too-diverged homologs, outside the range of operation of the "phylogeny-aware" methodology and we decided to test PRANK on simulations of smaller evolutionary distances. We simulated datasets based on the same trees after multiplying all branch lengths by a scale factor of 0.1, 0.2, or 0.4, and ran our iterative algorithm as described above. For datasets of less than 400 sequences, there was no significant improvement compared to a standard PRANK run. However, on larger datasets we achieved significantly improved alignment accuracy.

Figure 4.2a plots the alignment accuracy of the MSA produced in each iteration for five individual datasets of 1,000 sequences that were produced after scaling the trees by 0.2. Again the general trend is reduction of SP errors. Figure 4.2b plots the average accuracy improvement as a function of dataset size, for each scaling factor. Here we find a strong correlation between the number of sequences and the accuracy improvement with scale factor 0.2 and 0.4 (Pearson correlation $r = 0.946$, $p = 0.001$, and $r = 0.958$, $p = 0.010$, respectively).

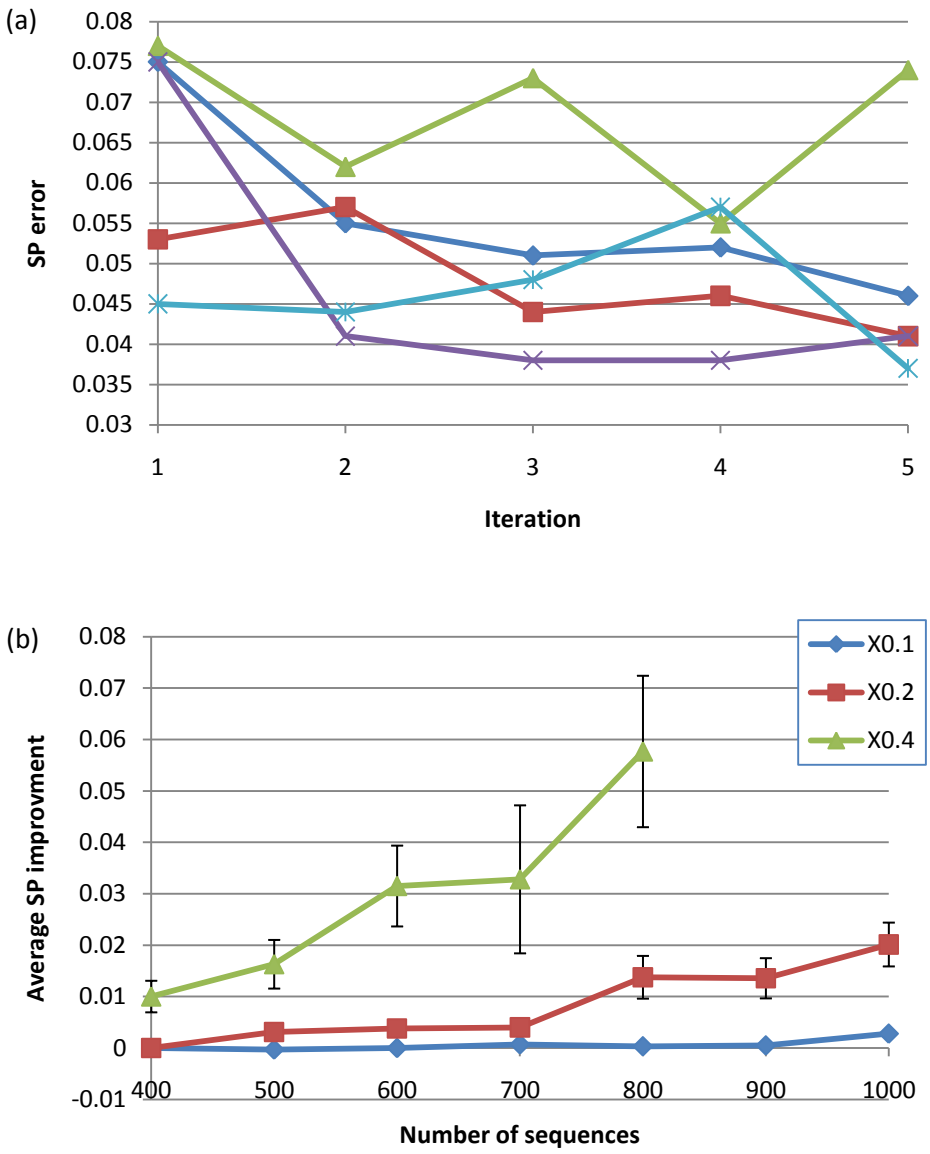

Figure 4.2: Reduced alignment errors following iterations with PRANK. (a) Alignment error rates measured by the SP score of the MSA produced at the end of each iteration, for five independent datasets of 1,000 sequences. **(b)** The average improvement in the SP score of the last MSA relative to the first one, as a function of the number of sequences. Each data point represents 10 independent datasets. The standard deviation is indicated by vertical bars.

4.4 Summary

This chapter presents an extension of the iterative phylogeny reconstruction scheme developed in Chapter 3. The contribution of improved topological accuracy was investigated within the widely used iterative scheme of phylogeny and alignment. All progressive alignment programs use a quick distance-based method to build their guide tree. Starting from the second iteration, an MSA is available for more sophisticated phylogenetic methods. The hybrid method described in Chapter 3 allows a more accurate yet efficient phylogeny reconstruction.

Previous studies concluded that alignment accuracy is rather robust to errors in the guide tree for datasets of 25 to 100 protein sequences (Nelesen et al. 2008). However, the simulation studies described here demonstrate that, for larger datasets, the improved phylogenetic accuracy leads to significant improvement in alignment accuracy. For diverged protein sequences aligned with CLUSTALW this effect was significant for all dataset sizes tested from 100 to 1,000 sequences. For less diverged sequences aligned with PRANK this effect was significant for 400 sequences or more. The magnitude of this effect in PRANK was correlated with both sequence number and degree of divergence (scaling of the tree).

Initially, we ran our iterative algorithm on real protein sequences from the widely-used BAliBASE benchmark (Thompson et al. 2005), which is based on structural alignments of real proteins (Section 1.6). However, such datasets usually consist of few sequences per MSA (no more than 25 sequences in the case of BAliBASE) and our hypothesis was that the effect of the tree is substantial only in large datasets of hundreds of sequences. In agreement with this

hypothesis, there was no change in the alignment accuracy by our iterative algorithm compared to a standard CLUSTALW run (data not shown). Since larger alignment benchmarks of real proteins are not available we turned to simulations of sequence evolution.

In parallel with our investigation, Liu et al. (2009) published similar results for simulated DNA sequences aligned in an iterative scheme combining the alignment algorithms MAFFT and MUSCLE with tree building using RAxML. See discussion in Section 7.1 regarding the comparison to RAxML. The general trends in that analysis are similar to the ones described here. They demonstrate significant improvement in alignment accuracy for datasets of 500 and 1,000 sequences, but not for datasets of 100 sequences.

To conclude, when approaching a challenge of comparative analysis of hundreds of sequences researchers will do better to invest effort in the application of more accurate phylogenetic methods than the simple distance-based methods commonly used for building guide trees. This conclusion complements the recent attention given to “phylogeny aware” progressive sequence alignment by Loytynoja and Goldman (2008).

5 An alignment confidence score capturing robustness to guide-tree uncertainty

This chapter is based on a published manuscript:
Penn, O.* Privman, E.*, Landan, G., Graur, D., and Pupko, T. *Mol Biol Evol, in press*.
* Equal contribution

5.1 Introduction

The investigation into the effect of the guide tree on progressive alignment led to an interesting observation. Even when the accuracy score of the alignment was not significantly affected by the accuracy of the guide tree (in small datasets) we observed that changes in the guide tree gave rise to dramatic changes in the MSA, which were not reflected in significant changes in the accuracy score. We hypothesized that these changes reflect uncertainty in the alignment rather than accuracy improvement. In this chapter I describe the investigation that followed, of the effect of guide tree uncertainty on alignment uncertainty.

As described in Section 1.1 above, an MSA is a prerequisite for virtually all comparative sequence analyses. All such analyses take the MSA input for granted, regardless of uncertainties in the alignment. Since errors in upstream methodologies tend to cascade downstream, alignment errors are an important concern in molecular data analysis. In the last decade, considerable efforts have been made to improve alignment accuracy (see Section 1.5). Nevertheless, studies based on structural-alignment benchmarks (described in Section 1.6 above) such as BAliBASE (Thompson et al. 2005) show that obtaining accurate alignments remain a challenging task. A recent evaluation of SP scores across the BAliBASE benchmark

concluded that the best alignment programs to date achieve only 76% average accuracy, i.e., a quarter of all residue pairs are incorrectly aligned (Nuin, Wang, and Tillier 2006). Therefore, distinguishing between accurate and noisy alignment regions is critical for MSA-dependent analyses, which should try to avoid alignment regions of low quality.

There are several possible sources for errors in sequence alignment. To begin with, all MSA programs use heuristic methods. In contrast to pairwise sequence alignment that can be optimally solved under a given scoring scheme, finding the optimal MSA is computationally prohibitive. Thus, MSA programs usually produce a sub-optimal alignment. Furthermore, even with optimal algorithms for pairwise sequence alignment there are often several co-optimal solutions, i.e., different alignments with the same maximal score. This issue affects all state-of-the-art MSA algorithms that are based on the "progressive alignment approach" (Feng and Doolittle 1987), because they use an optimal pairwise alignment algorithm for iteratively adding sequences to the MSA. Notably, while progressive alignment approaches differ in the manner according to which post-alignment corrections and refinements are made, the progressive alignment step is a critical component in all of them. Landan and Graur (2007; 2008) investigated this source of error and concluded that 80-90% of the columns and 40-50% of aligned residue pairs in a typical MSA are affected by uncertainty due to co-optimal solutions.

An additional point of concern is that the objective functions, which alignment algorithms attempt to maximize, are based on simplified models of the process of molecular sequence

evolution. Such approximations may yield high scores for unrealistic alignments. Therefore, even if we had unlimited computational power to find the set of MSAs with the optimal score, we cannot be confident that it includes the true alignment, since the true alignment may actually be sub-optimal. Additionally, the stochastic nature of sequence evolution introduces noise on top of the signal, and thus the true evolutionary history will often score less than the highest scoring alignment even if a perfect scoring function were available.

Finally, the alignment may be sensitive to errors in the guide tree, which is used for choosing the order in which the sequences are added to the growing MSA in the progressive alignment approach. As described in Section 1.5 above, the greedy nature of the progressive heuristic entails that early mistakes in pairwise alignments cannot be rectified with the addition of information from other sequences in latter stages. This problem may be aggravated when the topology of the guide tree is incorrect, leading to incorrect order of addition of sequences. Indeed, estimates of guide tree accuracy show that, on average, more than 10% of tree branches are topologically incorrect for datasets of 25 taxa, and this proportion increases with the number of taxa (Nelesen et al. 2008). Several studies measured alignment accuracy in terms of the percent of correctly aligned residues, by comparing a reconstructed MSA to a reference "true" benchmark MSA (e.g., Nelesen et al. 2008; Landan and Graur 2009). These studies concluded that the accuracy of the guide tree has a negligible effect on the accuracy score of the alignment. However, as we will show here, perturbations in the tree affect significant portions of the alignment, shifting residues one way or the other, even though the

overall accuracy score does not change significantly. Therefore, we argue that guide tree uncertainty is an important source of alignment uncertainty.

All the above factors contribute to substantial errors in alignments produced by state-of-the-art MSA algorithms. Equally troubling is the fact that, with the notable exception of TCOFFEE (Notredame, Higgins, and Heringa 2000), most of the widely-used MSA programs do not provide information regarding the reliability of different regions in the alignment, e.g., CLUSTALW (Thompson, Higgins, and Gibson 1994), MUSCLE (Edgar 2004), MAFFT (Katoh et al. 2005), and PRANK (Loytynoja and Goldman 2005).

Only a few confidence measures for alignments have been published. In phylogeny reconstruction it is common practice to remove alignment blocks suspect of low quality using the Gblocks program, which defines various cutoffs on the number of gapped sequences in an alignment column (Castresana 2000; Talavera and Castresana 2007). However, these criteria may excessively filter out regions with insertion/deletion events that can be aligned reliably. A few alignment algorithms output site-specific scores that allow the selection of high-confidence regions. Such a service was first offered by the SOAP program (Loytynoja and Milinkovitch 2001), which tests the robustness of each column to perturbation in the parameters of the popular alignment program CLUSTALW. The TCOFFEE web server (Poirot, O'Toole, and Notredame 2003) uses a library of alignments in the construction of the final MSA, and its output MSA is colored according to confidence scores that reflect the agreement between different alignments in the library regarding each aligned residue. Another alignment program

that can output an MSA with confidence scores is FSA (Bradley et al. 2009), which uses a statistical model that allows calculation of the uncertainty in the alignment. Similarly, the HoT (Heads-Or-Tails) score can be used as a measure of site-specific alignment uncertainty due to the co-optimal solutions problem mentioned above (Landan and Graur 2007; Landan and Graur 2008). However, none of these confidence measures account for uncertainties in the guide tree.

An alternative, more statistically justified approach to assess alignment uncertainty is the use of probabilistic evolutionary models for joint estimation of phylogeny and alignment (described in Section 1.7 above). The Bayesian approach allows calculation of posterior probabilities of estimated phylogeny and alignment, which is a measure of the confidence in these estimates across the whole solution space. In comparison, in the approach presented here and the previously published HoT score, perturbations are made to the input of greedy algorithms such as CLUSTALW, which were not designed to consider sub-optimal solutions. Therefore, in principle, we should prefer the Bayesian approach. However, in practice, the Bayesian approach is infeasible for all but the smallest datasets (Section 1.7).

Here we will show that uncertainties in the guide tree have a considerable effect on the robustness of the MSA. Subsequently, we develop a measure quantifying this effect as a confidence score for each column and for each residue in the alignment, based on the robustness of their alignment with respect to perturbations in the guide tree. Our measure is based on the bootstrap method, which is widely used for assigning confidence scores to

branches in reconstructed phylogenetic trees. Benchmark studies using BAliBASE as well as simulated sequences show that our alignment confidence scores are a good predictor of alignment accuracy, significantly improving on the HoT scores. Therefore, we conclude that guide tree uncertainty is an important source of error in sequence alignment, and that MSA-based analyses should take into account site-specific confidence scores, in order to avoid artifacts.

5.2 Methods

5.2.1 Construction of perturbed multiple sequence alignments

We begin with a standard MSA generated by any progressive alignment program, hereby termed "base MSA." Similar to the common practice in phylogeny reconstruction, we use the bootstrap (BP) approach (Felsenstein 1985) to obtain a set of trees that can be used as a proxy to a confidence interval around the inferred tree. These trees are obtained using the neighbor joining (NJ) algorithm (Saitou and Nei 1987). The pairwise distances used as input to the NJ algorithm are maximum likelihood estimates computed using the JTT amino acid replacement matrix (Jones, Taylor, and Thornton 1992). Next, each bootstrap tree is given as an input guide tree to the alignment program. The resulting set of perturbed MSAs is used for estimating the confidence level of the base MSA. As in the BP test for tree branches, the larger the number of perturbed guide trees, the more accurate is the estimated confidence score. In all of our analyses we used 100 BP replicates. The flow of the algorithm is shown in Figure 5.1.

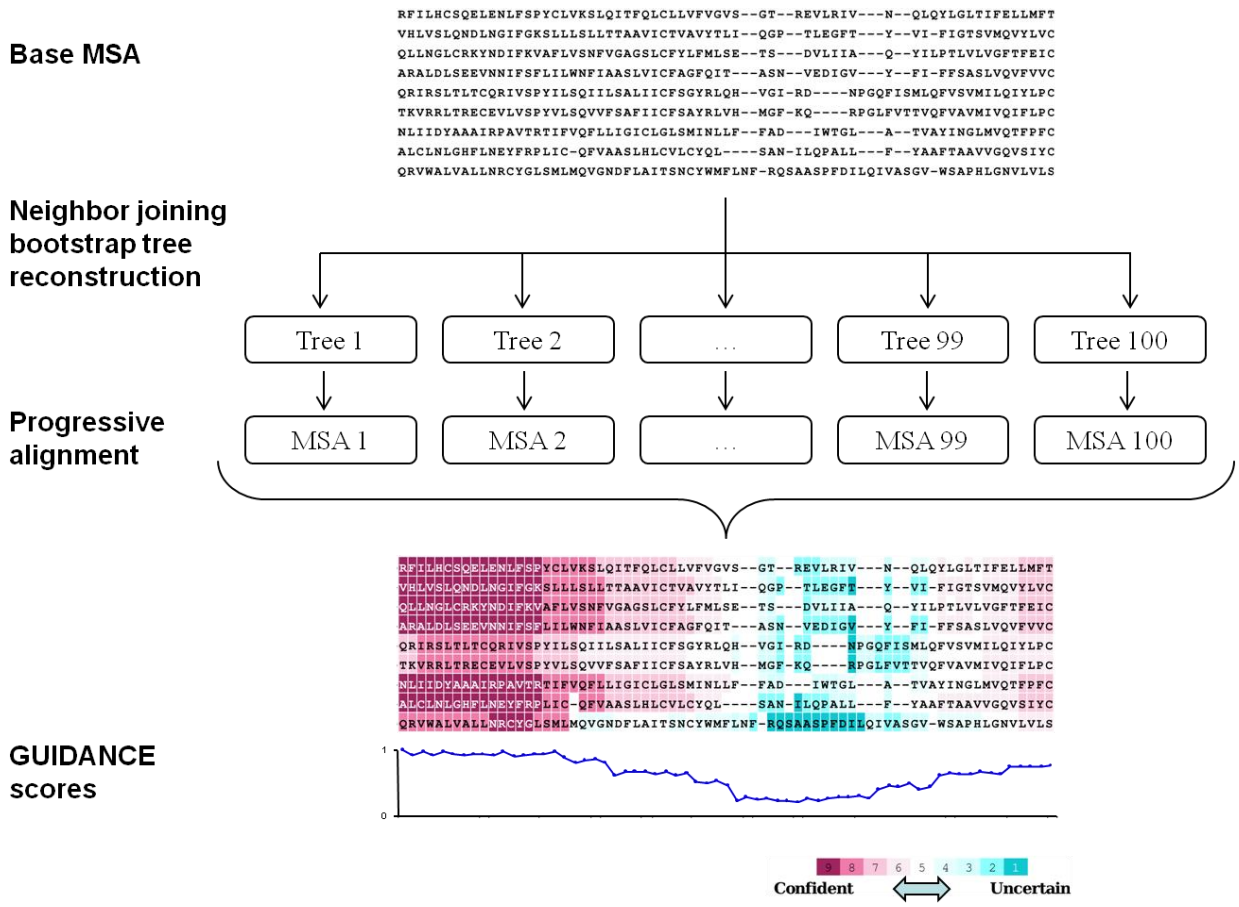

Figure 5.1: The "GUIDe tree based AligNment ConfidencE" (GUIDANCE) measure. A base MSA is produced by any progressive alignment method. Bootstrap neighbor joining (NJ) trees are reconstructed and given as guide trees to the progressive alignment program, producing a set of perturbed MSAs. Sum-of-pairs scores are then calculated by comparing each perturbed MSA to the base MSA, and are color coded on each residue in the alignment.

5.2.2 GUIDANCE confidence score calculation

The main goal of our method is to assign a confidence score for each column of the base MSA, which we name “GUIDe tree based AligNment ConfidencE” (GUIDANCE) scores. To this end, we

define a set of distances that measure the dissimilarity between a specific perturbed MSA and the base MSA. Specifically, three widely used distances are computed:

- Column score (CS): Each column of the base MSA that is identically aligned in the perturbed MSA is given a score of 1; all other columns are given the score 0.
- Sum-of-pairs score (SP): Each pair of residues in the base MSA that is identically aligned in the perturbed MSA is given a score of 1; all other residue pairs are given the score 0.
- Sum-of-pairs column score (SPC): The score of each column is simply the average of the SP scores over all pairs in it.

The CS score cannot distinguish between a column with one error and a column with many errors. In contrast, the SPC score can better quantify the difference between a column in the base MSA and a column in the perturbed MSA. Subsequently, unless stated otherwise, we only use SP and SPC.

Each residue pair in the base MSA can have a score of 1 or 0 in each of the perturbed MSAs. The average score over all perturbed MSAs is a measure of the confidence in aligning these two residues, and is termed here the GUIDANCE residue-pair score. The average SPC score over all perturbed MSAs is termed here the GUIDANCE column score.

Furthermore, we define a confidence score for a specific residue in a specific alignment column, the GUIDANCE residue score. This score is calculated by averaging the GUIDANCE

residue-pair scores over all pairs that include the residue in question. This score reflects the confidence of aligning this specific residue in this column.

5.2.3 Benchmark data

The BAliBASE benchmark database (Thompson et al. 2005) consists of MSAs that are based on structural alignments and are specifically designed for the evaluation and comparison of MSA programs. The database is categorized into several reference sets, according to types of alignment problems. Here we use BAliBASE reference sets 1-5, which include 218 datasets.

We applied the GUIDANCE method to each dataset, using the MAFFT alignment program (version 6.711), generating GUIDANCE residue-pair scores for each pair of aligned residues in the base MSA. We then used the BAliBASE reference alignments in order to assess the predictive power of the GUIDANCE score to identify alignment errors. Each aligned residue pair in the MAFFT base MSA was classified as correct/incorrect by comparing it to the reference MSA. A receiver operating characteristic (ROC) analysis (Green and Swets 1966; Fawcett 2006) was conducted using the R package ROCR (Sing et al. 2005), to evaluate the specificity and sensitivity of the GUIDANCE confidence measure. The performance of the GUIDANCE predictor was measured by the area under the ROC curve (AUC). The BAliBASE reference provides annotations of alignment regions for which the alignment is verified by superposition of protein structures, named core blocks. Therefore, we limited all the BAliBASE analyses to columns belonging to these core blocks only.

5.2.4 Simulations

The advantage of simulation is that the evolutionary history of insertion and deletion events is absolutely known. We used the ROSE program (Stoye, Evers, and Meyer 1998) to simulate protein alignments based on BAliBASE datasets. Each dataset of genuine protein sequences was used to reconstruct a phylogenetic tree using NJ. Site-specific evolutionary rates were estimated using the Bayesian method implemented in rate4site (Mayrose et al. 2004, http://www.tau.ac.il/~itaymay/cp/rate4site.html). We fed the tree and the rates as input to ROSE, thereby producing a simulated dataset for each of the original BAliBASE datasets, mimicking the biological characteristics of these proteins. These simulated datasets were used to conduct the ROC analysis as described above, except that here all columns in the reference alignment were used.

To supplement these simulations in an independent approach that is not based on the BAliBASE data, we also used the INDELible program (Fletcher and Yang 2009) to simulate 100 protein datasets of 50 sequences, using a root sequence length of 300, random trees, a power-law model of indel distribution with indelrate=0.1, gamma-distributed among site rate variation (*alpha*=1), and the LG replacement matrix.

5.2.5 Comparison to the Heads-or-Tails confidence measure

We compared the performance of the GUIDANCE measure to the HoT score, as described in Landan and Graur (2008), using the same MAFFT version (6.711). ROC analysis was performed as described above.

5.3 Results

5.3.1 Most alignment columns are sensitive to guide tree uncertainty

We applied the GUIDANCE method, using both MAFFT (Katoh et al. 2005) and CLUSTALW (Thompson, Higgins, and Gibson 1994), to an exemplary protein dataset consisting of 130 homologous chemoreceptors from *Drosophila melanogaster* (Robertson, Warr, and Carlson 2003). The purpose of this analysis was to study the effect of the guide tree on the resulting MSA, for a typical alignment problem. Figure 5.2 shows the level of agreement between the perturbed MSAs, generated by the GUIDANCE method, and the base MSA, generated by either CLUSTALW or MAFFT, using either column scores (CS) or sum-of-pairs scores (SP). For CLUSTALW, the CS scores vary between 0.029 and 0.11, with a median of 0.053 (Figure 5.2a). That is, in a typical perturbed MSA less than 6% of the columns are identically aligned as in the base MSA. For MAFFT alignments, the median is 11%. Taken together, these results suggest that alignment columns are highly sensitive to uncertainties in the guide tree. We next tested the sensitivity of aligned residue pairs, in terms of the average SP score of each perturbed MSA (Figure 5.2b). For CLUSTALW, the SP scores vary between 0.28 and 0.36, with a median of 0.31. For MAFFT, the SP scores vary between 0.31 and 0.43, with a median of 0.38. These results imply that in any perturbed MSA less than 50% of residue pairs are aligned as in the base MSA.

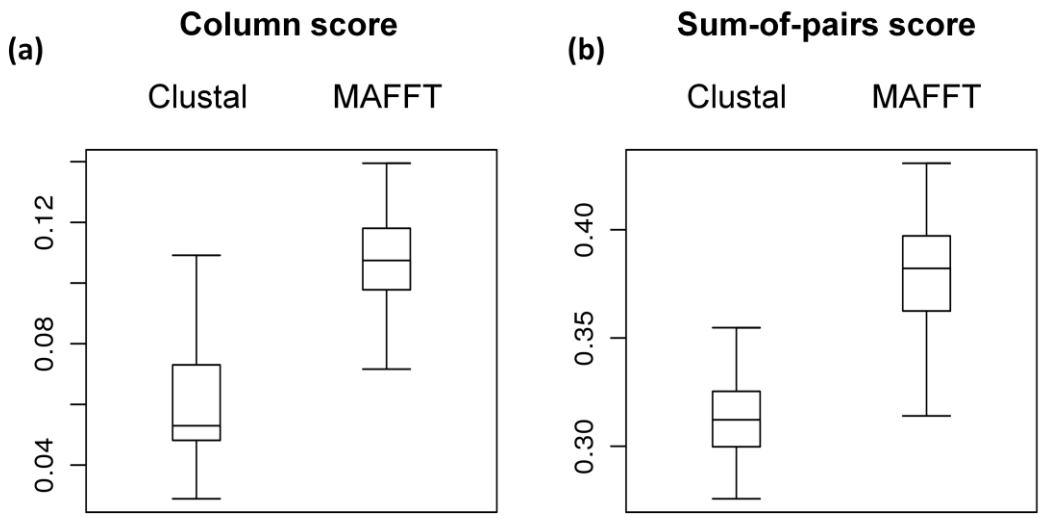

Figure 5.2: Agreement between MSAs built based on perturbed bootstrap trees and the base MSA for MAFFT and CLUSTALW alignments of *D. melanogaster* chemoreceptor sequences. Box plots summarize medians, quartiles, and range of column scores **(a)** and sum-of-pairs scores **(b)**.

5.3.2 GUIDANCE measure can identify alignment errors

Since uncertainty in the guide tree results in alignment uncertainty (as shown above), we hypothesized that alignment errors can be detected by searching for those alignment regions that are sensitive to guide tree perturbations. To this end, we used a continuous range of cutoffs for the GUIDANCE scores. The cutoff was used as a classification criterion to separate columns or residue pairs into reliable and unreliable. In order to test how well this classification correctly detects actual alignment errors, the columns and residue pairs of the inferred alignment should be compared to a known "true" one. Such comparison will reveal the proportions of true positive (correctly aligned residues that are marked as reliable by the

GUIDANCE classifier) and false positive (erroneously aligned residues that are marked as reliable by the GUIDANCE classifier) predictions. Since, in most cases, the "true" alignment is unknown, two approaches were used here to test the performance of the GUIDANCE classifier: (i) comparison against a reference benchmark of curated MSAs, and (ii) simulation studies. In addition, we compare the performance of the GUIDANCE classifier to the previously published HoT score, which was shown to be a highly accurate predictor of alignment errors (Landan and Graur 2008).

BAliBASE benchmark: We applied the GUIDANCE measure, using the MAFFT alignment algorithm, to the BAliBASE benchmark (see Section 5.2.3). Figure 5.3a presents a ROC analysis of GUIDANCE scores and HoT scores for residue pairs, as classifiers of alignment errors relative to the BAliBASE reference. Both methods accurately identified alignment errors, with an advantage to GUIDANCE over HoT, giving AUC values of 94.0% and 89.7%, respectively.

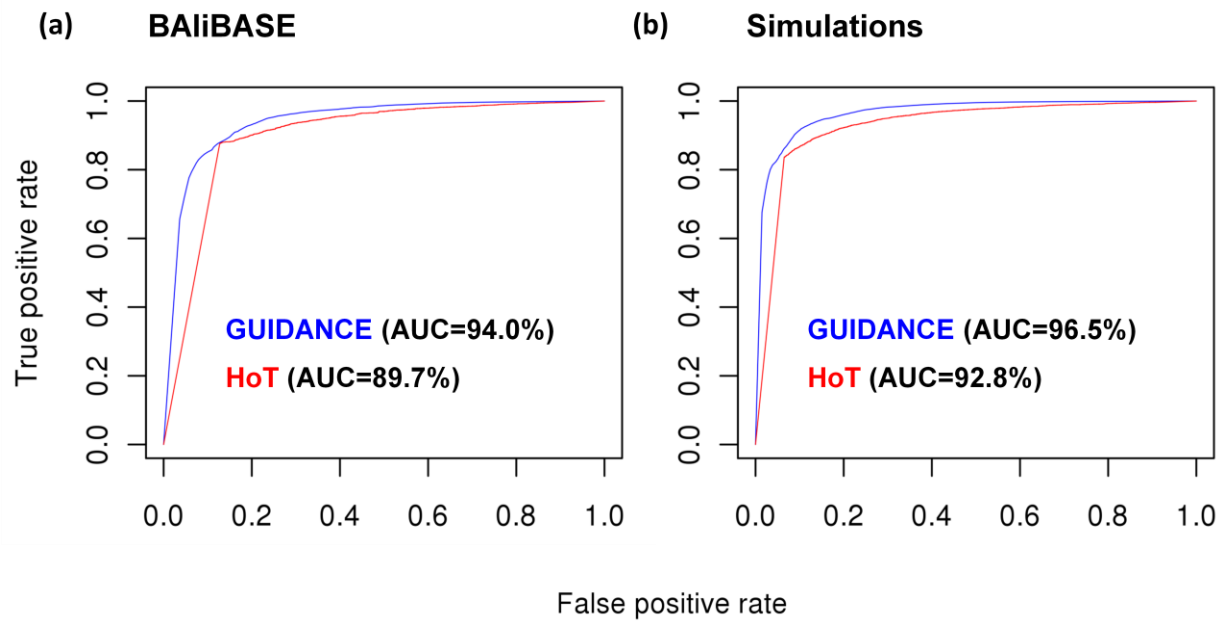

Figure 5.3: Accuracy of GUIDANCE scores in identifying alignment errors. ROC curves for HoT scores (red) and GUIDANCE scores (blue) of aligned residue pairs as predictors for alignment errors in: **(a)** the BAliBASE benchmark; **(b)** the simulations benchmark.

Simulations benchmark: Simulation studies provide further support for the higher accuracy of GUIDANCE scores compared to HoT (Figure 5.3b). As opposed to real protein benchmarks, in which one can never be absolutely sure of the true alignment, the exact locations of gaps are known with certainty in alignments of sequences generated by simulation. However, one has to make sure that the simulation settings reflect as much as possible true evolutionary dynamics. To this end, our simulations were based on the BAliBASE reference MSAs. That is, we simulated a reference alignment based on the phylogenetic tree and site-specific evolutionary rates inferred for each of the 218 datasets in BAliBASE, in order to replicate the

natural evolutionary dynamics of protein families. The GUIDANCE classifier accurately identified alignment errors with an AUC of 96.5%, improving on the 92.8% of the HoT classifier. An example demonstrating the difference between GUIDANCE and HoT is given in Figure 5.4, which plots the distribution of GUIDANCE and HoT column scores, compared to the actual alignment accuracy in the first 260 columns of a typical alignment of 11 simulated sequences. Both GUIDANCE and HoT scores correlate with the actual alignment errors, giving Pearson correlation coefficients of 0.81 and 0.50, respectively.

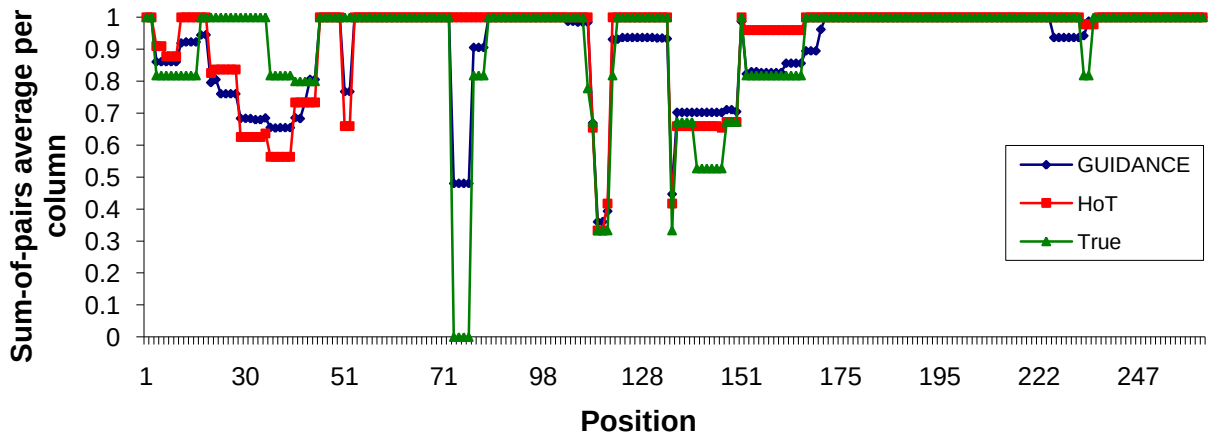

Figure 5.4: An example from the simulations benchmark. Distribution of GUIDANCE column scores (blue) compared to Heads-or-Tails (HoT) scores (red) and the actual alignment accuracy (green) in the first 260 columns of a typical simulated alignment.

Independent simulations of 100 datasets using the INDELilble program (Fletcher and Yang 2009), which were not based on BAliBASE data, gave comparable results – an AUC of 90.1% for

GUIDANCE and 88.4% for HoT. To summarize, the results obtained for the simulated data are in line with those obtained for the BAliBASE benchmark.

A combined GUIDANCE-HoT score: One would expect that GUIDANCE and HoT identify different types of alignment errors. We thus tried to combine the two scores to produce an even more powerful predictor. We investigated several approached in combining the two scores, including weighted average and a minimum function. However, they all produced similar ROC performance as the GUIDANCE measure alone.

Comparison to Gblocks: The Gblocks program (Castresana 2000) is design to eliminate poorly aligned regions of the MSA, effectively giving a binary score for every column. To compare the performance of Gblocks and our method, we run Gblocks on the simulation benchmark using two sets of parameters, "stringent" and "relaxed", as defined in Talavera et al. (2007). Figure 5.5 presents the false-positive and the true-positive rates of Gblocks, together with a ROC analysis of GUIDANCE column scores. The results show that for the same proportion of false-positives, GUIDANCE provides more true-positives, for both the stringent and the relaxed conditions.

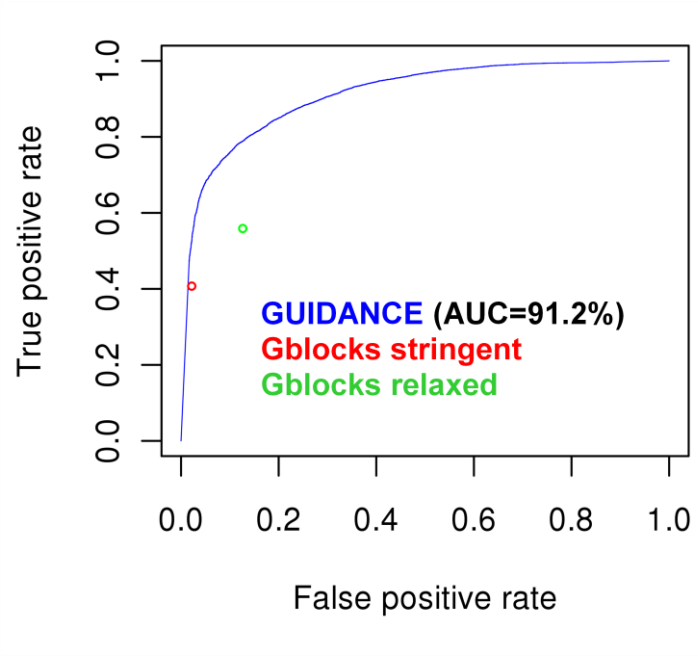

Figure 5.5: Comparison to Gblocks. The false-positive and true-positive rates of Gblocks "stringent" (red) and "relaxed" (green) parameter sets in comparison to a ROC curve for GUIDANCE column scores (blue), for the simulations benchmark.

Figure 5.6 summarizes the overlap between alignment errors that were detected by GUIDANCE and HoT scores as a Venn diagram. The total of 1,914,804 incorrectly aligned residue pairs in the MAFFT reconstruction of the BAliBASE benchmark were classified as detected by either method if their confidence score was less than 1. Almost 10% of the alignment errors were detected by GUIDANCE and not detected by HoT. In contrast, less than 1% of alignment errors were detected by HoT alone. Only 2.8% of alignment errors were not detected by either method.

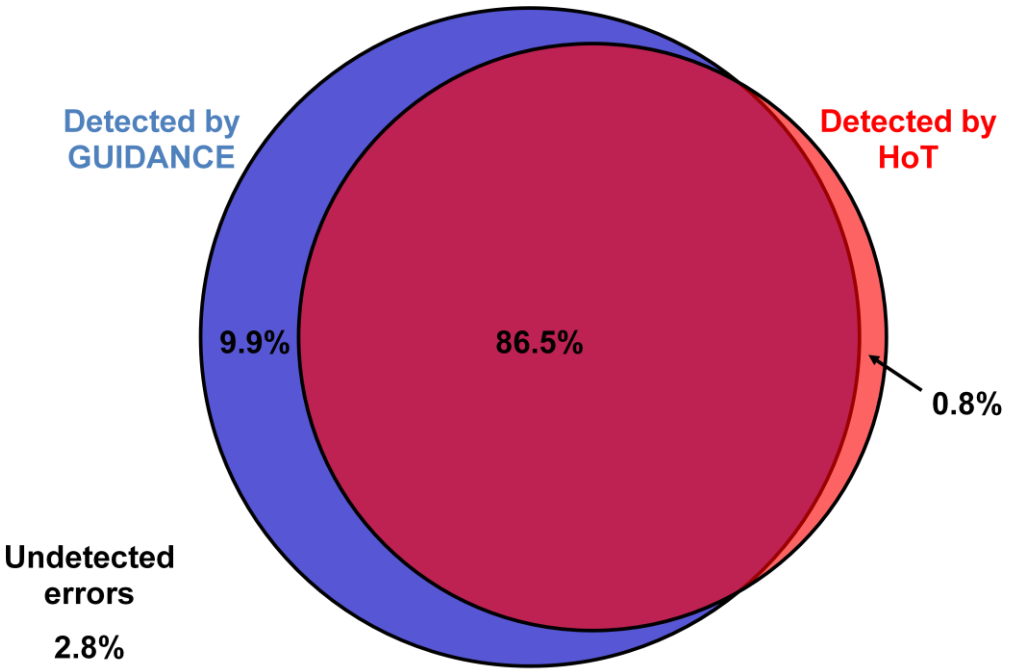

Figure 5.6: Venn diagram of alignment error detection by the GUIDANCE and Heads-or-Tails (HoT) scores. The total of 1,914,804 incorrectly aligned residue pairs in the BAliBASE benchmark were classified as detected by either method if their confidence score was less than 1. GUIDANCE detected 96.4% of the errors while HoT detected 87.3%. The HoT-detected-errors are nearly a subset of the GUIDANCE-detected-errors.

5.3.3 Visualization of alignment uncertainty

To facilitate examination of a specific MSA of interest, we suggest a graphic visualization of alignment uncertainty by coloring the MSA according to the GUIDANCE scores. As an example, Figure 5.7 shows a colored portion of the same MSA of chemoreceptors sequences that was used in Figure 5.2 above. The GUIDANCE residue scores are color coded on the MSA. This is a convenient way to inspect the implications of low-confidence regions for subsequent analysis.

Magenta colored residues can be considered reliable, while blue colored residues should be avoided. In addition, a plot of the GUIDANCE column scores is presented.

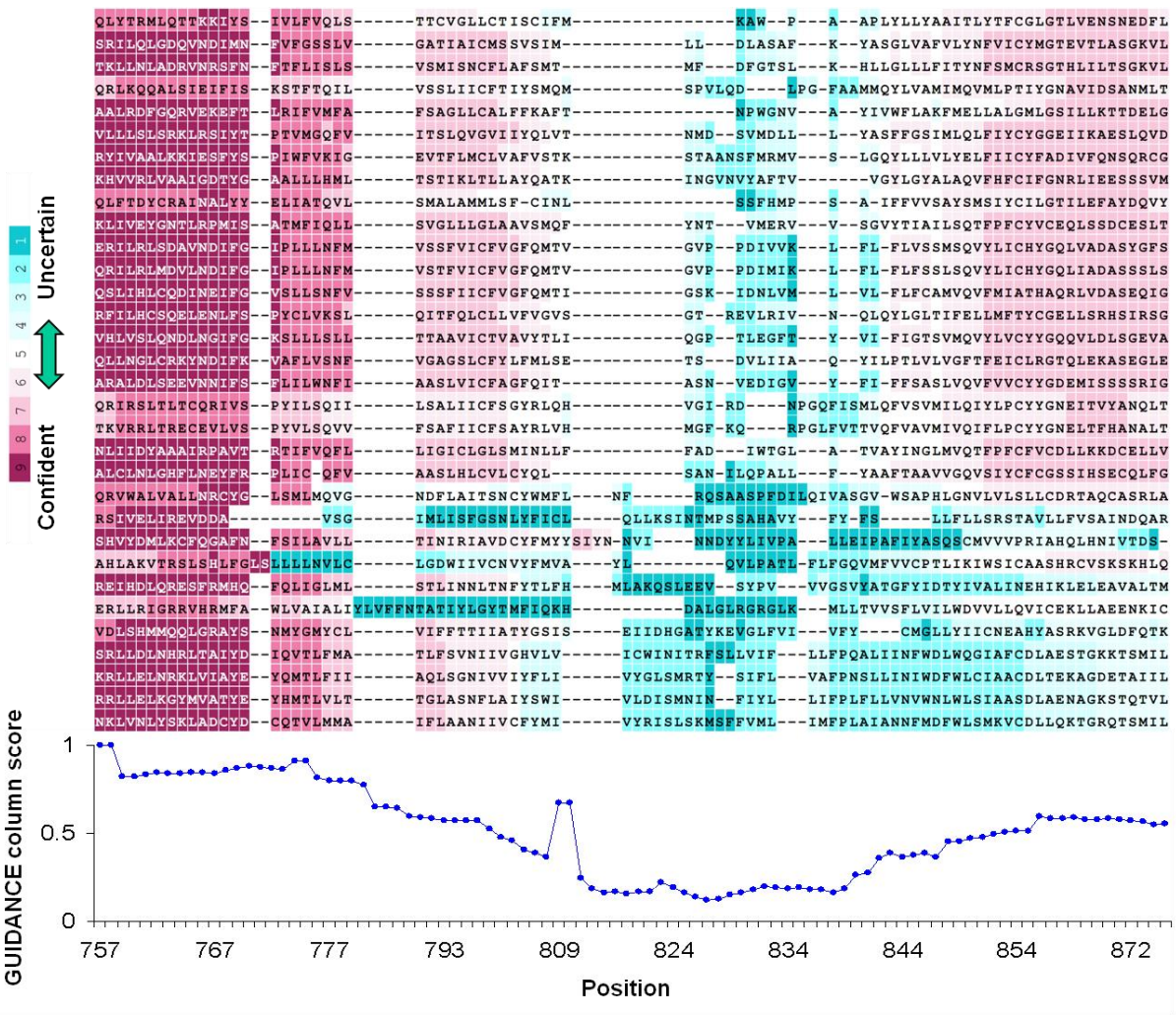

Figure 5.7: Color-coded GUIDANCE scores for *D. melanogaster* chemoreceptor sequences. A portion of the MSA is presented (columns 757-875 of 32 sequences). Confidently aligned residues are colored in shades of magenta and pink, while uncertain residues are colored in shades of blue. GUIDANCE column scores are plotted below the alignment.

As expected, wide gap-less blocks such as the first from the left score close to 100% confidence. Note the alignment is confident even though the sequences are variable. Downstream, the second and third blocks score significantly lower even though they similarly appear to be solid blocks. Furthermore, the GUIDANCE residue scores discriminate between the majority of sequences in the third block that are reliably aligned and two sequences that stand out in unreliable blue. Such a case of a divergent, badly-aligned sequence can be easily discovered using GUIDANCE.

5.4 Summary

The study reported here demonstrates that alignment reliability is dramatically affected by uncertainties in the guide tree. Based on this observation, a new measure for alignment confidence was devised. Bootstrap tree sampling, a proxy to a "confidence interval" around the guide tree, is used to perturb the progressive alignment and to quantify of the robustness of the alignment to such perturbations in the guide tree. Thereby, a measure for guide-tree uncertainty is translated into a measure of alignment uncertainty. This methodology produces the GUIDANCE confidence scores for each aligned residue, which can be summarized for each column or for each sequence in the MSA. The GUIDANCE scores facilitate consideration of alignment reliability of every residue in any downstream MSA-based analysis.

We evaluated the predictive power of GUIDANCE scores to identify alignment errors both for the BAliBASE benchmark of real protein alignments and for simulated alignments. We also compared the new GUIDANCE measure to the previously published HoT score, which is a

measure of alignment unreliability due to the co-optimal solutions problem (Landan and Graur 2007; Landan and Graur 2008). Notably, the HoT score was previously shown to be highly successful in predicting residue pairs that are erroneously aligned, and in the present study we report an AUC of 89.7% for HoT scores applied to the BAliBASE benchmark. The GUIDANCE scores make a substantial improvement on top of that, reaching an AUC value of 94.0%. Simply put, if we pick a point along the ROC plot in Figure 5.3a, we could use GUIDANCE scores to identify 80% of the correctly aligned residues in an average MSA, while "suffering" from only a 5% rate of false positives.

Interestingly, an average or a minimum of the two scores does not improve the AUC any further. This is surprising because one could expect some alignment columns that are uncertain in terms of co-optimal solutions, but not in terms of the robustness to the guide tree. If such columns existed in sufficient numbers then the combination of HoT and GUIDANCE measures should improve the prediction accuracy relative to the GUIDANCE measure alone. Since this is not the case, we conclude that most columns affected by the co-optimality issue are also affected by uncertainty in the guide tree. This does appear to be the case since less than one percent of alignment errors were detected by the HoT score and not by the GUIDANCE score (Figure 5.6). Clearly, while GUIDANCE focuses only on the effect of guide tree on alignment uncertainty, research on other sources of errors beside the guide tree can lead to better detection and quantification of alignment errors.

We conclude that the new alignment confidence measure is a highly accurate predictor for the correctness of specific MSA columns. As such, it is valuable for any MSA-based analysis. We encourage researchers to use the GUIDANCE confidence measure before any downstream analysis, rather than rely on alignments as unquestionable truths.

6 GUIDANCE: a web server for assessing alignment confidence scores

This chapter is based on a published manuscript:
Penn, O.* Privman, E.*, Ashkenazy, H., Landan, G., Graur, D., and Pupko, T. *Nucleic Acid Res* 38:W23-W28.
* Equal contribution

As a follow-up to the previous chapter, the present chapter describes the GUIDANCE web server, an implementation of the GUIDANCE confidence measure providing the following services: (i) producing MSAs accompanied with their site-specific confidence scores; (ii) graphically projecting these scores onto the MSA; and (iii) filtering and re-aligning low confidence regions. The server points to columns and sequences that are unreliably aligned and enables their automatic removal from the MSA, in preparation for downstream analyses. The GUIDANCE server has a user-friendly interface, intuitive graphical results, and is freely available for use at http://guidance.tau.ac.il with no requirement of log-in. Two algorithms for quantifying MSA uncertainties are implemented in the server. The GUIDANCE score measures the robustness of the MSA to guide-tree uncertainty as described in Chapter 55 above. The Heads-or-Tails (HoT) score measures alignment uncertainty due to co-optimal solutions (Landan and Graur 2007; Landan and Graur 2008).

Similar tools exist for assessing alignment confidence, such as T-COFFEE (Poirot, O'Toole, and Notredame 2003), SOAP(Loytynoja and Milinkovitch 2001), and MUMSA (Lassmann and Sonnhammer 2005b). The advantages of the web server implemented here are: (a) it is based on robust statistical measures of MSA reliability for quantifying two major sources of alignment

uncertainty (co-optimal solutions and guide-tree uncertainty) that are not addressed by other tools; (b) it allows the user to fine-tune the degree to which unreliable MSA parts are removed; (c) it implements a range of MSA algorithms and evolutionary models (for codons, amino-acids and nucleotides); (d) it is straightforward and easy to use.

6.1 Methods

The minimal input requirement for running the server is a set of DNA, RNA or protein sequences in FASTA format. The general flow of the program is as follows: (i) a standard MSA is generated, hereby termed "base MSA," by applying one of several progressive MSA algorithms; (ii) a set of perturbed MSAs is constructed according to the alignment confidence algorithm (HoT of GUIDANCE, see below); (iii) The set of MSAs is compared to the base MSA in order to estimate its confidence level. This comparison results in confidence scores between 0-1 for each residue, residue-pair, column, and sequence of the MSA, which are essentially different ways to average Sum-of-Pairs (SP) scores (Carrillo and Lipman 1988; Thompson, Plewniak, and Poch 1999); (iv) the confidence scores of all residues are projected onto the MSA, using a color-scale and the column scores are plotted below the alignment; (v) unreliable columns and sequences may be removed from the base MSA. The server currently supports three progressive alignment algorithms: CLUSTALW, MAFFT, and PRANK (Thompson, Higgins, and Gibson 1994; Katoh et al. 2005; Loytynoja and Goldman 2005).

The above procedure differs between GUIDANCE and HoT in the way that the set of perturbed MSA is created. GUIDANCE scores reflect the robustness of an alignment to guide tree

uncertainty. The GUIDANCE method perturbs the guide tree used to build the MSA, using bootstrap sampling (described in detail in Chapter 5). On the other hand, HoT scores reflect alignment uncertainty due to co-optimal solutions in the progressive alignment procedure. Here the set of perturbed MSAs is constructed by reversing the sequences at each of the pairwise alignment steps of the progressive alignment algorithm (Landan and Graur 2008).

6.1.1 Adjustable parameters

The server implements a few advanced options that are useful for fine-tuning the results. For the GUIDANCE algorithm, the number of bootstrap repeats can be set by the user (the default value is set to 100). The higher this number is, the more accurate the confidence score, but the running time increases linearly. The cutoffs according to which columns and sequences are filtered out for subsequent analysis are also adjustable. It is possible to change these cutoffs according to the proportion of columns\sequences that the user wishes to retain. The order of the sequences in the output MSA may be set according to the input file, or according to the alignment algorithm result file.

In addition, the server allows uploading a user MSA file instead of the sequences file. In this case, the input MSA is used as the base MSA and the confidence scores are calculated in the same way as described above. This option should be used with caution. It is useful for analyzing an MSA of interest, for example, an MSA that was generated using a more accurate guide-tree than the standard neighbor joining tree. However, it is important to remember that even when the base MSA is given as input, the alignment algorithm chosen is applied many

times in order to generate each of the perturbed MSAs. Therefore, supplying an MSA created by one program and inferring its confidence using another program may result in false predictions.

Advanced users can also alter the parameters passed on to the alignment program used. For example, by default, the server runs PRANK with the "+F" flag, but the experienced user may wish to remove that option in some cases (see http://www.ebi.ac.uk/goldman-srv/prank/). For MAFFT the user may enable the iterative refinement option and set the number of iterations in the MAXITERATE parameter. Additionally, an option to choose between the iterative refinement strategies genafpair, localpair, and globalair is provided when running MAFFT. See the MAFFT website for a description of these options and scenarios where their use is recommended (http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html).

6.1.2 Output

The main result of the GUIDANCE server is a graphical visualization of the confidence scores which consists of two parts (Figure 1a): (i) Color-scaled projection of the confidence scores of each residue onto the base MSA; (ii) A plot of the column scores. Text files are also produced containing the confidence scores for each column, residue, residue-pair, and sequence. In addition, the following MSA files are provided: (i) The base MSA; (ii) The MSA containing only reliable columns that passed a predefined threshold (this file may be used in downstream analyses such as phylogeny reconstruction); (iii) A sequence file of reliable sequences only (again for a predefined threshold). It is recommended to rerun GUIDANCE on the filtered

sequences as input, in order to re-align them without the disruptive effect of the badly aligned sequences (see example below). This can be done simply by clicking on the button next to the output file link.

6.2 Case study: the HIV Vpu accessory protein

We illustrate using GUIDANCE to identify unreliable alignment columns by analyzing an MSA of Vpu protein sequences from human and simian immunodeficiency viruses (HIV and SIV). These viruses are known for their high rate of evolution, which is attributed to an arms race between the virus and the host immune system (Rambaut et al. 2004). The Vpu protein has been recently shown to antagonize the host protein Tetherin, an innate immune factor, in order to promote viral release and replication (Neil, Zang, and Bieniasz 2008). Therefore, it is a natural candidate for many evolutionary analyses that rely on an MSA. Our purpose here was to demonstrate the use of GUIDANCE to evaluate reliability of the MSA, and the importance of this evaluation for the interpretation of downstream analyses.

Vpu is an accessory viral protein present in HIV-1 and SIV infecting chimpanzees and other primate species, yet absent in HIV-2. The protein contains ~80 amino-acids and has two known major functions, which are conducted by two distinct domains of the protein: (i) promotion of CD4 degradation via the cytoplasmic domain; and (ii) enhancement of virion release from host cells via the transmembrane domain, which has shown to be related to antagonism of Tetherin (Neil, Zang, and Bieniasz 2008; Nomaguchi, Fujita, and Adachi 2008).

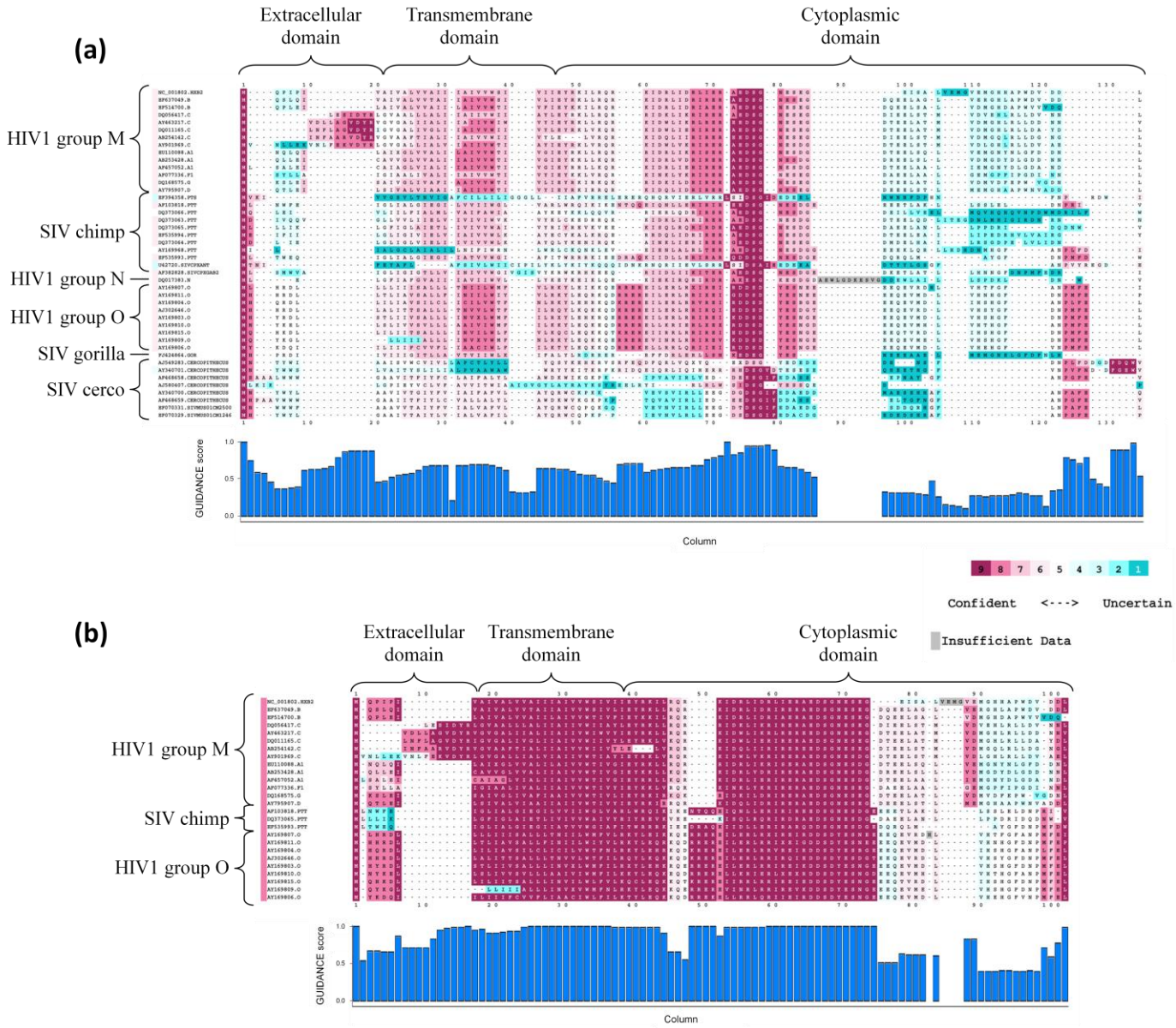

Figure 6.1: An example of the GUIDANCE output. (a) Residue confidence scores are projected onto the MAFFT alignment of Vpu protein sequences from human and simian immunodeficiency viruses (HIV and SIV). Confidently aligned residues are colored in shades of magenta and pink, while uncertain residues are colored in shades of blue. Column scores are plotted below the alignment. **(b)** Dramatically improved alignment confidence after filtering low scoring sequences and re-running GUIDANCE. Note the color-coding next to the sequence names before and after re-alignment.

We ran GUIDANCE on a sample of Vpu protein sequences from the three main HIV-1 groups (M, N and O) and SIV sequences from chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), and several *Cercopithecus* species, using MAFFT (Figure 6.1a). The results clearly show that the alignment of the cytoplasmic domain of Vpu is not robust to perturbations in the guide tree. The same applies to some residues in the transmembrane and extracellular domains. Looking at specific sequences, the SIV sequences from *Cercopithecus* and some of the sequences from *P. troglodytes* are shown to be badly aligned with the rest of the sequence set. By simply pressing a button, these sequences were filtered and GUIDANCE was rerun on the confidently aligned sequences only. The results demonstrate a dramatic improvement in MSA confidence (Figure 6.1b). The transmembrane domain and the 5' region of the cytoplasmic domain now receive almost perfect confidence scores. Note that although a clade of sequences was excluded by the GUIDANCE filter and the alignment is now considerably more condense, the remaining sequences are still highly variable and several gapped regions have been retained. The removal of the unconfidently-aligned sequences is necessary to avoid artifacts that they would have otherwise caused in downstream analyses such as inference of positive selection (Wong, Suchard, and Huelsenbeck 2008; Schneider et al. 2009).

Even after removing the low-scoring sequences, the alignment of the 3' region of the cytoplasmic domain is uncertain, thus, downstream analyses on these regions should be interpreted with caution. When appropriate, one may use the filtered MSA, provided by GUIDANCE, which contains only the reliable columns (e.g., for inference of positive selection).

This example demonstrates the importance of using GUIDANCE for removing badly aligned sequences that may disrupt the MSA and for noting which columns are suspect of alignment errors, which might affect downstream analysis.

6.2.1 Implementation

The GUIDANCE web server runs on a Linux cluster of 2.6GHz AMD Opteron processors, equipped with 4GB RAM per quad-core node. At the moment, our cluster will allocate up to 16 cores for GUIDANCE runs submitted through the web server, and the allocation of resources will grow with demand. The server runs up-to-date versions of the supported multiple alignment programs and an in-house implementation of neighbor joining bootstrap tree reconstruction. The HoT and GUIDANCE algorithms are implemented in Perl and C++. The source code of GUIDANCE is also available on the website, for large scale analyses, which users may want to run locally using their own computational resources (http://guidance.tau.ac.il/).

Running time depends on the dataset size (number and length of sequences) and (for GUIDANCE scores) on the number of bootstrap repeats. The major component of the running time is the multiple alignment program used, thus MAFFT runs will be fastest and PRANK runs slowest. To aid users with estimating running time for their datasets, we include a plot of average GUIDANCE and HoT running times using either MAFFT or PRANK for several dataset sizes, from 100 to 350 sequences, roughly 300 amino acids in length (Figure 6.2). Note that GUIDANCE was run with the default 100 bootstrap repeats, but this number can be reduced to

shorten the running time. HoT running time depends on the number of branches in the guide tree, which increases linearly with the number of sequences.

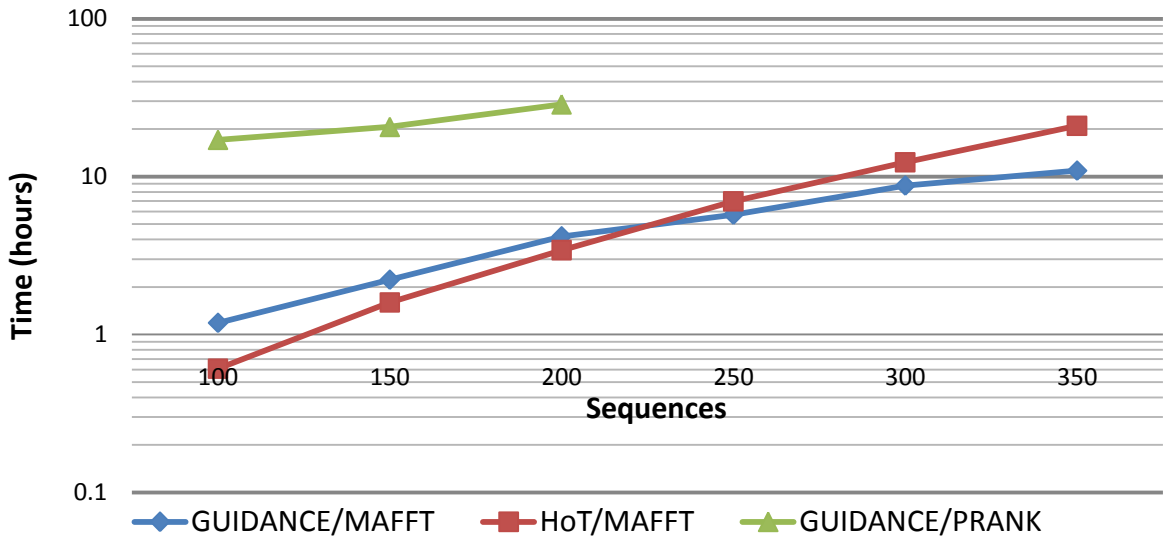

Figure 6.2: Average run-time performance as a function of the number of sequences. Simulated protein sequences roughly 300 amino acids long were aligned using MAFFT and analyzed by GUIDANCE (blue diamonds) or HoT (red squares). In addition, running time for GUIDANCE on PRANK alignments is plotted with green triangles. Each data point represents ten replicates.

7 Discussion

In this dissertation I have investigated computational methodologies for comparative sequence analysis. The methods developed attempt to address challenges posed on the one hand by the multiple revolutions in sequencing technologies that repeatedly multiplied the breadth of our knowledge of sequences from single genes, to whole genomes, and to vast collections of genomes, and on the other hand by the realization that reliable processing and analysis of these data require complex models and sophisticated algorithms. These two trends stretch the limits of computational power that grows at a much more modest rate of doubling approximately every two years (Moore 2005). By comparison, the number of nucleotides stored in the EMBL Nucleotide Sequence Database doubles approximately every 16 months (Goldman and Yang 2008). The increased complexity of models and algorithms for sequence analysis aggravate this growing chasm, because they require more processing power per nucleotide.

The algorithmic enhancements developed here take the leading computational approaches in terms of accuracy and efficiency and attempt to combine and build upon their respective strengths. Thereby, the proposed methods may provide the most accurate reconstruction of trees and MSAs that is computationally feasible for large sequence datasets. Naturally, others have developed independent performance improvements for phylogeny and alignment. As discussed below, many of these improvements can be combined with those developed here. For this reason, the methods developed in these studies were implemented as modular tools,

allowing for maximum reusability. All software is freely distributed as open source libraries to encourage integration with other bioinformatics packages, and extensive user interfaces were developed for comfortable control of most features. As much as possible, support is given to multiple operating systems to allow wide distribution. Ultimately, a web server was set up in order to facilitate wide usage by non-computational researchers in molecular biology by providing an intuitive, graphical user interface and automatic processing of user uploaded data.

7.1 Efficient and accurate phylogeny reconstruction

Perhaps the greatest difficulty is in algorithms with exponential relationship between sequence number and computation time for a given dataset of homologous sequences. This generally holds for state-of-the-art phylogeny reconstruction by probabilistic methods (ML and Bayesian approaches, reviewed in Section 1.3) because they attempt to search across the solution space, or the set of all possible tree topologies, which grows rapidly with the number of sequences (Cavalli-Sforza and Edwards 1967).

This limitation of standard probabilistic approaches is the primary motivation for the phylogenetic methods developed in Chapter 3. A hybrid ML-NJ approach was developed to take the best from two worlds – the accurate evolutionary modeling of the probabilistic (ML) paradigm and the computational efficiency of the distance-based paradigm (NJ is used here). The hybrid approach can be viewed as a speed improvement for probabilistic model-based tree reconstruction. It can also be viewed as an accuracy improvement for distance-based tree

reconstruction. In his discussion of distance-based methods, Felsenstein (2004) eluded to their limitation compared to ML methods:

> *"When evolutionary rates vary from site to site in molecular sequences, distances can be corrected for this variation… In likelihood methods, the correction can use information from changes in one part of the tree in inform the correction in others. Once a particular part of the molecule is seen to change rapidly in the primates, this will affect the interpretation of that part of the molecule among the rodents as well. But a distance matrix method is inherently in capable of propagating the information in this way. Once one is looking at changes within rodents, it will forget where changes were seen among primates. Thus distance matrix methods must use information about rate variation substantially less efficiently then likelihood methods. This casts a cloud over their use, one which may prove hard to dispel."* Felsenstein (2004, p. 175)

The integration of ML methodology into distance-based phylogeny reconstruction aims to address exactly this weakness. The ML approach is used to fit a rich evolutionary model to a fixed input (MSA and tree topology) and then this model is used to efficiently reconstruct a more accurate tree using the distance-based NJ method. Iterations of tree building and model fitting allow quick convergence on a more accurate tree without the need to calculate the computationally intensive likelihood function for many topologies, as in ML tree searches. In fact, since most of the contribution to accuracy is in the second iteration then one round of likelihood optimization could suffice for the largest, most computationally challenging datasets.

Simulation studies demonstrated that models accounting for site-specific rates significantly improve distance estimation for proteins with average to high levels of rate variation. This result demonstrates the detrimental effect of methods and models that make over-simplifying assumptions regarding sequence evolution. And since distance-based methods such as NJ are statistically consistent (Atteson 1997), that is, will give an accurate tree when supplied with accurate distances, then we can expect significant improvement of the tree due to a significant improvement of the distance estimation. This effect is evident in the evaluations of tree reconstruction by the ML-NJ hybrid. Distance-based tree reconstruction achieved improved accuracy while still retaining the ability to process thousands of sequences in reasonable running time. To summarize, the hybrid approach allows use of complex models for accurate tree reconstruction in an efficient manner that is not feasible with the standard probabilistic approaches such as ML tree search and Bayesian MCMC.

7.1.1 Rapid maximum likelihood tree search

In parallel with the development of our hybrid approach, dramatic efficiency improvements have been accomplished with the standard ML approach. Most notably, the RAxML package has been developed over the years, implementing several significant efficiency enhancements that reduce the cost of each likelihood computation, make the search for the ML tree more efficient to reduce the number of likelihood computations, and to allow parallelization across multi-core platforms that are now widely available (Stamatakis, Ludwig, and Meier 2004; Stamatakis, Ludwig, and Meier 2005; Stamatakis 2006; Stamatakis, Hoover, and Rougemont

2008). Rapidly reconstructed RAxML trees are still equally accurate or better than other more time-consuming implementations. Thereby RAxML allows ML tree search for datasets as large as 25,000 sequences (Stamatakis 2006). Another noteworthy implementation is GARLI (Zwickl 2006) that is also capable of processing thousands of sequences. Having said that, efficient ML tree search can still benefit from higher accuracy of the distance-based tree that is used as the starting point for the search, as discussed in Section 1.3 above. Improved starting point will shorten the search for the ML tree, and possibly avoid local maxima in some scenarios. Furthermore, many of the efficiency enhancements developed for RAxML are also applicable to the use of probabilistic models in the hybrid method, and pairwise distance estimation can be easily parallelized because each pair is considered independently.

Efficient distance-based reconstruction may still be the preferable solution for even larger trees. For example, a tree of 200,000 rRNA sequences was reconstructed by an approximate distance-based method (Katoh and Toh 2007), and the Ribosomal Database Project (Cole et al. 2009) now contains 1,237,963 small subunit rRNA sequences (Release 10, Update 20, as of May 19, 2010) which theoretically could be analyzed as a single MSA. Therefore, it seems that larger and larger datasets will continue to be limited to distance-based tree reconstruction. The accuracy of such trees may be improved by the hybrid approach described here.

7.1.2 Applications of the hybrid method

The application of the SSRV model to bacterial phylogeny (Section 3.4) demonstrates the modularity of the ML-NJ hybrid and its potential for the integration of more advanced

evolutionary models in order to relieve the bias and error when model assumptions are violated by certain more challenging biological datasets. Hopefully, this potential of the hybrid approach developed in this thesis will be used in future studies.

Therefore, special attention was given to the distribution of the source code for the hybrid ML-NJ implementation in a manner that will maximize its usability by the wider scientific community. The algorithm was implemented as part of the SEMPHY package – an extensive, freely distributed, open source C++ library of probabilistic models and algorithms for phylogeny (http://compbio.cs.huji.ac.il/semphy/). The library implements the range of complex algorithms in a modular, object-oriented fashion that maximizes code reusability. For the non-programmer, a user interface was designed to allow control of many of the algorithmic variations, model choice, and parameterization that may be beneficial in different scenarios. Over the past eight years the SEMPHY package has been maintained and freely distributed. According the ISI Web of knowledge, SEMPHY has been cited by 44 publications (http://apps.isiknowledge.com). One such opportunity for the application of SEMPHY to the study of a specific phylogeny developed into a collaboration between the laboratories of Prof. Tal Pupko and Prof. Sara Lavi. Our paper on the phylogeny of the protein phosphatase 2C superfamily is attached as an appendix to this manuscript.

7.2 The effect of guide-tree accuracy on MSA accuracy

The mutual dependency of alignment and phylogeny appears very one-sided in the literature of comparative sequence analysis. Current phylogenetic studies rely on MSAs of nucleotide or

amino-acid sequences and phylogeneticists take great care in preparing their MSAs to apply the best alignment algorithms, often followed by manual corrections, and filtering of reliable alignment blocks. However, no such efforts are made to supply progressive alignment algorithms with the best possible guide trees. MSAs are aligned according to guide trees built by inaccurate NJ or UPGMA. Chapter 4 described an investigation of the contribution of improved phylogenetic accuracy using the hybrid ML-NJ within the widely used iterative scheme of phylogeny reconstruction and progressive alignment. Results of simulation studies demonstrate that improved guide tree accuracy becomes significant when the number of sequences increases to the hundreds (this thesis and Liu et al. 2009).

As with the iterative ML-NJ hybrid, the iterative scheme of phylogeny and alignment enjoys the advantage of modularity. Although the simulation studies described here are limited to CLUSATW and PRANK, any progressive alignment program can be easily used with any phylogeny reconstruction method. Hence, it is advisable to choose the most accurate phylogenetic method that is computationally feasible for the dataset in question. For tens of thousands of rRNA sequences one might be limited to distance-based methods, but for a thousand sequences or less one can take advantage of the fast ML search algorithms mentioned in Section 7.1 above. When the number of sequences drops to a hundred or less, the significance of the guide tree diminishes and simple NJ will do.

7.3 Alignment confidence

Errors in MSAs cascade into downstream analyses that depend on them. The reliance on an alignment as a fixed foundation for comparative sequence analysis is a noteworthy flaw in the vast majority of MSA-based research, with the exception of a handful of studies that address the uncertainty in the alignment using Bayesian methodology, especially MCMC (see Section 1.7). The severity of this flaw is intensified by the frequency of alignment errors. It was estimated that a quarter of the residues are incorrectly aligned, even when the most accurate alignment algorithms are used (Nuin, Wang, and Tillier 2006). Therefore, comparative sequence research is in dire need of reliable measures for alignment confidence that will enable identification of error-prone alignment regions.

Chapter 5 described the development and evaluation of the GUIDANCE confidence measure. Guide tree uncertainty was shown to be a major factor in MSA uncertainty. This observation gave rise to the GUIDANCE algorithm that uses bootstrap tree sampling to quantify the sensitivity of MSAs to guide tree uncertainty, and to estimate site-specific alignment confidence scores. Evaluation using simulations and real protein benchmark data demonstrated the predictive power of GUIDANCE scores to accurately identify alignment errors.

7.3.1 Limitations of the GUIDANCE method

The use of bootstrap trees as guide trees for progressive sequence alignment may seem at first ill advised. The bootstrap sampling technique deliberately introduces noise into the

reconstruction of the tree, creating trees with some errors in the branching order of the internal nodes. When the process of progressive alignment reaches an erroneously reconstructed internal node, the alignment attempts to represents an ancestral sequence that did not exist in the true evolutionary history. However, the fundamental assumption of our approach is that the conventionally used guide-tree most often contains numerous errors (e.g., Nelesen et al. 2008). Therefore, the bootstrap sampling of perturbed trees provides a statistically justified representation of the level of error in the guide tree.

Ideally, alignment and trees should be reconstructed simultaneously taking into account uncertainties in all related parameters: tree topology, branch lengths, indel probabilities and indel length distribution, substation models, rate variation, etc. Bayesian methods that use MCMC provide a suitable solution to achieve exactly that (see Section 1.7). A by-product of the MCMC is a confidence measure in terms of the posterior probabilities of each alignment column. Our approach can be viewed as related to MCMC, except that only uncertainty in tree topology is accounted for (and all other parameters are fixed). In our method, the set of bootstrap trees is a sample from the space of possible tree topologies. A trivial modification on our method would be to use a set of trees sampled using MCMC as guide trees instead of the bootstrap trees used in GUIDANCE. The posterior probabilities of the MCMC sampling may be used to weight the different trees. While this approach enjoys a stronger statistical justification, it is likely to increase the computational burden and prohibit the use of our method for large datasets.

Another point worth noting is that the GUIDANCE confidence score is absolutely dependant on uncertainty in the guide tree. In principle, it is possible to have 100% bootstrap support for the guide tree, in which case the GUIDANCE confidence will be 100% for every alignment column. This may be true for some trees of very few sequences. In such scenarios the HoT score (which is also implemented in the web server) may still be capable of detecting unreliably aligned regions because it is not affected by the guide tree. However, in practice, one rarely sees 100% support for all tree branches. Indeed, this does not happen in any of the 218 datasets in the BAliBASE benchmark, even though many of them contain fewer than ten sequences. Therefore, in general, it is recommended to use the GUIDANCE method that outperforms HoT on both the BAliBASE benchmark and simulations studies (Section 5.3.24).

A practical consideration with our approach is the increased running time required for (typically 100) bootstrap repeats, reconstructing many guide trees and MSAs. However, since we use simple NJ bootstrap trees, and the relatively fast MAFFT alignment algorithm, this increased running time will often be negligible in comparison to the running time of downstream analysis, such as Bayesian phylogeny reconstruction or positive selection inference.

7.3.2 The GUIDANCE web server and usage in downstream MSA-based analyses

Chapter 6 described the GUIDANCE web server and usage of GUIDANCE confidence scores in preparation for MSA-based analyses. The Vpu case study (Section 6.2) was chosen to demonstrate the utility of the GUIDANCE method. The field of HIV genomics sets many challenges for comparative sequence analysis. First and foremost, the virus is a rapidly evolving

pathogen, which is a major obstacle for disease control, but also an asset and a challenge for evolutionary investigation via comparative sequence analysis. As evident from GUIDANCE results for the MSA of Vpu sequences, the alignment of HIV and SIV sequences is often difficult and can be expected to harbor error-prone regions. Second, and as a consequence of rapid evolution, the intensive efforts to characterize HIV genetically yielded tens of thousands of genomic sequences that represent the extensive variance of these viruses. Effective utilization of such vast datasets of homologous sequences requires both accurate and computationally feasible methodologies. Third, the evolutionary perspective has been used extensively to derive insights into HIV biology and the development of the global pandemic. Both the (repeated) zoonotic transfers from apes to humans and the adaptation and diversification of HIV subtypes in humans were investigated using evolutionary approaches. Specifically, phylogenetic analyses were used to infer selection forces acting on specific genes and specific residues in the viral genome (e.g., Leitner et al. 1996; Nielsen and Yang 1998; Crandall et al. 1999; Zanotto et al. 1999; Draenert et al. 2004; Leslie et al. 2004; Penn et al. 2008). However, such methodologies, especially site-specific positive selection inference, are sensitive to alignment errors that inflate their predictions (Wong, Suchard, and Huelsenbeck 2008; Schneider et al. 2009).

The GUIDANCE server offers researchers tools to deal with these challenges. Variation on server usage in preparation for different types of phylogenetic analysis may include: (i) choice between the most accurate alignment by PRANK and the most efficient alignment by MAFTT

that can handle thousands of sequences (further tuning of the compromise between accuracy and speed is offered via advanced parameters passed to MAFFT); (ii) manual review of a color-coded MSA for quick visual assessment of large datasets, or numerical tables of scores for automatic processing; (iii) Automatic removal of low-scoring sequences re-alignment of the dataset to avoid the disruptive effect of these sequence; (iv) filtering reliable columns in preparation to subsequent analysis such as phylogeny reconstruction.

GUIDANCE is recommended for use in conjunction with any and all MSA-based studies, since virtually all MSAs are affected by some degree of uncertainty. The type of downstream analysis may dictate different modes of running GUIDANCE and of usage of GUIDANCE scores. It is generally recommended to use GUIDANCE to filter out badly aligned sequences and re-align the data, since such sequences usually disrupt the alignment among the other sequences, which could be reliably aligned otherwise. However, the option for removing alignment columns may or may not be used, depending on the expected sensitivity of the downstream analysis to alignment errors.

Perhaps the most widely used MSA-based analysis is phylogeny reconstruction. It is common practice to filter gap-less blocks in the alignment and only use those columns for phylogeny reconstruction. Gblocks (Castresana 2000; Talavera and Castresana 2007) is usually used for this purpose, but a comparative evaluation (in Section 5.3.2 above) demonstrated that the accuracy of filtering columns containing alignment errors by GUIDANCE is superior over Gblocks. The merits of removing columns for phylogeny reconstruction may vary between

different datasets and different evolutionary scenarios because of the delicate balance between filtering noise and loss of evolutionary information. Therefore, it is debated whether columns should be removed for phylogeny reconstruction (Gatesy, DeSalle, and Wheeler 1993; Giribet and Wheeler 1999; Aagesen 2004). Furthermore, the choice of cutoff on the confidence scores clearly affects the tradeoff between the sensitivity and the specificity in the identification of alignment errors. There are no specific recommended values for these cutoffs because their effect on the alignment varies considerably among datasets. The web server provides a list of cutoffs with their respective effects on the remaining proportion of sequences/columns and users are encouraged to experiment with several cutoffs, especially when removing sequences and re-aligning the dataset.

Several phylogenetic methodologies attempt to detect evolutionary patterns in a site-specific manner, and these should be considered in light of the evolutionary phenomena that are sought after. On the one hand, Bayesian methods for site-specific rate inference as implemented in the Consurf web server (Landau et al. 2005) are usually robust to a few badly aligned residues in a column. It stands to reason that a column corresponding to a conserved site will still be inferred as conserved as long as most of the data are correctly aligned. On the other hand, as mentioned above, site-specific prediction of positive selection using the Ka/Ks measure, as in the Selecton web server (Stern et al. 2007), may be more vulnerable to alignment errors (Wong, Suchard, and Huelsenbeck 2008; Schneider et al. 2009). These methods seeks fast evolving sites that are generally more difficult to align. A few badly aligned

residues may inflate the Ka/Ks estimate for the column and lead to false inference of positive selection. Moreover, Ka/Ks inference of positive selection is only considered if the whole gene passes the statistical significance threshold in a likelihood ration test (LRT). Therefore, the inclusion of badly aligned columns in this test may be detrimental for certain genes that erroneously pass the LRT threshold due to the inflated Ka/Ks scores in these columns. Similar considerations of alignment confidence may be also applicable to other analyses involving site-specific rate inference such as rate-shift detection (Gu 1999; Moreira, Le Guyader, and Philippe 1999; Knudsen and Miyamoto 2001; Wang and Gu 2001; Pupko and Galtier 2002; Abhiman and Sonnhammer 2005; Penn et al. 2008).

The removal of a whole column because of a subgroup of badly aligned residues results in loss of reliable information from other confidently aligned residues in the same column. Less radical use of the GUIDANCE scores may minimize information loss. GUIDANCE calculates confidence scores for individual residues in the MSA. Low-scoring residues may be masked, for example by substituting them with “missing data” or gap characters. In phylogenetic analyses this procedure will prevent biases due to the badly aligned residues, such as the tendency for over-prediction of positive selection (Schneider et al. 2009), while still allowing use of the evolutionary information in rest of the column. Thereby, a finer separation of signal from noise can be achieved.

7.3.3 Wide distribution of GUIDANCE in the scientific community

The theme of modularity is common to all the hybrid methodologies developed in this thesis. The GUIDANCE confidence measure is applicable to any and all MSA algorithms based on the progressive alignment technique. This strength distinguishes it from other confidence measures discussed in Section 5.1 above (e.g., TCOFFEE) which are specific to individual alignment programs. The only requirements that GUIDANCE makes is that the alignment program receives an input guide tree. Some programs do not offer this feature, but this should be easy to implement in any algorithm that is based on progressive alignment. Therefore, GUIDANCE enjoys a broad relevance.

As with the distribution of the hybrid ML-NJ as part of the SEMPHY package (Section 7.1.2 above) considerable efforts were invested in making GUIDANCE available to the scientific community. Due to its general applicability to all MSA-based studies, the GUIDANCE measure should be relevant to wider population of potential users. Here we were able to provide a web server to avoid the need for software installation, provide a more user-friendly graphical interface and extensive utilities such the automatic alignment filtering by sequences and/or columns. A share of a large Linux cluster was dedicated to running server jobs managed by a high-throughput queuing system. Despite this investment the use of GUIDANCE is offered free of charge. Therefore, we hope to target a wider audience. Since its publication five weeks ago, the site received 65 visitors from 13 countries. 51 jobs were submitted from 19 unique IP

addresses. We are doing our best to promote usage of GUIDANCE by conducting workshops and presentations in local and international conferences.

We also made all software involved in this project publicly available as open source. A downloadable, open-source program allows researchers conducting large scale computational analysis to run analyses on their high-throughput infrastructures. Furthermore, the open source enables fellow bioinformaticians to reuse and integrate GUIDANCE in their development of alignment algorithms, which is especially important for taking advantage of the modularity and generality discussed above.

7.4 Concluding remarks

The value of advances in bioinformatics methodology goes beyond any specific applied study such as the reconstruction of the bacterial phylogeny (Section 3.4), the PP2C superfamily (Appendix A), or the analysis of the Vpu protein (Section 6.2). The full value of a computational tool is found in the sum of small contributions to all applied studies that use it, and also in subsequent methodological developments that are based on it. During the studies described in this thesis I have worked with that intention in mind. I can only hope that some of the innovations proposed here will prove valuable to future research in molecular biology.

References

Aagesen, L. 2004. The information content of an ambiguously alignable region, a case study of the trnL intron from the Rhamnaceae. Mol Phylogenet Evol **4**:35-49.

Abhiman, S., and E. L. Sonnhammer. 2005. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins **60**:758-768.

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol Biol Evol **20**:255-266.

Aloy, P., E. Querol, F. X. Aviles, and M. J. Sternberg. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol **311**:395-408.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**:3389-3402.

Ansorge, W. J. 2009. Next-generation DNA sequencing techniques. N Biotechnol **25**:195-203.

Atteson, K. 1997. The performance of neighbor-joining algorithms of phylogeny reconstruction. Computing and Combinatorics:101-110.

Baba, M. L., L. L. Darga, M. Goodman, and J. Czelusniak. 1981. Evolution of cytochrome C investigated by the maximum parsimony method. J Mol Evol **17**:197-213.

Barford, D., Z. Jia, and N. K. Tonks. 1995. Protein tyrosine phosphatases take off. Nat Struct Biol **2**:1043-1053.

Batzoglou, S. 2005. The many faces of sequence alignment. Brief Bioinform **6**:6-22.

Behal, R. H., D. B. Buxton, J. G. Robertson, and M. S. Olson. 1993. Regulation of the pyruvate dehydrogenase multienzyme complex. Annu Rev Nutr **13**:497-520.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2004. GenBank: update. Nucleic Acids Res **32**:D23-26.

Bolen, J. B. 1995. Protein tyrosine kinases in the initiation of antigen receptor signaling. Curr Opin Immunol **7**:306-311.

Bradley, R. K., A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. 2009. Fast statistical alignment. PLoS Comput Biol **5**:e1000392.

Brochier, C., and H. Philippe. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature **417**:244.

Brown, J. R., and W. F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci U S A **92**:2441-2445.

Bruno, W. J., N. D. Socci, and A. L. Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol **17**:189-197.

Burnett, G., and E. P. Kennedy. 1954. The enzymatic phosphorylation of proteins. J Biol Chem **211**:969-980.

Carrillo, H., and D. Lipman. 1988. The multiple sequence alignment problem in biology. SIAM Journal on Applied Mathematics **48**:1073-1082.

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

Cavalli-Sforza, L. L., and A. W. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet **19**:233-257.

Cheng, A., P. Kaldis, and M. J. Solomon. 2000. Dephosphorylation of human cyclin-dependent kinases by protein phosphatase type 2C alpha and beta 2 isoforms. J Biol Chem **275**:34744-34749.

Cheng, A., K. E. Ross, P. Kaldis, and M. J. Solomon. 1999. Dephosphorylation of cyclin-dependent kinases by type 2C protein phosphatases. Genes Dev **13**:2946-2957.

Cohen, P. 1989. The structure and regulation of protein phosphatases. Annu. Rev. Biochem. **58**:453-508.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res **37**:D141-145.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res **16**:10881-10890.

Cowan, K. J., and K. B. Storey. 2003. Mitogen-activated protein kinases: new signaling pathways functioning in cellular responses to environmental stress. J Exp Biol **206**:1107-1115.

Crandall, K. A., C. R. Kelsey, H. Imamichi, H. C. Lane, and N. P. Salzman. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol **16**:372-382.

Daubin, V., M. Gouy, and G. Perriere. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res **12**:1080-1090.

Daubin, V., M. Gouy, and G. Perriere. 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform **12**:155-164.

Do, C. B., M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res **15**:330-340.

Draenert, R., S. Le Gall, K. J. Pfafferott, A. J. Leslie, P. Chetty, C. Brander, E. C. Holmes, S. C. Chang, M. E. Feeney, M. M. Addo, L. Ruiz, D. Ramduth, P. Jeena, M. Altfeld, S. Thomas, Y. Tang, C. L. Verrill, C. Dixon, J. G. Prado, P. Kiepiela, J. Martinez-Picado, B. D. Walker, and P. J. Goulder. 2004. Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. J Exp Med **199**:905-915.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**:214.

Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press.

Eck, R. V., and M. O. Dayhoff. 1966. Atlas of protein sequence and structure 1966. National Biomedical Research Foundation, Silver Spring, Maryland.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792-1797.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A **93**:13429-13434.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern recognition letters **27**:861-874.

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

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Pp. 401.

Felsenstein, J. 1995. PHYLIP (phylogeny inference package), version 3.57 c. Department of Genetics, University of Washington, Seattle.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates Sunderland, MA.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-376.

Felsenstein, J. 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics **5**:164-166.

Feng, D. F., and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol **25**:351-360.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic zoology **20**:406-416.

Fletcher, W., and Z. Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol **26**:1879-1888.

Forterre, P., and H. Philippe. 1999. Where is the root of the universal tree of life? BioEssays **21**:871-879.

Friedman, N., M. Ninio, I. Pe'er, and T. Pupko. 2002. A structural EM algorithm for phylogenetic inference. J Comput Biol **9**:331-353.

Gaits, F., K. Shiozaki, and P. Russell. 1997. Protein phosphatase 2C acts independently of stress-activated kinase cascade to regulate the stress response in fission yeast. J Biol Chem **272**:17873-17879.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol **18**:866-873.

Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**:220-221.

Gardner, P. P., A. Wilm, and S. Washietl. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res **33**:2433-2439.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol **14**:685-695.

Gatesy, J., R. DeSalle, and W. Wheeler. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol Phylogenet Evol **2**:152-157.

Germot, A., and H. Philippe. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J Eukaryot Microbiol **46**:116-124.

Giribet, G., and W. C. Wheeler. 1999. On gaps. Mol Phylogenet Evol **13**:132-143.

Goldman, N., and Z. Yang. 2008. Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. Philosophical Transactions of the Royal Society B **363**:3889-3892.

Goodman, M., G. W. Moore, J. Barnabas, and G. Matsuda. 1974. The phylogeny of human globin genes investigated by the maximum parsimony method. J Mol Evol **3**:1-48.

Goodman, M., and J. F. Pechere. 1977. The evolution of muscular parvalbumins investigated by the maximum parsimony method. J Mol Evol **9**:131-158.

Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol **264**:823-838.

Green, D. M., and J. A. Swets. 1966. Signal detection theory and phycophysics. John Wiley & Sons, New York.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol **16**:1664-1674.

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

Guindon, S., F. Lethiec, P. Duroux, and O. Gascuel. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res **33**:W557-559.

Halanych, K. M. 2004. The new view of animal phylogeny. Annu. Rev. Ecol. Evol. Syst. **35**:229-256.

Hanada, M., J. Ninomiya-Tsuji, K. Komaki, M. Ohnishi, K. Katsura, R. Kanamaru, K. Matsumoto, and S. Tamura. 2001. Regulation of the TAK1 signaling pathway by protein phosphatase 2C. J Biol Chem **276**:5753-5759.

Hardison, R., and W. Miller. 1993. Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. Mol Biol Evol **10**:73-102.

Hasegawa, M., and M. Fujiwara. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol Phylogenet Evol **2**:1-5.

Hasegawa, M., H. Kishino, and N. Saitou. 1991. On the maximum likelihood method in molecular phylogenetics. J Mol Evol **32**:443-445.

Hein, J. 2001. An algorithm for statistical alignment of sequences related by a binary tree. Pac Symp Biocomput **6**:179–190.

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

Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet **4**:275-284.

Holmquist, R., T. H. Jukes, H. Moise, M. Goodman, and G. W. Moore. 1976. The evolution of the globin family genes: concordance of stochastic and augmented maximum parsimony genetic distances for alpha hemoglobin, beta hemoglobin and myoglobin phylogenies. J Mol Biol **105**:39-74.

Howe, K., A. Bateman, and R. Durbin. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics **18**:1546-1547.

Huang, B., R. Gudi, P. Wu, R. A. Harris, J. Hamilton, and K. M. Popov. 1998. Isoenzymes of pyruvate dehydrogenase phosphatase. DNA-derived amino acid sequences, expression, and regulation. J Biol Chem **273**:17680-17688.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst Biol **44**:17-48.

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

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310-2314.

Husmeier, D., and F. Wright. 2001. Detection of recombination in DNA multiple alignments with hidden Markov models. J Comput Biol **8**:401-427.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8**:275-282.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Just, W. 2001. Computational complexity of multiple sequence alignment with SP-score. J Comput Biol **8**:615-623.

Karlin, S., and H. M. Taylor. 1975. A first course in stochastic processes. Academic Press, New York.

Karow, J. 2009. Leerink Report: About 900 Next-Gen Sequencers Deployed to Date; Market Poised for Growth. GenomeWeb: http://www.genomeweb.com/leerink-report-about-900-next-gen-sequencers-deployed-date-market-poised-growth.

Kashiwaba, M., K. Katsura, M. Ohnishi, M. Sasaki, H. Tanaka, Y. Nishimune, T. Kobayashi, and S. Tamura. 2003. A novel protein phosphatase 2C family member (PP2Czeta) is able to associate with ubiquitin conjugating enzyme 9. FEBS Lett **538**:197-202.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res **33**:511-518.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**:3059-3066.

Katoh, K., and H. Toh. 2007. PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics **23**:372-374.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst Biol **45**:363-374.

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics **167**:1513-1524.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16**:111–120.

Kluge, A. G., and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. Systematic zoology **18**:1-32.

Knudsen, B., and M. M. Miyamoto. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci U S A **98**:14512-14517.

Koh, C. G., E. J. Tan, E. Manser, and L. Lim. 2002. The p21-activated kinase PAK is negatively regulated by POPX1 and POPX2, a pair of serine/threonine phosphatases of the PP2C family. Curr Biol **12**:317-321.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980-984.

Komaki, K., K. Katsura, M. Ohnishi, M. Guang Li, M. Sasaki, M. Watanabe, T. Kobayashi, and S. Tamura. 2003. Molecular cloning of PP2Ceta, a novel member of the protein phosphatase 2C family. Biochim Biophys Acta **1630**:130-137.

Krause, D. S., and R. A. Van Etten. 2005. Tyrosine kinases as targets for cancer therapy. N Engl J Med **353**:172-187.

Kumar, S., K. Tamura, and M. Nei. 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci **10**:189-191.

Labes, M., J. Roder, and A. Roach. 1998. A novel phosphatase regulating neurite extension on CNS inhibitors. Mol Cell Neurosci **12**:29-47.

Landan, G., and D. Graur. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol **24**:1380-1383.

Landan, G., and D. Graur. 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. Pac Symp Biocomput **13**:15-24.

Landan, G., and D. Graur. 2009. Characterization of pairwise and multiple sequence alignment errors. Gene **441**:141-147.

Landau, M., I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res **33**:W299-302.

Larget, B., and D. L. Simon. 1999. Markov chasin Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol **16**:750.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**:1095-1109.

Lassmann, T., and E. L. Sonnhammer. 2005a. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics **6**:298.

Lassmann, T., and E. L. Sonnhammer. 2005b. Automatic assessment of alignment quality. Nucleic Acids Res **33**:7120-7128.

Lawson, J. E., S. H. Park, A. R. Mattison, J. Yan, and L. J. Reed. 1997. Cloning, expression, and properties of the regulatory subunit of bovine pyruvate dehydrogenase phosphatase. J Biol Chem **272**:31625-31629.

Leitner, T., D. Escanilla, C. Franzen, M. Uhlen, and J. Albert. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci U S A **93**:10864-10869.

Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S. A.

Thomas, A. St John, T. A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B. D. Walker, and P. J. Goulder. 2004. HIV evolution: CTL escape mutation and reversion after transmission. Nat Med **10**:282-289.

Leung-Hagesteijn, C., A. Mahendra, I. Naruszewicz, and G. E. Hannigan. 2001. Modulation of integrin signal transduction by ILKAP, a protein phosphatase 2C associating with the integrin-linked kinase, ILK1. Embo J **20**:2160-2170.

Lewis, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol Biol Evol **15**:277.

Li, M. G., K. Katsura, H. Nomiyama, K. Komaki, J. Ninomiya-Tsuji, K. Matsumoto, T. Kobayashi, and S. Tamura. 2003. Regulation of the interleukin-1-induced signaling pathways by a novel member of the protein phosphatase 2C family (PP2Cepsilon). J Biol Chem **278**:12013-12021.

Lipman, D. J., S. F. Altschul, and J. D. Kececioglu. 1989. A tool for multiple sequence alignment. Proc Natl Acad Sci U S A **86**:4412-4415.

Liu, K., S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science **324**:1561-1564.

Lopez, P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covarion model. J Mol Evol **49**:496-508.

Loytynoja, A., and N. Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A **102**:10557-10562.

Loytynoja, A., and N. Goldman. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science **320**:1632-1635.

Loytynoja, A., and M. C. Milinkovitch. 2001. SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics **17**:573-574.

Lunter, G., I. Miklos, A. Drummond, J. L. Jensen, and J. Hein. 2005. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics **6**:83.

Mailund, T., G. S. Brodal, R. Fagerberg, C. N. Pedersen, and D. Phillips. 2006. Recrafting the neighbor-joining method. BMC Bioinformatics **7**:29.

Mann, D. J., D. G. Campbell, C. H. McGowan, and P. T. Cohen. 1992. Mammalian protein serine/threonine phosphatase 2C: cDNA cloning and comparative analysis of amino acid sequences. Biochim Biophys Acta **1130**:100-104.

Manning, G., D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. 2002. The protein kinase complement of the human genome. Science **298**:1912-1934.

Mao, M., M. C. Biery, S. V. Kobayashi, T. Ward, G. Schimmack, J. Burchard, J. M. Schelter, H. Dai, Y. D. He, and P. S. Linsley. 2004. T lymphocyte activation gene identification by coregulated expression on DNA microarrays. Genomics **83**:989-999.

Marley, A. E., A. Kline, G. Crabtree, J. E. Sullivan, and R. K. Beri. 1998. The cloning expression and tissue distribution of human PP2Cbeta. FEBS Lett **431**:121-124.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol **21**:1781-1791.

Meskiene, I., E. Baudouin, A. Schweighofer, A. Liwosz, C. Jonak, P. L. Rodriguez, H. Jelinek, and H. Hirt. 2003. Stress-induced protein phosphatase 2C is a negative regulator of a mitogen-activated protein kinase. J Biol Chem **278**:18945-18952.

Moore, G. E. 2005. Moore's Law at 40. Moore's Law at 40: Chemistry and the Electronics Revolution,. Chemical Heritage Foundation, Philadelphia.

Moreira, D., H. Le Guyader, and H. Philippe. 1999. Unusually high evolutionary rate of the elongation factor 1 alpha genes from the Ciliophora and its impact on the phylogeny of eukaryotes. Mol Biol Evol **16**:234-245.

Morgenstern, B., K. Frech, A. Dress, and T. Werner. 1998. DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics **14**:290-294.

Murphy, W. J., T. H. Pringle, T. A. Crider, M. S. Springer, and W. Miller. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research **17**:413-421.

Murray, M. V., R. Kobayashi, and A. R. Krainer. 1999. The type 2C Ser/Thr phosphatase PP2Cgamma is a pre-mRNA splicing factor. Genes Dev **13**:87-97.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48**:443-453.

Neil, S. J., T. Zang, and P. D. Bieniasz. 2008. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature **451**:425-430.

Nelesen, S., K. Liu, D. Zhao, C. R. Linder, and T. Warnow. 2008. The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. Pac Symp Biocomput **13**:25-36.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.

Nomaguchi, M., M. Fujita, and A. Adachi. 2008. Role of HIV-1 Vpu protein for virus spread and pathogenesis. Microbes Infect **10**:960-967.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302**:205-217.

Nuin, P. A., Z. Wang, and E. R. Tillier. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics **7**:471.

Ofek, P., D. Ben-Meir, Z. Kariv-Inbal, M. Oren, and S. Lavi. 2003. Cell cycle regulation and p53 activation by protein phosphatase 2C alpha. J Biol Chem **278**:14299-14305.

Pang, A., A. D. Smith, P. A. Nuin, and E. R. Tillier. 2005. SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. BMC Bioinformatics **6**:236.

Penn, O., A. Stern, N. D. Rubinstein, J. Dutheil, E. Bacharach, N. Galtier, and T. Pupko. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. PLoS Comput Biol **4**:e1000214.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc Biol Sci **267**:1213-1221.

Poirot, O., E. O'Toole, and C. Notredame. 2003. Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res **31**:3503-3506.

Prajapati, S., U. Verma, Y. Yamamoto, Y. T. Kwak, and R. B. Gaynor. 2004. Protein phosphatase 2Cbeta association with the IkappaB kinase complex is involved in regulating NF-kappaB activity. J Biol Chem **279**:1739-1746.

Pupko, T., and N. Galtier. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci **269**:1313-1316.

Pupko, T., and I. Mayrose. 2010. Probabilistic methods and rate heterogeneity *in* H. M. Lodhi, and S. H. Muggleton, eds. Elements of Computational Systems Biology. Wiley.

Raghava, G. P., S. M. Searle, P. C. Audley, J. D. Barber, and G. J. Barton. 2003. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics **4**:47.

Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes. 2004. The causes and consequences of HIV evolution. Nat Rev Genet **5**:52-61.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol **43**:304-311.

Redelings, B. D., and M. A. Suchard. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst Biol **54**:401-418.

Robertson, H. M., C. G. Warr, and J. R. Carlson. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proc Natl Acad Sci U S A **100 Suppl 2**:14537-14542.

Robinson, D., and L. Foulds. 1979. Comparison of weighted labelled trees. Pp. 119-126. Combinatorial Mathematics VI. Springer, Berlin / Heidelberg.

Saitou, N., and T. Imanishi. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol **6**:51.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406-425.

Sankoff, D. 1975. Minimal mutation trees of sequences. SIAM Journal on Applied Mathematics **28**:35-42.

Sankoff, D., C. Morel, and R. J. Cedergren. 1973. Evolution of 5S RNA and the non-randomness of base replacement. Nat New Biol **245**:232-234.

Schlessinger, J. 2000. Cell signaling by receptor tyrosine kinases. Cell **103**:211-225.

Schneider, A., A. Souvorov, N. Sabath, G. Landan, G. H. Gonnet, and D. Graur. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biology and Evolution **2009**:114.

Seroussi, E., N. Shani, D. Ben-Meir, A. Chajut, I. Divinski, S. Faier, S. Gery, S. Karby, Z. Kariv-Inbal, O. Sella, N. I. Smorodinsky, and S. Lavi. 2001. Uniquely conserved non-translated regions are involved in generation of the two major transcripts of protein phosphatase 2Cbeta. J Mol Biol **312**:439-451.

Sheneman, L., J. Evans, and J. A. Foster. 2006. Clearcut: a fast implementation of relaxed neighbor joining. Bioinformatics **22**:2823-2824.

Shenolikar, S. 1994. Protein serine/threonine phosphatases - new avenues for cell regulation. Annu. Rev. Cell. Biol. **10**:56-86.

Shokat, K. M. 1995. Tyrosine kinases: modular signaling enzymes with tunable specificities. Chem Biol **2**:509-514.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics **21**:3940-3941.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol **147**:195-197.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy: the principles and practice of numerical classification. WH Freeman San Francisco.

Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin **38**:1409-1438.

Sonnhammer, E. L., S. R. Eddy, and R. Durbin. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins **28**:405-420.

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

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol **57**:758-771.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21**:456-463.

Stamatakis, A., T. Ludwig, and H. Meier. 2004. New fast and accurate heuristics for inference of large phylogenetic trees. Parallel and Distributed Processing Symposium.

Steppan, S. J., B. L. Storz, and R. S. Hoffmann. 2004. Nuclear DNA phylogeny of the squirrels (Mammalia: Rodentia) and the evolution of arboreality from c-myc and RAG1. Mol Phylogenet Evol **30**:703-719.

Stern, A., A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach, and T. Pupko. 2007. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res **35**:W506-511.

Stoye, J., D. Evers, and F. Meyer. 1998. Rose: generating sequence families. Bioinformatics **14**:157-163.

Strimmer, K., and A. Von Haeseler. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol **13**:964-969.

Strovel, E. T., D. Wu, and D. J. Sussman. 2000. Protein phosphatase 2Calpha dephosphorylates axin and activates LEF-1-dependent transcription. J Biol Chem **275**:2399-2403.

Suchard, M. A., and B. D. Redelings. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics **22**:2047-2048.

Sullivan, J., Z. Abdo, P. Joyce, and D. L. Swofford. 2005. Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. Mol Biol Evol **22**:1386-1392.

Sun, H., and N. K. Tonks. 1994. The coordinated action of protein tyrosine phosphatases and kinases in cell signaling. Trends Biochem Sci **19**:480-485.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci U S A **99**:16138.

Takekawa, M., M. Adachi, A. Nakahata, I. Nakayama, F. Itoh, H. Tsukuda, Y. Taya, and K. Imai. 2000. p53-inducible wip1 phosphatase mediates a negative feedback regulation of p38 MAPK-p53 signaling in response to UV radiation. Embo J **19**:6517-6526.

Takekawa, M., T. Maeda, and H. Saito. 1998. Protein phosphatase 2Calpha inhibits the human stress-responsive p38 and JNK MAPK pathways. Embo J **17**:4744-4752.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol **56**:564-577.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol **9**:678-687.

Tamura, S., K. R. Lynch, J. Larner, J. Fox, A. Yasui, K. Kikuchi, Y. Suzuki, and S. Tsuiki. 1989. Molecular cloning of rat type 2C (IA) protein phosphatase mRNA. Proc Natl Acad Sci U S A **86**:1796-1800.

Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol **11**:261-277.

Terasawa, T., T. Kobayashi, T. Murakami, M. Ohnishi, S. Kato, O. Tanaka, H. Kondo, H. Yamamoto, T. Takeuchi, and S. Tamura. 1993. Molecular cloning of a novel isotype of Mg(2+)-dependent protein phosphatase beta (type 2C beta) enriched in brain and heart. Arch Biochem Biophys **307**:342-349.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**:4673-4680.

Thompson, J. D., P. Koehl, R. Ripp, and O. Poch. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins **61**:127-136.

Thompson, J. D., F. Plewniak, and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res **27**:2682-2690.

Thorne, J. L., H. Kishino, and J. Felsenstein. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol **33**:114-124.

Thorne, J. L., H. Kishino, and J. Felsenstein. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol **34**:3-16.

Travis, S. M., and M. J. Welsh. 1997. PP2C gamma: a human protein phosphatase with a unique acidic domain. FEBS Lett **412**:415-419.

Van Walle, I., I. Lasters, and L. Wyns. 2005. SABmark--a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics **21**:1267-1268.

Wang, L., and T. Jiang. 1994. On the complexity of multiple sequence alignment. J Comput Biol **1**:337-348.

Wang, Y., and X. Gu. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics **158**:1311-1320.

Waterman, M. S., and M. D. Perlwitz. 1984. Line geometries for sequence comparisons. Bulletin of Mathematical Biology **46**:567-577.

Wenk, J., H. I. Trompeter, K. G. Pettrich, P. T. Cohen, D. G. Campbell, and G. Mieskes. 1992. Molecular cloning and primary structure of a protein phosphatase 2C isoform. FEBS Lett **297**:135-138.

Wera, S., and B. A. Hemmings. 1995. Serine/threonine protein phosphatases. Biochem J **311 (Pt 1)**:17-29.

Wheeler, W. C., and D. S. Gladstein. 1994. MALIGN: a multiple sequence alignment program. Journal of Heredity **85**:417.

Woese, C. R. 1987. Bacterial evolution. Microbiol Rev **51**:221-271.

Wong, K. M., M. A. Suchard, and J. P. Huelsenbeck. 2008. Alignment uncertainty and genomic analysis. Science **319**:473-476.

Wuyts, J., G. Perriere, and Y. Van De Peer. 2004. The European ribosomal RNA database. Nucleic Acids Res **32**:D101-103.

Yamaguchi, H., G. Minopoli, O. N. Demidov, D. K. Chatterjee, C. W. Anderson, S. R. Durell, and E. Appella. 2005. Substrate specificity of the human protein phosphatase 2Cdelta, Wip1. Biochemistry **44**:5285-5294.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367-370.

Yang, Z. 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39**:306-314.

Yang, Z. 1994b. Estimating the pattern of nucleotide substitution. J Mol Evol **39**:105-111.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol **10**:1396-1401.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol **11**:316-324.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol **14**:717-724.

Yeh, T. C., and S. Pellegrini. 1999. The Janus kinase family of protein tyrosine kinases and their role in signaling. Cell Mol Life Sci **55**:1523-1534.

Zanotto, P. M., E. G. Kallas, R. F. de Souza, and E. C. Holmes. 1999. Genealogical evidence for positive selection in the nef gene of HIV-1. Genetics **153**:1077-1089.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J Mol Evol **39**:315-329.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

Appendix A: Evolution of the Metazoan Protein Phosphatase 2C Superfamily

This appendix is based on a published manuscript:
Stern, A., Privman, E., Rasis, M., Lavi, S., Pupko, T. 2007. *J. Mol. Evol.* 64: 61-70

A.1 Introduction

The development of improved accuracy phylogenetic methods led me to a collaboration with the laboratory of Prof. Lavi, that studies protein phosphatase 2C superfamily. Reversible protein phosphorylation is a major regulatory mechanism of cellular functions such as stress-activated signal transduction, mitogenic signal transduction, and cell cycle control (e.g., Sun and Tonks 1994; Hanada et al. 2001). Protein kinases have been in the spotlight for several decades (Burnett and Kennedy 1954; Bolen 1995; Shokat 1995; Yeh and Pellegrini 1999; Schlessinger 2000; Cowan and Storey 2003) leading to the recent development of drug therapy using kinase inhibitors (reviewed in Krause and Van Etten 2005). Yet far less focus has been given to the kinases counterparts in cell regulation, the protein phosphatases. It is becoming clear that the interplay between kinases and phosphatases is quite complex. Thus, in order to fully understand the process of phosphorylation, it is imperative to focus research on phosphatases as well as on kinases.

Protein serine/threonine phosphatases are divided into four structurally distinct superfamilies (Cohen 1989; Shenolikar 1994; Barford, Jia, and Tonks 1995; Wera and Hemmings 1995): PP1, PP2A, PP2B, and PP2C. The PP2C superfamily of phosphatases (also referred to as PPM) is

defined by distinct amino acid sequence and three-dimensional (3D) structure (Tamura et al. 1989; Mann et al. 1992; Wenk et al. 1992). The PP2C superfamily does not seem to be evolutionary related to PP1, PP2A and PP2B, which are all multi-subunit enzymes. This study focuses on the PP2C superfamily as an embodiment of the evolutionary diversity of protein phosphatases.

A.1.1 PP2C Functions

PP2C is a monomeric enzyme which displays broad substrate specificity. Distinguishing characteristics of PP2C are its: i) absolute requirement for divalent cations, mainly Mg^{+2} or Mn^{+2}, ii) distinctive structural features, iii) insensitivity to inhibition by Okadaic acid (Barford, Jia, and Tonks 1995; Wera and Hemmings 1995) . At least 15 distinct PP2C human paralogs have been found in mammalian cells (Table A.1). All of these PP2Cs have Mg^{+2} and/or Mn^{+2} dependent phosphatase activity against artificial substrates *in vitro* (Komaki et al. 2003).

The majority of PP2Cs are involved in regulation of stress activated protein kinase (SAPK) cascades which relay signals in response to external stimuli (Meskiene et al. 2003). These cascades are a subfamily of the mitogen–activated protein kinase (MAPK) cascades. Different PP2Cs negatively regulate SAPK pathways at different levels. For instance, PP2Cβ inhibits the TAK1 pathway (Hanada et al. 2001; Li et al. 2003) and is involved in the NF-kappaB pathway (Prajapati et al. 2004). PP2Cα inactivates the p38 pathway and the c-Jun amino-terminal kinase (JNK) pathway (Takekawa, Maeda and Saito, 1998), and is additionally involved in the Wnt signalling pathway (Strovel, Wu, and Sussman 2000).

Recent studies have demonstrated that the PP2C superfamily is also associated with eukaryotic cell cycle processes, which are controlled by the ordered activation and inactivation of cyclin-dependent protein kinases (CDKs). Reversible protein phosphorylation is one of the mechanisms through which extra-cellular and intra-cellular signals regulate CDKs (Cheng et al. 1999; Cheng, Kaldis, and Solomon 2000). Additionally, we have previously reported that over-expression of PP2Cα activates the expression of the tumour suppressor gene TP53/p53, which leads to G2/M cell cycle arrest and apoptosis (Ofek et al. 2003). Thus, the PP2C superfamily appears to be directly involved in several cell regulation and cell signalling processes. In fact, many PP2C members have been reported to inactivate CDK and MAPK family kinases by dephosphorylating a conserved threonine residue on the so called T-loop of these kinases (Marley et al. 1996; Cheng et al. 1999; Takekawa et al. 2000), implicating that the PP2C superfamily may be general T-loop phosphatases.

Table A.1: The 15 human PP2C paralogs and their known cellular function

PP2C variants[(a)]	***Function \ Cellular process***	***Expression***[(b)]	***Reference***
PP2Cα P35813	- Negative regulation of p38-MAPK and JNK cascades via dephosphorylation of MKK4, MKK6 and p38 - Involvement in cell cycle regulation via interaction with CDKs - Activation of the p53-pathway (cell cycle arrest)	Ubiquitous	(Takekawa, Maeda, and Saito 1998; Cheng, Kaldis, and Solomon 2000; Ofek et al. 2003)
PP2Cβ O75688	Negative regulation of p38-MAPK and JNK cascades via dephosphoralytion of TAK1	Skeletal muscle↑ Heart↑ (of specific splice variants)	(Terasawa et al. 1993; Gaits, Shiozaki, and Russell 1997; Marley et al. 1998; Hanada et al. 2001; Seroussi et al. 2001)
PP2Cγ O15355	Necessary for the formation of a functional spliceosome	Widely expressed. Testis, skeletal muscle, and heart↑	(Travis and Welsh 1997; Murray, Kobayashi, and Krainer 1999)
Wip1 O15297	Mediation of a negative feed-back loop of the p38-MAPK-p53 pathway: inactivation of p53, relief of cell-cycle arrest		(Takekawa et al. 2000; Yamaguchi et al. 2005)
PP2Cε Q5SGD2	A possible role of PP2Cε is in the regulation of the IL-1-TAK1 signaling pathway		(Li et al. 2003)
PP2Cζ Q810X6	The exact function of this gene is not yet known. It has been shown that PP2C ζ is able to associate with ubiquitin conjugating enzyme 9	Testis↑	(Kashiwaba et al. 2003)
PP2Cη Q96M16	Putative nuclear localization signal (NLS) suggests that PP2C η dephosphorylates a unique substrate(s) in the cell nucleus		(Komaki et al. 2003)
POPX1 Q7LAF3	Inactivation of the p21 (Cdc42/Rac)-activated kinase PAK	Brain↑ Testis↑	(Koh et al. 2002)

PP2C variants[(a)]	*Function \ Cellular process*	*Expression*[(b)]	*Reference*
FEM2/ POPX2 P49593	- Inactivation of the p21 (Cdc42/Rac)-activated kinase PAK - Activation of Calmodulin-dependent kinase II - Promotion of apoptosis	Ubiquitous	(Koh et al. 2002)
ILKAP Q9H0C8	Modulation of cell adhesion and growth factor signaling through association with integrin linked kinase		(Leung-Hagesteijn et al. 2001)
TA-PP2C NP_644812.1	Co-expression with IL-2 in T-cells		(Mao et al. 2004)
PPM1K Q56AN8	Function unknown. Annotated as PP2C mitochondrial phosphatase	Mitochondria	
PPM1H/ NERPP Q6PI86	Signal transduction pathway for neuronal inhibitory factors in CNS myelin	Brain↑	(Labes, Roder, and Roach 1998)
PDP1 Q9P0J1	Dephosphorylation and concomitant reactivation of the alpha subunit of the E1 component of the pyruvate dehydrogenase complex	Skeletal muscle↑	(Behal et al. 1993; Lawson et al. 1997; Huang et al. 1998)
PDP2 Q9P2J9	Dephosphorylation and concomitant reactivation of the alpha subunit of the E1 component of the pyruvate dehydrogenase complex	Liver↑	(Huang et al. 1998)

[(a)] The first row shows the name of the human gene while the second row shows the SWISS-PROT (http://us.expasy.org/sprot/) or GenBank (Benson et al.; Guindon et al. 2005) identifiers.
[(b)] ↑ indicates up-regulation in the specified tissue. Cells were left blank when no expression data were available for the gene.

A.1.2 Phylogenetic Study of PP2C

An evolutionary comparison of kinomes across species (Manning et al. 2002) has demonstrated the value of a phylogenetic study of kinases. This enabled mapping kinases specific to each

lineage, as well as delineating the pathways involving kinases shared throughout various evolutionary lineages. Similar to families of protein kinases, the multiplicity of PP2C proteins suggests a broad functional diversity of these proteins. Thus, the aim of this study is to conduct a comprehensive genomic evolutionary analysis of all known members of the PP2C superfamily in Metazoa. This is the first analysis of its kind undertaken in the study of the PP2C superfamily, and as such promises to be highly informative in characterizing protein isoform diversification. Consequently, we performed an extensive search of all genomic databases for PP2C genes. We then conducted a phylogenetic analysis of all PP2C members found, with the aim of assigning paralogy and orthology relations to each sequence found. These assignments enabled us to predict functions of previously unidentified genes in the superfamily, as well as explore the differences between the families within the PP2C superfamily and estimate the relative dates of the diversification events. We thereby explore the breadth of PP2C functional conservation throughout the metazoan kingdom.

A.2 Methods

A.2.1 Search for PP2C Members in Metazoa

Sequences were retrieved from the following databases: Genbank (www.ncbi.nlm.nih.gov) (Benson et al. 2005), ENSEMBL (www.ensembl.org) (Hubbard et al. 2005), FLYBASE (www.fruitfly.org) (Stapleton et al. 2002) and WORMBASE release WS130 (www.wormbase.org) (Harris et al. 2004). Initially, all fully sequenced eukaryote genomes (*Homo sapiens, Pan troglodytes, Mus musculus, Rattus norvegicus, Canis familiaris, Bos taurus, Monodelphis*

domestica, Gallus gallus, Xenopus tropicalis, Danio rerio, Takifugu rubripes, Tetraodon nigroviridis, Drosophila melanogaster, Anopheles gambiae, Apis mellifera, Caenorhabditis elegans, Saccharomyces cerevisiae) were screened for PP2C sequences using BLAT (Kent 2002), TBLASTN (Altschul et al. 1990), and the Orthologue Prediction section of ENSEMBL. Sequences uncovered in these stages were filtered according to two criteria: (i) due to an unreliable alignment containing excess gaps, all yeast sequences were discarded, (ii) orthologous sequences for which there was a significant deviation from the PP2C signature defined in PROSITE (Bairoch and Bucher 1994; entry PS01032) were discarded. Accession numbers of all sequences used in the study are available as supplementary material.

A.2.2 Sequence Alignment and Phylogenetic Reconstruction

Multiple sequence alignment (MSA) was performed using the MUSCLE program version 3.52 . Maximum likelihood based phylogenetic reconstruction was performed with the PhyML program (Halanych 2004) using among site rate variation with 4 discrete rate categories, and the JTT model of sequence evxolution. Node supports were determined by performing 100 bootstrap replicates. Bootstrap values higher than 70% were considered significant.

Due to the extended evolutionary time spanned by the sequences in this analysis and the low similarity between different paralogs, the resulting alignment includes regions which are difficult to align as well as many gaps. This is a general problem when analysing superfamily genes, and one must verify that the inferred phylogeny does not depend on those parts of the alignment which are uncertain. In order to test the robustness of the results to the alignment's

validity, we ran the Gblocks program on the alignment, with the aim of removing poorly aligned regions, and then performed the phylogenetic analysis on the resulting positions. The new tree was found to be essentially identical to the tree which we reported based on the full alignment. All minor differences between the trees were supported by low bootstrap values in the new tree. The settings used to run Gblocks, together with the resulting reduced alignment and resulting phylogenetic tree are given in the supplementary material.

A specific sequence was considered part of a group if it belonged to a monophyletic clade in the tree. For this purpose, it was assumed that the root of the tree is on one of the branches leading to one of the more anciently derived groups (as is later defined in the Results section).

A.2.3 Site-Specific Evolutionary Rate Analysis

The conservation pattern of the PP2C superfamily was analyzed by estimating site-specific evolutionary rates, using the Bayesian approach of the Consurf server (Mayrose et al. 2004; Landau et al. 2005). The analysis was conducted using the reconstructed PhyML tree. Site-specific positive selection was analysed using the Selecton server (Doron-Faigenboim et al. 2005), using the Bayesian method (Yang et al. 2000). Once again, the reconstructed PhyML tree was given as input.

A.3 Results

The search for PP2C members in the human genome led to the identification of 15 members (table A.1). Each of these members was found throughout the vast majority of the vertebrate genomes searched, with the exception of a few orthologs not found. It is unclear whether these exceptions stemmed from incomplete sequencing or from gene losses in these organisms. All the 15 groups of orthologs clustered as monophyletic groups in the phylogenetic tree reconstructed (Figure A.1), providing firm phylogenetic support for the hypothesis whereby the PP2C superfamily arose following a series of duplication events, and for the classical classification of the PP2C family members. Furthermore, the phylogenetic tree enabled assigning annotation to previously unidentified genes according to their location on the phylogenetic tree.

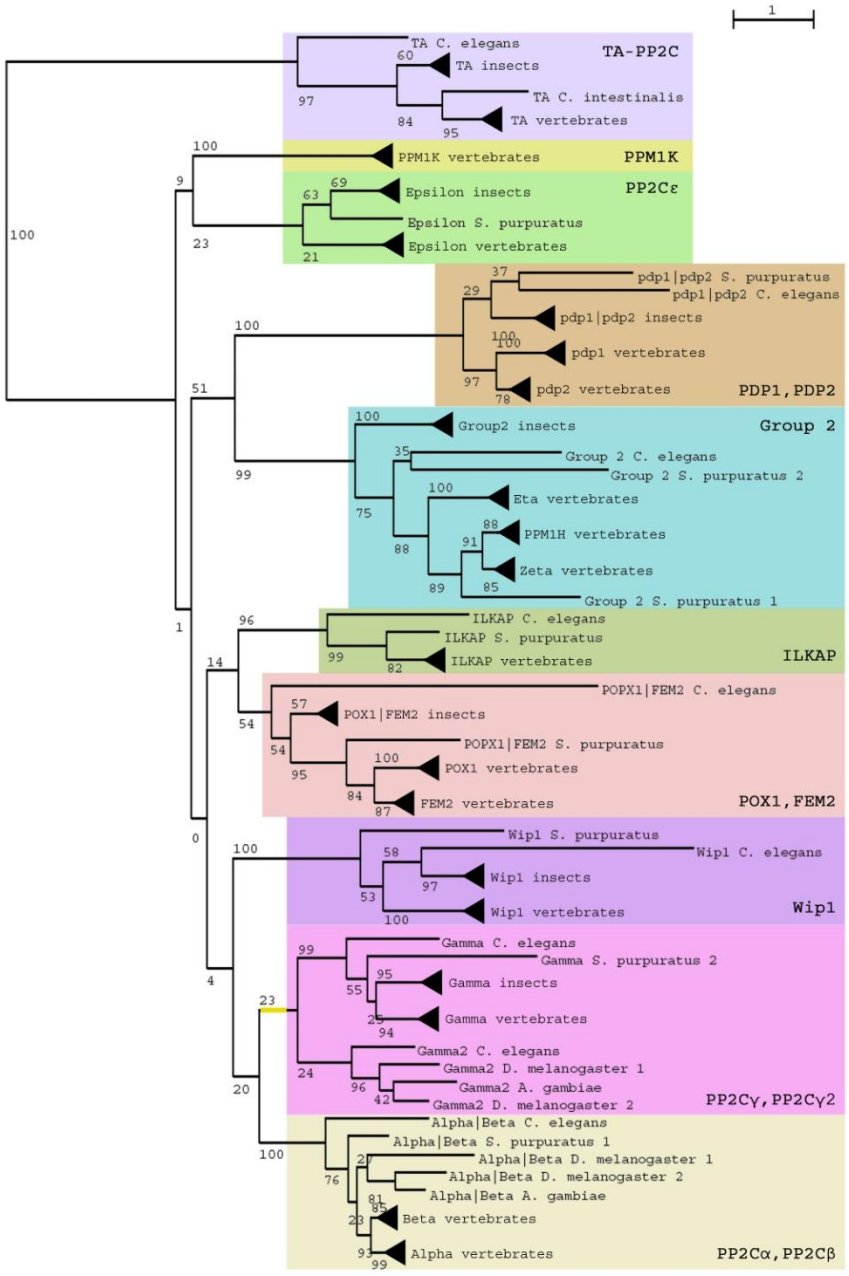

Figure A.1: **Maximum likelihood phylogenetic tree of PP2C, with the TA group used as an outgroup.** For brevity, the tree presented here shows a collapsed version of the full tree, where all clades representing vertebrate orthologs were collapsed, as were all clades of orthologous insects (collapsed clades are shown as black triangles). The lengths of the branches leading to the collapsed groups are the lengths of the branches leading to the ancestral groups in the original tree. Major PP2C groups are shaded in color for clarity. The branch marked in yellow represents a putative gene duplication event with subsequent gene loss in vertebrates in one of the duplicants. A full tree in Newick format is available in the supplementary material.

A.3.1 Two Types of PP2C

Mammalian PP2Cs have been previously classified into two subgroups according to differences in amino acid sequence motifs. Group 2 consists of PP2Cη, PP2Cζ, and PPM1H, while group 1

includes all other PP2Cs (Komaki et al. 2003). TA-PP2C, discovered only a year later (Mao et al. 2004), was not classified as belonging to any of the two groups . Groups 1 and 2 differ in their PP2C signature, as well as in the other sequential motifs found to characterize the PP2C superfamily (Komaki et al. 2003). All residues forming the catalytic domain (Das et al. 1996) are part of the PP2C signature and these additional motifs.

The phylogenetic reconstruction of the PP2C family (Figure A.1) shows that the three families belonging to group 2 (PP2Cη, PP2Cζ, and PPM1H) form a monophyletic clade (100% bootstrap), supporting the previous finding whereby group 2 is characterized by a unique sequence (Komaki et al. 2003). Since the differences between groups 1 and 2 are displayed in residues which surround the catalytic site, this may hint at a functional divergence of the PP2C superfamily into two functionally distinct groups.

A.3.2 Mapping of PP2C Duplications

Nine different clades in the tree comprise members in protostomes. These sequences include sequences belonging to monophyletic PP2C groups (TA-PP2C, ILKAP, PP2Cγ, Wip1, and PP2Cε), as well as sequences assumed to have been derived from ancestral forms of PP2C prior to their duplication. For example, a worm sequence which forms an outgroup of the PDP1 and PDP2 vertebrate clades is assumed to be derived from the ancestral sequence of PDP1 and PDP2 prior to their duplication, and will hereby be referred to as the PDP1|PDP2 ancestrally derived sequence. Similarly, there is a protostome ancestrally derived sequence of POPX1|FEM2, PP2Cα|PP2Cβ, group 2 (PP2Cη|PP2Cζ|PPM1H), and PDP1|PDP2. Consequently, it appears that

much of the PP2C diversification occurred before bilaterian diversification. On the other hand, the nine paralogous groups PP2Cα, PP2Cβ, POPX1, FEM2, PP2Cζ, PPM1H, PPM1K, PDP1, and PDP2 are only present in vertebrates. Therefore, the mapping of the emergence of each one of the PP2C groups during evolution is possible, as depicted in Figure A.2. Nine groups emerged prior to the divergence of protostomes, whilst another nine were created by duplications prior to the divergence of vertebrates.

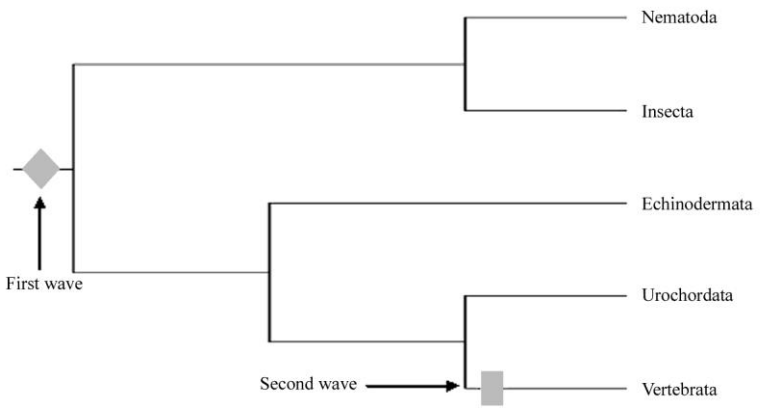

Figure A.2: The two active periods of gene diversification by duplication are marked on the currently accepted species tree by arrows. The rectangle represents the putative emergence of the groups: PP2Cα, PP2Cβ, PPM1K, FEM2, POPX1, PP2Cζ, PPM1H, PDP1, and PDP2. The diamond represents the latest possible dating of the emergence of the groups: Wip1, ILKAP, PP2Cγ, TA-PP2C, PP2Cε , PP2Cα|β, PDP1|PDP2, group 2, and FEM2|POPX1.

More precise mapping and relative dating of the different gene duplication events proved to be more difficult. A majority of the more ancient duplications were supported by very low

bootstrap values on the tree (Figure A.1). Thus, it is currently impossible to map which members were created by each ancient duplication event, and to determine the order of duplications. This is further aggravated by lack of available genomes, specifically in non-vertebrate chordates and poriferans. However, the formation of four more recent duplications was supported by very high bootstrap values. These duplications (PP2Cα versus PP2Cβ, PDP1 versus PDP2, PP2Cζ versus PPM1H, and POPX1 versus FEM2) occurred before vertebrate diversification, and are supported by a single form in non-vertebrates (sea urchin and/or insects). For example, the clades of PDP1 and PDP2 do not include any sea urchin sequences, yet a sea urchin sequence forms an outgroup. An additional duplication, between PP2Cη and the PP2Cζ|PPM1H ancestor, is more difficult to map precisely. According to the tree, the duplication occurred after the speciation of insects, and before the speciation of vertebrates. However, it is unclear whether this duplication occurred before or after the speciation of sea urchin, due to the location of the two sea urchin sequences (Figure A.1; denoted as Group 2 S. purpuratus 1 and 2).

A.3.3 Protostome-Specific Duplications

The phylogenetic tree supports an ancient duplication event in the ancestral PP2Cγ (Figure A.1; marked in yellow) where the gene was apparently lost in the lineage leading to vertebrates, yet remained in protostomes. This novel paralog further underwent another duplication in insects. Yet another insect-specific duplication is evident, as is apparent by the existence of both Alpha|Beta *D. Melanogaster* 2 and Alpha|Beta *D. Melanogaster* 1 (see Figure A.1). In both

these insect-specific duplications, no sequence is available from *A. Gambiae*, which may indicate other a loss in this lineage or missing data in the *A. Gambiae* genome. Little is known of these expansions in insects, yet they are supported by EST evidence in *D. Melanogaster* (UCSC genome browser; (Kent et al. 2002)), ruling out the hypothesis whereby these sequences represent pseudogenes.

A.3.4 Mapping Functional Regions in PP2C

We performed a comprehensive analysis of the pattern of amino acid conservation throughout the MSA of the PP2C superfamily. This analysis took into account the phylogenetic tree, thus enabling more precise inference of conservation of amino acid sites (Pupko et al. 2002). Two crystal structures of PP2C members exist in the Protein Data Bank (Berman et al. 2000) – one of the human PP2Cα (Das et al. 1996) and one of the Mycobacterium Tuberculosis PP2C (Pullen et al. 2004). The conservation pattern obtained for the PP2C superfamily was mapped onto the Van-der-Waals surface of PP2Cα, since it is the only metazoan PP2C with a known 3D structure. A clear pattern of high conservation is apparent throughout the N-terminus of the protein, whilst the C-terminus of the protein is highly variable. The variability of the C-terminus is expected, since the C-terminus of the PP2Cα is unique to this family, and may serve as a substrate recognition domain in the cleft that is created between it and the catalytic domain (Das et al. 1996). However, the relatively high conservation pattern of the N-terminus is more surprising. On the one hand, the results reinforce the previous knowledge of functionally important sites in PP2C, showing that the nine catalytic residues found in PP2Cα (Das et al.

1996) are indeed highly conserved. More surprisingly, over 50 additional sites appeared to be highly conserved. Whilst some of these sites cluster around the catalytic site in the globular N-terminus, others form the cleft between the N-terminus and the C-terminus and form part of the bulk above the catalytic region (according to the orientation in Figure A.3). The high conservation of the N-terminal part of the cleft suggests that there may be a shared mechanism of this cleft throughout all the PP2C families.

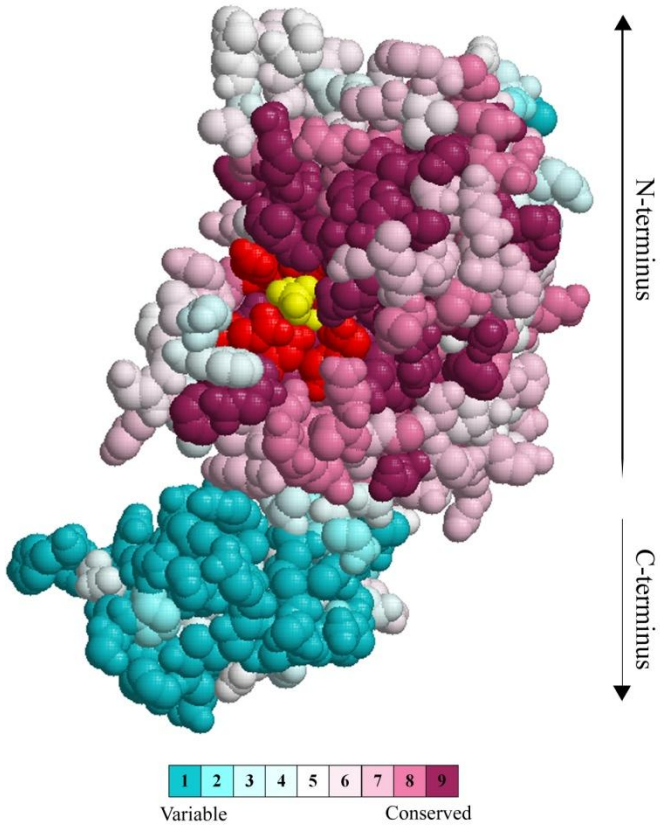

Figure A.3: The conservation pattern of the PP2C superfamily as inferred by Consurf. Conservation scores are color-coded onto the Van der Waals surface of PP2Cα, where bordeaux corresponds to maximal conservation, white corresponds to average conservation and turquoise corresponds to maximal variability. The Mg^{2+} ions and associated water molecules are shown in yellow, and the nine previously identified catalytic sites (Das et al. 1996) are shown in red. These nine sites also attained the highest level of conservation in the analysis. Arrows show the globular N-terminus region, and the C-terminus tail.

In order to study the differences between the PP2C families, pairs of PP2C families were analyzed for site-specific positive selection using the Selecton server (Doron-Faigenboim et al. 2005). The underlying assumption was that following the gene duplication events which

created the two families, both genes underwent a specialization process. Such a process may have led to a rapid fixation of mutations due to positive selection forces. Thus, all pairs of PP2C families which most recently diverged (PP2Cα-PP2Cβ, PDP1-PDP2, FEM2-POPX1, PPM1H-PP2Cζ) were analysed. The analysis of pairs of sequences as opposed to an analysis of the entire superfamily enables a more reliable MSA at the codon level. However, in all of these families no site-specific or global positive selection was observed. This may be due to the fact that purifying selection within each of the families obscures the footprint of positive selection which the families underwent, and due to the small species sampling.

Additional evidence for gene specialization may be obtained by analysing insertion and deletion events in the different gene families. To this end, the MSA was utilized as a rough indicator of insertion and deletion events (using visual inspection). A schematic drawing was created depicting PP2C domains common to all groups, as well as those unique to specific groups (Figure A.4). Only such blocks which are well defined and clearly observed when viewing the alignment were depicted. Furthermore, these blocks were flanked by anchors of regions which are conserved throughout the entire alignment. This analysis suggests that the evolution of the PP2C family included several significant insertion or deletion events which may have led to the specialization of the duplicants.

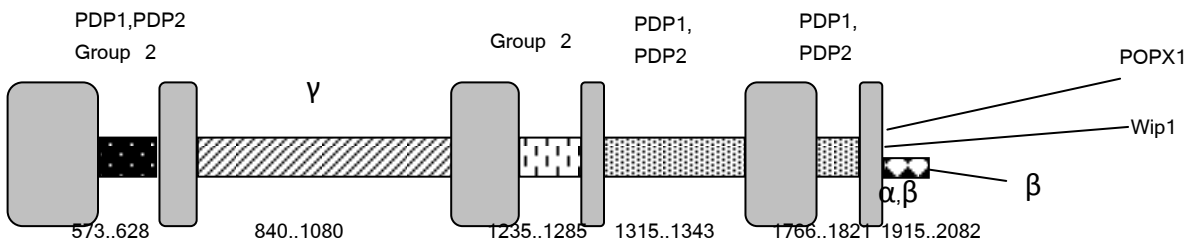

Figure A.4: A schematic drawing of the PP2C superfamily alignment. Bars represent blocks which are unique to specific groups. Grey rectangles represent the rest of the alignment, including all regions which are common to all PP2C proteins. Coordinates of the blocks (according to the MSA) appear underneath each block. The MSA is available as part of the supplementary material.

A.4 Discussion

In this study, we present an analysis of PP2C evolution. The reconstructed phylogeny displays the relationship between PP2C paralogs and orthologs throughout Metazoa, revealing the existence of at least 15 PP2C groups which were created via gene duplication. Analysis of the PP2C superfamily suggests that two waves of duplications were responsible for the creation of the majority of the PP2C members. The first wave of duplications presumably led to the formation of functionally different groups which specialized in different catalytic processes. Our analysis suggests that this wave took place before the divergence of bilaterians. Presumably, these ancient duplications occurred successively in a short time frame (represented by short branches between the different groups; see Figure A. 1), rendering it difficult to determine the order of the duplication events (as is evident from the low bootstrap support on these branches). The second wave of duplications presumably led to the formation

of tissue specific groups, and most likely took place at the beginning of vertebrate evolution. More precise timing of these duplication events is still beyond reach due to the lack of sequence data in poriferans.

Extensive gene duplication events during early chordate evolution have been previously reported (e.g. Miyata and Suga 2001; McLysaght, Hokamp, and Wolfe 2002). These extensive events may have been the result of a whole genome duplication, or a series of partial chromosomal block duplications. These duplications, which took place in early vertebrate evolution, are thought to account for the existence of the four HOX gene clusters (Larhammar, Lundin, and Hallbook 2002; Prince 2002), as well as for the four different MHC gene clusters (Abi-Rached et al. 2002). Numerous studies have found protein families with a pattern of evolution similar to the pattern of PP2C found here (e.g. Miyata and Suga 2001; Wakeham et al. 2005). Many of these families include kinase and phosphatase families involved in cell signaling. Furthermore, a comparative study comparing kinome catalogs throughout different species (Manning et al. 2002) revealed that the creation of functionally distinct kinase families occurred during early metazoan evolution. Thus, we postulate that the evolution of signal transduction occurred in two major active periods. The first, occurring before metazoan radiation, may have been driven by the increase in complexity of multicellular organisms. This required more sophisticated signaling between and within cells. Furthermore, the more complex developmental mechanisms also required a more elaborate network of signaling proteins. This suggests that the more anciently derived PP2Cs evolved following a requirement for new

signaling pathways. For example, the more anciently derived Wip1 and ILKAP display this pattern: whereas Wip1 evolved as part of the cell-cycle pathway, ILKAP evolved to participate in cell adhesion and growth factor signaling.

The second active period, before vertebrate diversification, occurred concomitantly with the development of tissues such as skeletal muscle and the nervous system. Indeed, the transition between non-vertebrates and vertebrates is believed to be one of the major leaps in complexity during evolution, involving the evolution of cells such as the neural crest, the brain, and the spinal cord (Gilbert 2001). Insights into this phenomenon are evident throughout all PP2C groups found to have diverged in this second active period. All PP2C vertebrate-specific duplications characterized in this study show specialized tissue-specific expression patterns. In fact, in each of these duplications, one duplicant is uniquely expressed in either skeletal muscle or nervous system tissue. Whereas PP2Cα and FEM2 are ubiquitous in the cell, their duplication partners PP2Cβ and POPX1 are tissue specific. PP2Cβ splice variants were shown to display skeletal muscle and heart specific tissue expression (Marley et al. 1998; Seroussi et al. 2001). Similarly, POPX1 displays brain specific expression (Koh et al. 2002). PP2Cζ is displayed in the testicular germ cells (Kashiwaba et al. 2003), while PPM1H is expressed in the brain and is involved in neuronal inhibitory pathways (Labes, Roder, and Roach 1998). Finally, PDP1 and PDP2 both catalyze the same reaction, but differ in tissue distribution. PDP1 is highly expressed in skeletal muscle, whereas PDP2 is expressed in liver (Huang et al. 1998).

The evolutionary conservation pattern of PP2C suggests that the highly conserved catalytic domain and the surrounding core are shared by all PP2C members. Of specific interest are those residues which have not been previously identified as critical to the enzymatic function of the protein. Since these sites are abundant on the protein surface, they may represent novel protein or ligand binding sites. As the PP2C superfamily plays an important role in delicate signals relayed across the cell, it is likely that the different PP2C proteins are further bound by tight regulation. Thus, the novel conserved sites found may be the key to understanding these regulatory mechanisms.

Aside from the shared conserved domain, several PP2C families have a unique appendage which may be the specificity determinant of this family. For instance, in the C-terminal region, PP2Cα and PP2Cβ share a unique tail (Figure A.4). These differences may be responsible for the fact that the two paralogs differ in their substrate binding abilities. Further investigation of the differences between the families could also focus on the gene-regulation level, for instance by comparing the promoters and non-translated regions of the different paralogs. We have previously reported that the non-translated regions of PP2Cβ are highly conserved throughout orthologous genes (Seroussi et al. 2001), indicating a significant regulatory role for these regions. Furthermore, different PP2Cβ transcripts were found to differ by alternative splicing and alternative promoters (Ohnishi et al. 1999). It thus seems that the diversification which gave rise to new PP2C families continued with the creation of variants which differ at the transcriptional level. Pinpointing the precise functional differences among and within the PP2C

families remains a challenge which may play a pivotal role in our understanding of complex signal networks in cells.

5 לפיתוח שיטה חדשה למדידת החסינות של כל עמודה ב-MSA בפני חוסר הוודאות שבעץ-המדריך. בחינה באמצעות מאגרי רצפים המשמשים אמת מידה (benchmark) לדיוק של שיטות עימוד מגלה שהמדד החדש מזהה במדויק אזורים בלתי-מהימנים ב-MSA ומאפשר את הפרדתם. פרק 6 מתאר מימוש המדד החדש בשרת GUIDANCE, המאפשר זיהוי מדויק של טעויות עימוד, כמו גם כלים לניפוי והתמודדות עם טעויות אלה. שרת זה מספק לחוקרים אמצעים מונעים המאפשרים הגנה על מחקרים מבוססי-MSA מפני הנזקים הנגררים בעקבות עימוד שגוי.

במהלך כל המחקרים הללו השיפורים האלגוריתמיים המוצעים נבחנים לפי מאגרי benchmark מקובלים של רצפים אמיתיים ומדומים (simulation studies) גם יחד. המתודולוגיות המפותחות מאפשרות שימוש במודלים הסתברותיים מתקדמים של אבולוציה של רצפים בשילוב עם האלגוריתמים המובילים לפילוגנזה ועימוד רצפים. למרות זאת, ניתן דגש לתכנון אלגוריתמי יעיל המאפשר ניתוח כמויות הנתונים המיוצרים ע"י טכנולוגיות הריצוף המתקדמות במהירות. מתוך כך ניתן יהיה לנתח ברמת דיוק גבוהה אוספים גדולים המונים אלפים רבים של רצפים, אשר בעבר ניתן היה לנתח רק באמצעות שיטות פשטניות. החידושים האלגוריתמיים נבנו ככלים מודולאריים כדי לאפשר את שילובם עם פיתוחים מתודולוגיים מקבילים, והם מופצים בקרה הקהילה המדעית. השיטות שפותחו במחקר זה משתלבות ביסודות המתודולוגיה לניתוח משווה של רצפים וצפויות לתרום לדיוק ומהימנות של מחקרים עתידיים בביולוגיה מולקולארית.

תקציר

מהפכה אחר מהפכה בטכנולוגיות ריצוף אפשרו צבירת אוספים גדולים של רצפי דנ"א הומולוגיים. כיום, מאמץ מרכזי בחקר ביולוגיה מולקולארית מכוון לניצול מאגרי המידע הללו ע"י ניתוח משווה, לשם הסקת תובנות לגבי התפקידים הביולוגיים של הרצפים. טווח רחב של ניתוחים השוואתיים של רצפים, החל מפילוגנזה מולקולארית וכלה בחיזוי מבני חלבונים תלת-ממדיים, נבנים על עימוד רצפים, Multiple Sequence Alignment (MSA), ושיחזור עצים פילוגנטיים כמבני הנתונים היסודיים בניתוח רצפים. אלגוריתמים מורכבים פותחו עבור שתי המשימות החישוביות הללו, אך בפועל לוקים אזורים נרחבים של ה-MSA וגם של העץ בחוסר מהימנות. הקשיים בכל אחת משתי הבעיות בנפרד מוגברים ע"י התלות ההדדית בינן כיוון שהן מזינות זו את זו ונבנות זו על זה – עצים משמשים להדרכת תהליך העימוד, ואילו שיטות לשחזור עצים מסתמכות על MSA. התלות ההדדית הזו יוצרת בהכרח שרשרת של העברת טעויות הלוך וחזור בין שני השלבים האלו בניתוח רצפים.

המחקרים האגודים בעבודת הגמר הזו מתמודדים עם האתגרים של שיפור הדיוק בשחזור עצים ועימוד רצפים, ושל בחינת העברת השגיאות ההדדית בין השניים. בפרק 3 מפותחות שיטות מעורבות לשחזור עצים בכדי לשלב את היתרונות של מידול אבולוציוני מדויק באמצעות שיטות בייזיאניות (Bayesian) יחד עם אלגוריתמים מבוססי-מרחק יעילים. תרומה משמעותית לדיוק מושגת באמצעות שימוש במודלים של שונות בקצב האבולוציוני, ובהמשך במודלים מתקדמים יותר, דמוי-covarion. יישומן של שיטות אלו מודגם תוך שחזור וניתוח שתי פילוגנזות לדוגמא – עץ מינים של חיידקים ועץ של משפחת אנזימים בבעלי חיים. בהמשך, בפרק 4 משמשת תבנית חוזרנית לחקר התרומה של עץ-מדריך משופר לדיוק של תוכנות עימוד שונות.

לצד המאמצים לצמצם טעויות בשחזור, חשוב לא פחות להבין ולאפיין את גורמי השגיאות השונים שנותרים. ניתוח התלות ההדדית מראה שחוסר וודאות בעץ-המדריך (guide-tree) המשמש שיטות לעימוד הדרגתי של רצפים (progressive sequence alignment) הם מקור עיקרי לחוסר וודאות בעימוד. תובנה זאת מובילה בפרק

שיטות מעורבות בהשראת התלות ההדדית בין

עימוד רצפים ושחזור עצים פילוגנטיים

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

מאת

איל פריבמן

הוגש לסנאט של אוניברסיטת תל-אביב

יוני, 2010