ERROR AND UNCERTAINTY IN COMPUTATIONAL PHYLOGENETICS

by

VICTOR HANSON-SMITH

A DISSERTATION

Presented to the Department of Computer and Information Science
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2011

DISSERTATION APPROVAL PAGE

Student: Victor Hanson-Smith

Title: Error and Uncertainty in Computational Phylogenetics

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Computer and Information Science by:

John Conery	Chair
Daniel Lowd	Member
Sara Douglas	Member
Joseph W. Thornton	Outside Member

and

Kimberly Andrews Epsy	Vice President for Research and Innovation/ Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2011

DISSERTATION ABSTRACT

Victor Hanson-Smith

Doctor of Philosophy

Computer and Information Science

December 2011

Title: Error and Uncertainty in Computational Phylogenetics

The evolutionary history of protein families can be difficult to study because necessary ancestral molecules are often unavailable for direct observation. As an alternative, the field of computational phylogenetics has developed statistical methods to infer the evolutionary relationships among extant molecular sequences and their ancestral sequences. Typically, the methods of computational phylogenetic inference and ancestral sequence reconstruction are combined with other non-computational techniques in a larger analysis pipeline to study the inferred forms and functions of ancient molecules. Two big problems surrounding this analysis pipeline are computational error and statistical uncertainty. In this dissertation, I use simulations and analysis of empirical systems to show that phylogenetic error can be reduced by using an alternative search heuristic. I then use similar methods to reveal the relationship between phylogenetic uncertainty and the accuracy of ancestral sequence reconstruction. Finally, I provide a case-study of a molecular machine in yeast, to demonstrate all stages of the analysis pipeline.

This dissertation includes previously published co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Victor Hanson-Smith

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Seattle University, Seattle, WA

DEGREES AWARDED:

Doctor of Philosophy in Computer and Information Science, 2011, University of Oregon
Master of Science in Computer and Information Science, 2007, University of Oregon
Bachelor of Science in Computer Science and Software Engineering, 2003, Seattle University

AREAS OF SPECIAL INTEREST:

Evolution, genomic systems, high-performance computing, scientific software engineering

PROFESSIONAL EXPERIENCE:

Graduate Research Fellow, National Science Foundation IGERT program in evolution, development, and genomics, 2008-2011

Graduate Teaching Assistant, Department of Computer and Information Science, University of Oregon, 2007-2009

Graduate Research Fellow, University of Oregon Neuroinformatics Lab, 2006-2007

Software Architect, Authora Inc., Seattle, WA, 2003-2005

Undergraduate Fellow, Microsoft Corp., Seattle, WA, 2002-2003

Undergraduate Fellow, Intel Corp., Folsom, CA, 2000

GRANTS, AWARDS AND HONORS:

National Science Foundation IGERT training grant, 2008 - 2011

Intel Foundation Science and Engineering Scholar, 1999 - 2003

PUBLICATIONS:

G.C. Finnigan, V. Hanson-Smith, T.H. Stevens, and J.W. Thornton. Evolution of increased complexity in a molecular machine *In press, Nature*, 2011.

G.C. Finnigan, V. Hanson-Smith, B.D. Houser, H.J. Park, and T.H. Stevens. The reconstructed ancestral subunit a functions as both V-ATPase isoforms vph1p and stv1p in S. Cerevisiae. *Molecular Biology of the Cell*, 22(17), July 2011.

V. Hanson-Smith, B. Kolaczkowski, Bryan and J.W. Thornton. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*, 27(9), Apr 2010.

ACKNOWLEDGEMENTS

Special thanks to John Conery for facilitating interdisciplinary research at the University of Oregon between the departments of biology and computer and information Science. John's graduate-level courses in bioinformatics and computational science sparked my initial interest in this field of research. Thanks to Joe Thornton for allowing me to work in his lab, and for constantly challenging me to pursue excellence – it has made me a better thinker, writer, and scientist. Thanks to members of the Thornton Lab for constructive comments on my work, and for fostering an overall supportive and productive work environment over the past four years. Thanks to members of the Cresko and Phillips labs for comments.

For the experiments discussed in chapters 2 and 3, I thank Geeta Eick, Christopher Baker, Eric Gaucher, Belinda Chang, and Mikhail Matz for curating and sharing sequence data. Thanks to Katrina Ray for working with me to develop an early software prototype. For chapter 4, I thank Greg Finnigan for his collaboration on a challenging – but ultimately rewarding – project.

This work has been supported by the NSF IGERT training grant DGE-9972830 to the University of Oregon.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION TO COMPUTATIONAL PHYLOGENETICS

Over the last two centuries of human thought, the discovery of biological evolution profoundly changed – and continues to change – our perception of the living world (Darwin (1859); Huxley (1942); Lewontin (1972)). Combined with a modern understanding of molecular biology and genetic architecture, an evolutionary perspective allows us to investigate the genesis and function of diverse biological forms (Raff (1996); Carroll et al. (2005)). Further, an evolutionary perspective is useful to learn how the overall ecology of our planet is interconnected, and how our own bodies interact with that ecology. The study of evolution can be challenging, however, because many interesting and relevant biological systems evolved over timescales that are vastly longer than the length of a human lifetime. The challenge is that necessary ancestral forms are often unavailable for direct study because they existed millions – or billions – of years ago.

The Role of Computer Science in Evolutionary Studies

We can time travel, in a sense, using computational models of molecular evolution. The field of computational phylogenetics has developed Markov models to infer evolutionary history from contemporary molecular sequence data. These types of Markov models are used to reconstruct the phylogenetic history of gene families and to reconstruct ancestral gene sequences. These two *in silico* techniques – phylogenetic inference and ancestral reconstruction – are often combined with *in vivo* or *in vitro* molecular techniques to generate and then test hypotheses about

the evolutionary trajectory of protein families. This combination of computational analysis with wet-lab experimentation is the cornerstone to an emerging paradigm for studying functional molecular evolution (Dean and Thornton (2007)).

A multi-algorithm analysis pipeline combines the methods of phylogenetic inference and ancestral reconstruction with non-computational molecular techniques (Fig. 1.) (Thornton (2004)). This pipeline begins with a family of molecular sequences whose evolution we wish to investigate. Sequences are typically chosen whose functions vary across a family, and we wish to know how those functions shifted over evolutionary time. Here I briefly describe the pipeline stages. We first align homologous sites in the sequences, infer the phylogeny that give rise to the sequences, and then reconstruct sequences for ancestral species. An ancestral gene sequence that has been computationally reconstructed can be physically "resurrected" by synthesizing its coding sequence onto a plasmid, transfecting that plasmid into a living cell, and then allowing the cell's native genetic machinery to transcribe and translate the reconstructed gene into real protein product. Ancestral proteins can be functionally characterized using many different assays; the appropriate assay depends on the type of protein under scrutiny. Overall, this analysis pipeline allows us to observe ancient protein functions before and after significant milestones in evolutionary history.

Error and Uncertainty

Two big problems surrounding this analysis pipeline are computational error and statistical uncertainty. Error can be introduced at every stage in the pipeline; the methods of sequence alignment, phylogenetic inference, and ancestral reconstruction use heuristic algorithms that are known to produce errored results

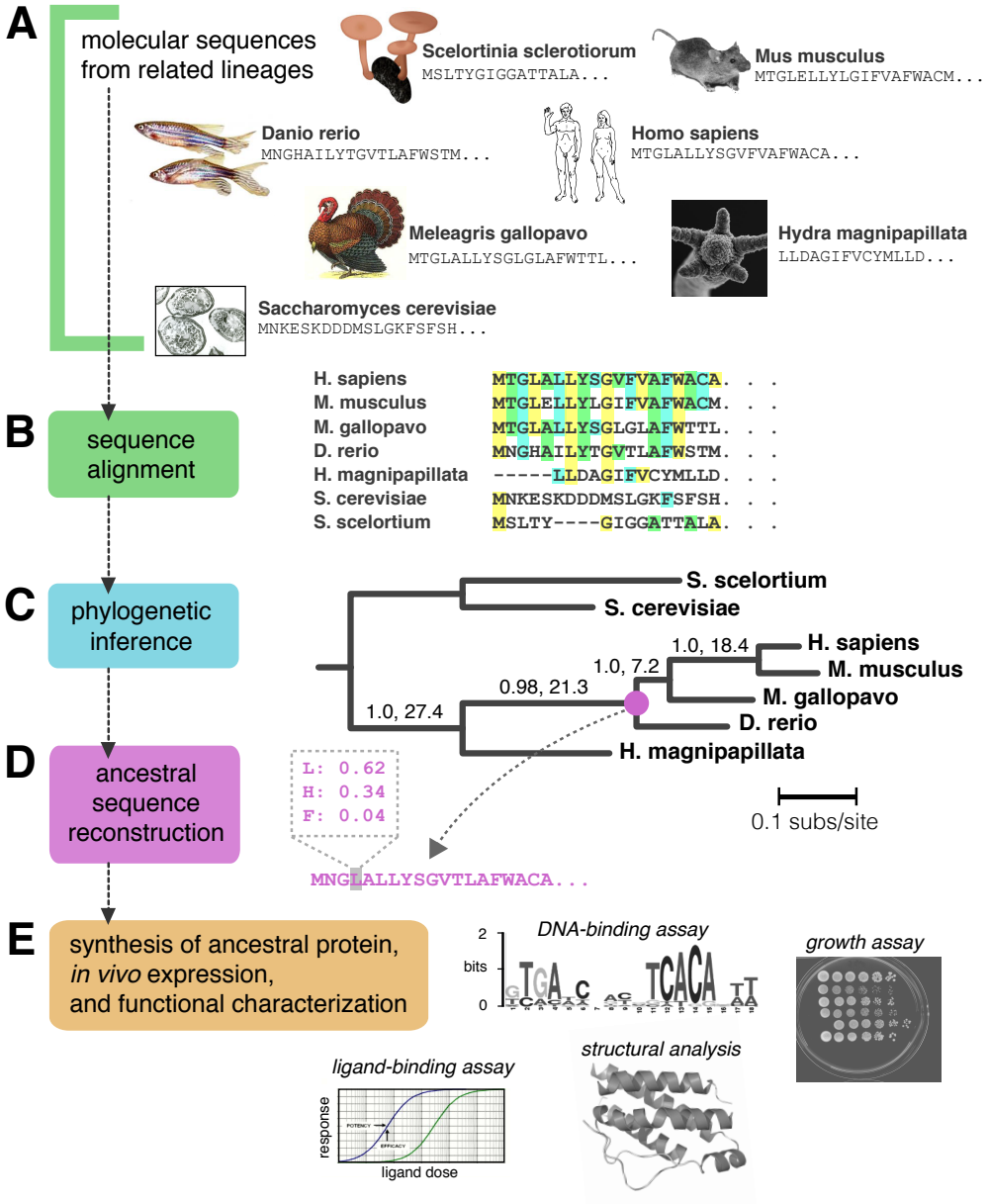


FIGURE 1. A multi-step analysis pipeline for evolutionary studies. This pipeline is used to infer the evolution and function of ancestral proteins. Shown here is the analysis of seven eukaryotic amino acid sequences, arbitrarily chosen for this example. See text in chapter I for description of the pipeline steps.

in some conditions. Identifying and eliminating sources of error is critical because inaccurate inferences made at early stages in the pipeline will lead to inaccurate inferences at downstream stages. Related to computational error, statistical uncertainty measures the degree to which we think a particular inference is accurate. The presence of significant uncertainty implies that alternate solutions should be considered in addition to the best solution. Statistical uncertainty, like error, can emerge at every stage in the pipeline; uncertainty comes in the form of mismatch costs (for alignments), likelihood scores (for trees), and posterior probabilities (for ancestral sequences). These types of uncertainty can be explicitly propagated down the pipeline by performing each downstream stage on the distribution of possible inputs from the upstream stage. Uncertainty propagation, however, incurs non-trivial computational costs and becomes intractable when taken to its philosophical extreme. It is therefore important to know when uncertainty matters, and when uncertainty can be ignored.

This dissertation addresses the role of error and uncertainty within this pipeline. I am broadly interested in two questions: *(i) How do we make the results of the pipeline more accurate? (ii) When is it appropriate to propagate uncertainty from an early pipeline stage to downstream stages?* In this introduction, I describe each stage of the analysis pipeline in more detail. In chapter II, I discuss improving the accuracy of phylogenetic inference, the role of heuristic search algorithms in introducing ML phylogenetic error, and I propose a more accurate heuristic. In chapter III, I discuss the role of phylogenetic uncertainty on the accuracy of ancestral sequence reconstruction; I show that phylogenetic uncertainty can be ignored due to a seemingly paradoxical relationship between trees and ancestral sequences. The material in chapter III was previously published with co-authors

(Hanson-Smith et al. (2010)). In chapter IV, I provide a case-study demonstrating an analysis using all stages of the pipeline. The material in chapter IV was is co-authored with collaborators in Tom Stevens lab at University of Oregon, and – at the time of this writing – is currently in-press at Nature (2011). Curious readers may also find appendix C useful, in which I show there is a complex, but significant, relationship between alignment accuracy and ancestral reconstruction accuracy.

A: Sequence Sampling

The first step of the pipeline is to collect molecular sequences – typically nucleotides or amino acids – that are evolutionary related and whose encoded functions are of experimental interest (Fig. 1.A). The amino acid sequences shown in Fig. 1.A encode an arbitrary protein fragment in seven Eukaryotic species: S. sclerotiorum (pathogenic plant fungus), D. rerio (zebrafish), M. gallopavo (turkey), S. cerevisiae (budding yeast), M. musculus (house mouse), H. sapiens (humans), and H. magnipapillata (fresh water polyp). Collecting molecular sequences is labor intensive, and most evolutionary studies use sequences that have been previously uploaded to sequence repositories. The database GenBank is the dominant worldwide repository, storing millions of protein sequences from thousands of species across the tree of life (Burks et al. (1992); Benson et al. (1999)).

B: Sequence Alignment

The second step of the analysis pipeline is to infer the relatedness – or homology – between individual sites within the set of collected sequences (Fig. 1. B). Families of sequences drift away from each other over evolutionary time, and

it may not be clear how two or more related sequences are, in fact, related. This problem is typically solved with dynamic string-matching algorithms, of which there exist many varieties (Batzoglou (2005); Notredame (2007)). The result of sequence alignment is a matrix of size $M \times N$, where M is the length of the longest sequence in the collection, and N is the number of sequences. Each cell in this matrix contains a single evolutionary character, and all the characters in each column are assumed to be homologous. Over the course of evolution, insertion and deletion events alter the length of a molecular sequence, such that some members of a sequence family can have extra characters (in the event of insertions) or have missing characters (in the event of deletions). Alignment algorithms place "gap" characters to indicate the location of insertions and deletions.

C: Phylogenetic Inference

Given an alignment of sequences, the next step in the pipeline is to infer the phylogeny that gave rise to the alignment (Fig. 1.C). A primitive approach is to cluster sequences according to their pairwise distances, measured as percentage sequence dissimilarity (Cavalli-Sforza and Edwards (1967); Sokal and Sneath (1963); Saitou and Nei (1987)). However, distance-based methods have limited utility because they compress the sequence alignment into a matrix of pairwise distances between sequences, and thus discard potentially useful information about evolutionary constraints at individual sequence sites. Rather, the dominant paradigm for phylogenetic inference is to begin with a distance-based tree and then optimize this tree using a Markov model in a likelihood framework. Unlike simpler distance-based approaches, Markov models explicitly consider the substitutional

process by which sequences evolved. The statistical foundations of molecular Markov models are described in more detail in Appendix A.

Phylogenetic Markov models include parameters whose values are typically unknown. These parameters include the phylogenetic topology, branch lengths on that topology, the relative substitution rates between sequence states, and—depending on the particular model—other parameters to account for various evolutionary processes. A likelihood function is used to calculate the likelihood of a particular set of parameter values (Felsenstein (1981)); a search function is then used to find the set of values with the maximum likelihood. Given a sequence alignment D, the likelihood $L(t, \theta|D)$ of a topology t with model parameters θ, is defined as the probability $P(D|t, \theta)$ of observing D given t and θ. Likelihoods are calculated using an algorithm that recursively traverses the phylogenetic tree (described in Appendix B). The goal of maximum likelihood (ML) phylogenetics is to find values for t and θ that maximize the function $L(t, \theta|D)$. The search strategies used to find ML phylogenies are the subject of Chapter II.

D: Ancestral Sequence Reconstruction

Once a phylogeny is found, the next step in the pipeline is to infer the sequences for ancestral nodes (Fig. 1.D). Although any ancestor on a phylogeny can be reconstructed, usually only a few ancestors are experimentally relevant, depending on the specific hypothesis under scrutiny. Most ancestral queries target historical shifts in protein function; these types of studies require at least two ancestors: one immediately before the shift and another immediately after the shift took place. Other hypotheses may require more than two ancestors. For example, more than a dozen reconstructed opsin protein ancestors were

used to reveal that vertebrates historically evolved through a variety of aquatic environments (Yokoyama et al. (2008)). In another example, eleven reconstructed elongation factor protein ancestors were used to infer a historical shift in Earth's paleotemperature (Gaucher et al. (2003, 2007)).

Ancestral sequences are reconstructed using the same types of Markov models that are used to infer ML phylogenies. However, rather than searching for the ML tree, ancestral reconstruction uses a fixed ML tree and searches for the ML ancestral states. The computational mechanics of ML ancestral reconstruction are discussed in Chapter III.

E: Functional Characterization

After ancestral sequences have been computationally reconstructed, the function of those ancestors can be experimentally observed using molecular techniques (Fig. 1.E) (Thornton (2004); Liberles (2007)). Specifically, ancestral sequences can be physically synthesized, subcloned onto plasmids, and transfected into living cells. The cells' native genetic machinery will then transcribe and translate the plasmid sequence into actual proteins. In situations where the cell naturally contains a contemporary descendant of the ancestral sequence, the cell's native copy can be disabled such that the only functional copy is the ancestral gene on the plasmid. Once an ancestral protein is expressed, its function can be studied using a variety of assays. The appropriate assay depends on the protein family. Transcription factor proteins, for example, can be assayed to determine their binding preference for different DNA motifs (Stormo and Zhao (2010)). Nuclear proteins can be assayed to determine their binding-response to variable ligand doses (Bridgham et al. (2006)). Some protein families can be studies at a coarse

level, where simple cell growth can be taken as a proxy for function (*Finnigan and Hanson-Smith 2011*). Finally, some proteins are amenable to structural analysis – using techniques like X-ray crystallography – and the biophysical determinants of their specific functions can be investigated (Ortlund et al. (2007); Harms and Thornton (2010)). Taken together, this interdisciplinary analysis pipeline opens a window into the evolutionary past and allows for the direct observation of hypothesized protein functions that have not existed in millions – or billions – of years.

CHAPTER II

REDUCING ERROR IN PHYLOGENETIC INFERENCE

Phylogenetic tree structures are the de facto representation for evolutionary relationships among related molecular sequences. Knowing the correct tree is the necessary first step for many useful downstream analyses, including ancestral sequence reconstruction (Liberles (2007)), phylogeography (Avise et al. (1987); Avise (1998)), and estimation of species divergence times (Taylor and Berbee (2006)). The correct phylogeny, however, is often unknowable because the ancestral species necessary to determine historical branching patterns are typically unavailable. Rather, phylogenies are usually inferred computationally from contemporary sequences. The dominant paradigm for phylogenetic inference is to use a parametric Markov model that describes the relative substitution rates between different molecular states — typically, nucleotides, amino acids, or codons. Given a molecular sequence alignment, the likelihood of a particular tree and model equals the probability of observing the alignment, given the tree topology, branch lengths on that topology, and specific values for substitution rates between states (Felsenstein (1981); Bryant et al. (2005)). The true values for the tree, branches, and parameters are typically unknown, and search and optimization algorithms are necessary to find the combination with the maximum likelihood (ML) value. To the extent that the likelihood function correlates with the accuracy of evolutionary history, the ML phylogeny is the most probable explanation for the evolution of a given sequence alignment.

The ML phylogeny, however, is not always easy to find because the space of possible trees and parameter values is so immense that brute-force search strategies

are computationally intractable for all but trivial-sized problems. Instead, search heuristics are employed to constrain the ML exploration to high-likelihood regions of parameter space (Felsenstein (2004)). The search for an ML tree is typically decomposed into two nested problems: optimizing tree topologies, and optimizing continuous parameters (Figure 2.). The primary problem is to search the space of tree topologies and find the topology with the highest likelihood. The search for the ML topology typically begins with an initial tree constructed using a neighbor-joining algorithm (Saitou and Nei (1987); Gascuel (1997)). From this initial tree, the space of possible tree topologies can be traversed by swapping tree branches in order to transform one topology into a different topology. The secondary problem is to optimize the branch lengths and other model parameters on each explored tree. Virtually everyone solves the secondary problem using an approach I refer to as Unimax, in which free parameters are sequentially optimized individually. Unimax is typically implemented using the van Wijngaarden-Deker-Brent method, which combines three unique hill-climbing algorithms – inverse quadratic interpolation, root bracketing, and bisection – to find the maximum of a function with one free parameter (Brent (1972)). Unimax is the default option in popular phylogenetic software packages PhyML (Guindon et al. (2010)), RaxML (Stamatakis (2006); Ott et al. (2007)), Garli (Zwickl (2006)), PAML (Yang (2007)), and PAUP (Swofford (2003)).

Unimax assumes free parameters are separable. In other words, Unimax assumes that the ML solution can be found by individually optimizing each parameter while holding all other parameters constant. For branch lengths, at least, this assumption seems to be incorrect because the likelihood of a phylogeny is computed via a postorder traversal of the tree – using Felsenstein's so-called

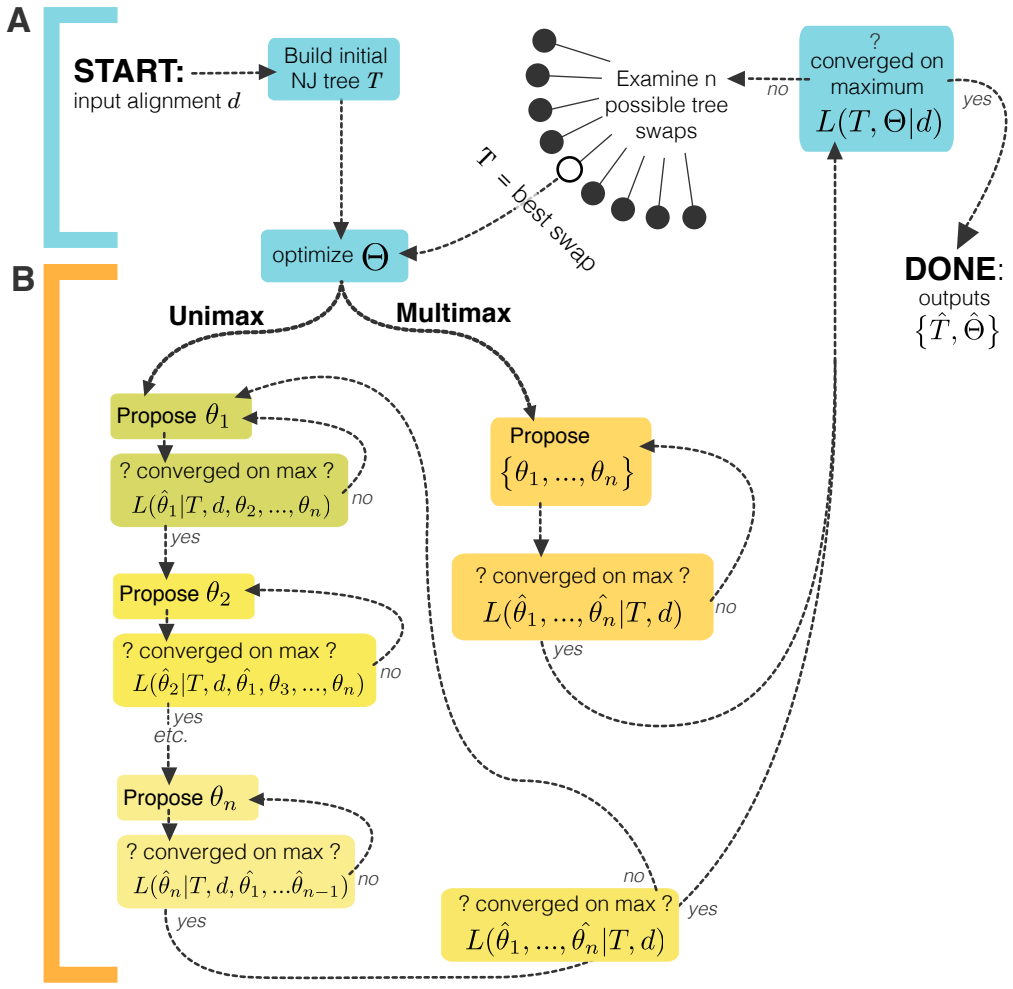


FIGURE 2. Flowchart of Unimax and Multimax optimization. Unimax and Multimax use different approaches to optimize model parameters. Finding the ML phylogeny involves two nested problems. (A) The high-level problem is find the ML topology T. Starting with a multiple sequence alignment d, propose a topology T, and optimize all free model parameters and branch lengths Θ on T. This process stops when we converge on an ML T. Otherwise, we examine possible topological rearrangements to T (tree swaps), and repeat the overall process. (B) The low-level problem is optimize all continuous parameters on a fixed T. Unimax sequentially optimizes each free parameter $\theta_1, \theta_2, ..., \theta_n \in \Theta$, using fixed values for all other free parameters. Multimax simultaneously optimizes all parameters $\theta_i \in \Theta$, at each iteration proposing a new set of parameter values.

pruning algorithm – in which conditional probabilities of ancestral states at internal nodes are propagated up the tree (Felsenstein (1981)); adjusting one branch length affects the conditional probabilities at nearby internal nodes, which ultimately affects the likelihood of nearby branch lengths. The recursive relationship between branches means that branch lengths may be correlated parameters. Unimax ignores this recursive relationship, and instead optimizes each parameter in isolation.

Stated formally, Unimax assumes that the ML solution for a parameter θ_i can be found using fixed values for the alignment d, the topology t, and all other free parameters θ_j in the set of parameters Θ (Eq. 2.1).

$$P(\hat{\theta_i}|d, t, \theta_j \in \Theta, i \neq j) \tag{2.1}$$

In contrast, a non-separable optimization method simultaneously seeks the ML solutions for all free parameters $\hat{\theta_1}, ..., \hat{\theta_n}$, using a fixed alignment d and topology t (Eq. 2.2).

$$P(\hat{\theta_1}, ..., \hat{\theta_n} \in \Theta|d, t) \tag{2.2}$$

Although previous scholarship has recognized faulty logic in Unimax's assumption of separability, the degree to which this assumption impairs the accuracy of ML phylogenetic inference has not been systematically investigated (Yang (2000); Bryant et al. (2005)).

In order to assess the effect of separability on phylogenetic accuracy, I implemented an alternative ML optimization algorithm that optimizes all parameters simultaneously rather than separably. This approach, referred to here as Multimax, is formally based on the conjugate-gradient approach of the Broyden-

Fletcher-Goldfarb-Shanno (BFGS) algorithm (Press et al. (1992); Nocedal and Wright (1999)). BFGS operates by estimating the first- and second-derivates of the likelihood function locally around the current parameter values. BFGS then uses these derivates to estimate an ML solution. BFGS jumps to this estimated optimum, and recomputes the local derivatives. If the functional gradient at the new point is zero or within an acceptable margin, BFGS has converged upon an ML solution. Otherwise, BFGS uses the new gradient to estimate an updated ML solution. BFGS repeats this process until it converges on an optimum or reaches a user-specified maximum number of iterations.

I compared the performance of Multimax (MM) to Unimax (UM) under a range of conditions, both empirical and simulated. Across these conditions, I observed that UM's assumption of separability significantly impaired the accuracy of phylogenetic inference. UM was less accurate than MM because UM leads to poor ML branch lengths, which ultimately drives the tree search algorithm into suboptimal regions of tree space.

Materials and Methods

Unimax

I used PhyML's implementation of Unimax using Brent's method, as described in chapter 9.3 of Numerical Recipes in C (Press et al. (1992)).

Multimax

Multimax using BFGS works as follows. Given a sequence alignment d, a fixed tree topology T, and a set of starting values for all free parameters Θ, BFGS estimates the first- and second-derivate gradients of the likelihood function

$L(\Theta|T,d)$. My software implementation estimates these gradients as a Hessian matrix, constructed with local secant approximation (Nocedal and Wright (1999)). The likelihood gradients are used to find a multidimensional uphill direction p in parameter space (Eq. 2.3).

$$Hp = -\nabla L(\Theta|T,d) \tag{2.3}$$

where H is the Hessian matrix of second-order partial derivates between all parameters $\theta \in \Theta$, p is the direction of the functional optima relative to our current values of Θ, and ∇L is the first-derivate of the likelihood function. After solving for p, BFGS performs a line search in the direction of p in order to propose a new optimum. BFGS then jumps to this proposed point. BFGS repeats these four steps – calculating gradients, solving equation 2.3, proposing an optimum, and then jumping – until it arrives at a point whose gradient is zero or 200 iterations have been performed.

Tree Swapping

I used a combination of nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) to swap branches and propose new topologies (Swofford and Olsen (1990)). I used the default implementations of NNI and SPR within PhyML version 3.0. PhyML's swap algorithm calculates a swap score for every possible topology rearrangement; this score estimates the potential likelihood increase of accepting the swapped tree relative to the pre-swapped tree. Swap scores were calculated as the estimated likelihood of the swapped tree (L_{t+1}) minus the likelihood of the current tree (L_t). The values for L_{t+1} were estimated by optimizing only the four branches directly affected by the swap.

Simulated Sequence Alignments

In order to compare the accuracy of UM and MM on simplified conditions, I simulated amino acid and nucleotide sequences evolving on randomly generated trees whose sizes varied from 8 to 1024 terminal branches. For each tree size, I generated twenty random trees and generated a random amino acid sequence at the root of the tree. I simulated the ancestor evolving across the branches, without insertion/deletion events, to produce an alignment of descendant sequences. All branches were individually drawn from the uniform distribution [0.00,0.05] for pinnate trees and [0.00,0.50] for balanced trees. Nucleotide sequences were simulated using the JC69 model to create alignments with 1000 sites. Amino acid sequences were simulated using the JTT model to create alignments with 400 sites. Simulations were performed using INDELible with insertion/deletion events disabled (Fletcher and Yang (2009)).

Empirical Sequence Alignments

In order to determine if UM biased our evolutionary interpretation of real sequence data, I used MM and UM to infer ML trees for sequence alignments from six gene families: (i) steroid-hormone receptor ligand binding domains from across Metazoa (Bridgham et al. (2008)), (ii) Mcm1 transcription factors from twelve species of Fungi Ascomycete Saccharomycotina (Baker et al. (2011)), (iii) vacuolar ATP-ase subunits c, c', and c" sampled broadly from Opisthokonts (cite XX), (iv) thioredoxins from species across the tree of life (Perez-Jimenez et al. (2011)), (v) ribosomal 16S nucleotide sequences sampled from across proteobacteria (cite XX), and (vi) TMO-4c4 gene nucleotide sequences, sampled broadly from the mitochondria of ray-finned fish (Scorpaeniformes) (Smith and Wheeler (2004)).

I compared the MM and UM ML trees using four criterion: their congruity with our a priori expectations, their maximized likelihood scores, their complementary topological differences, and their unique paths taken through tree space during tree search.

Phylogenetic Inference

I inferred ML phylogenies for empirical alignments using our own in-house modifications to PhyML version 3.0. Tree search began with the neighbor-joined tree, as implemented in PhyML. The search was then driven by either UM or MM until convergence on an optimum. The best-fitting Markov model was found by repeating the search with different substitution matrices and levels of heterogeneity, and then using the Akaike Information Criterion to find the model with the highest likelihood without overparameterization (Akaike (1973)). Using the best-fitting model, ML phylogenetic inference was performed with full tracing enabled, in which PhyML records path taken through tree space.

Reciprocal Restart Analysis

In order to determine if UM or MM could further optimize the other method's ML tree, I restarted UM from the MM ML tree and restarted MM from the UM ML tree. In order to determine if UM or MM could optimize the other method's ML branch lengths on the initial topology (before any swaps had been performed), I disabled topology search and restarted MM from the UM ML branch lengths on the initial topology and restarted UM from the MM ML branch lengths.

Ancestral Sequence Reconstruction

I reconstructed ancestral sequences using maximum likelihood as implemented in PAML version 4.2 and an in-house GUI – named Lazarus – that controls PAML (Yang (2007); Hanson-Smith et al. (2010)).

Tree Error

For every simulated alignment, I measured relative tree error as the symmetric difference between the neighbor-joined topology and the true tree, divided by the number of branches in the tree. I computed symmetric differences using the function *dendropy.Tree.symmetric_difference* as implemented in the DendroPy library (Sukumaran and Holder, 2010). Relative tree error, informally, measures the proportion of spurious clades in a tree. In order to determine statistical significance between Unimax and Multimax tree error, I used a paired two-tailed T-test to measure if the mean tree errors were significantly different. T-values and derived P-values were computed using the function *cogent.stats.math.t_paired* in the PyCogent library (Knight et al., 2007).

I measured the error in overall tree length by dividing the sum of branch lengths on each ML tree by corresponding sum on its true tree. This product is reported in this paper as length error. In order to determine statistical significance between Unimax and Multimax tree length error, I used a paired two-tailed T-test to measure if the mean tree errors were significantly different. T-values and derived P-values were computed using the function *cogent.stats.math.t_paired* within the PyCogent library (Knight et al., 2007).

Results

Separable ML optimization impairs evolutionary interpretation

Unimax (UM) and Multimax (MM) drove the tree search to find different ML trees, so the assumption of separability does matter. For sequences simulated under controlled conditions, the difference between UM and MM ML trees was greater for large trees and for pinnate-shaped trees (Fig. 3.). For sequences of gene families evolved under real conditions, UM and MM drove the tree search algorithm to find disagreeing topologies for five out of six families (Figs. 4. - 8.). In most cases, there was strong support for the disagreeing branches; in a typical analysis, the placement of these discordant branches would not be interpreted as uncertain, nor would their placement warrant further investigation regarding sequence choice.

MM was superior in finding trees with high likelihood scores and with less error. This means the assumption of separability not only matters, but actually impairs phylogenetic inference. For empirical sequences, MM-driven search led to trees with higher likelihoods, indicating a more effective search of space (Fig. 10.). For simulated sequences, in which the true is known, I found MM trees had fewer erroneously placed clades than UM trees (Fig. 9.).

Separable ML optimization leads to suboptimal trees

I next sought why UM led to poor ML solutions. Was it that UM foiled the search by getting stuck on ridges, saddles, or other non-peak features in parameter space? Or, did UM irreversibly drive the search into poor regions of tree space whose highest summits were non-global optima? I disqualified the first theory by performing reciprocal restart analysis on ML trees (Fig. 11.). For the five empirical

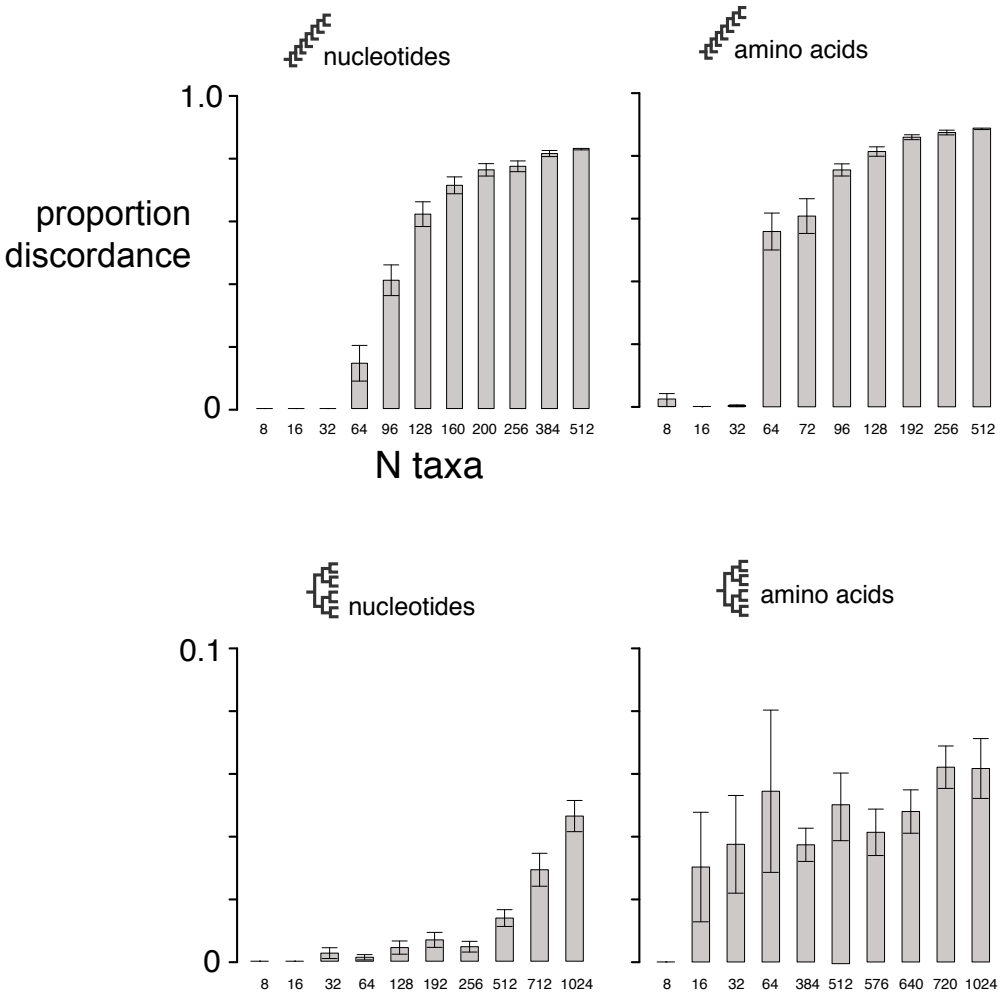


FIGURE 3. Discordance between Unimax and Multimax trees. Unimax and Multimax found discordant ML topologies in many cases. N taxa is the number of sequences in the simulated alignment. Proportion discordance, the mean proportion of clades that differed between the UM and MM ML trees.

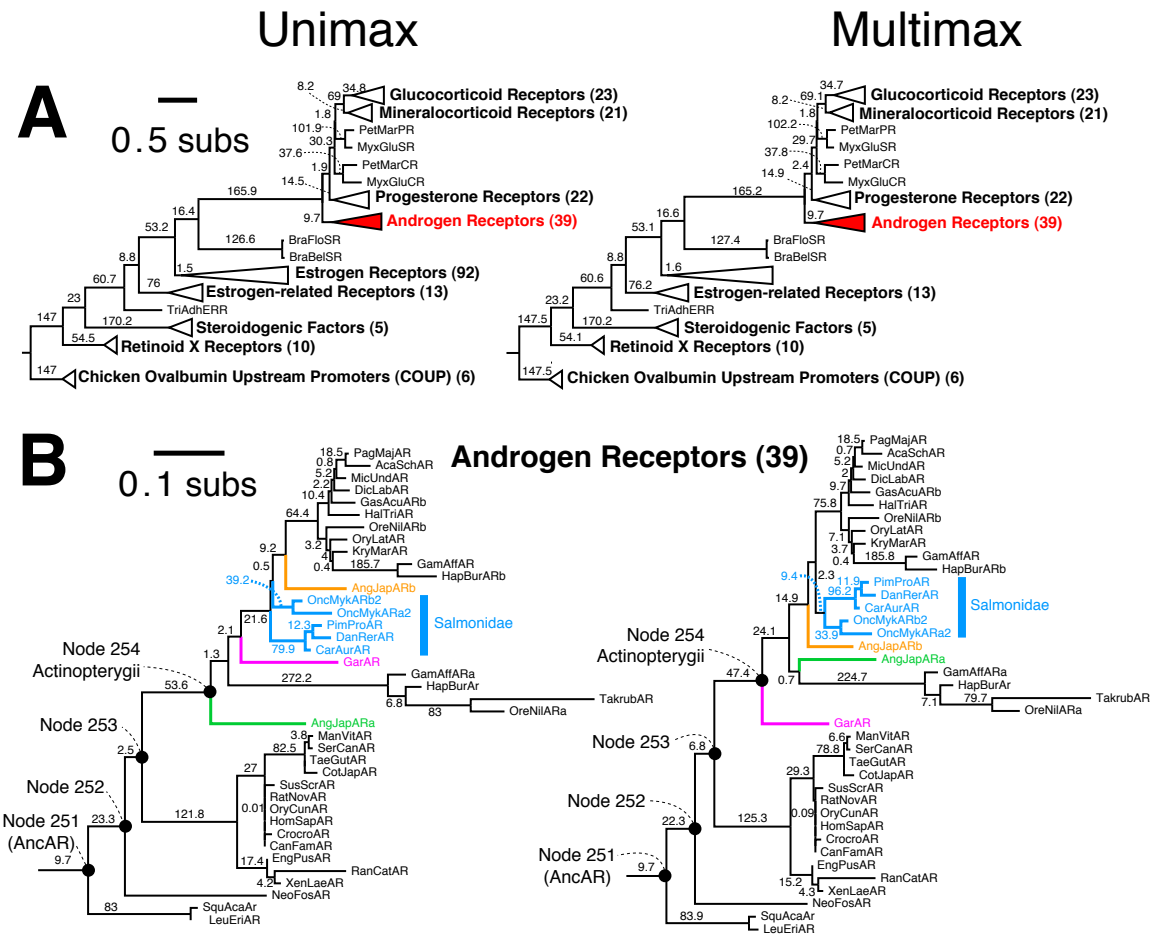


FIGURE 4. Phylogenies of steroid-hormone receptor family. Shown also are related outgroup genes. (A) The ML tree inferred using Unimax is shown on the left, and the ML tree inferred using Multimax is on the right. The clade in red is expanded in part (B). In the clades of androgen receptors, colors indicate branches that were placed differently on the two trees. All support values on internal branches are approximate likelihood ratio test values. Relevant ancestral nodes (see main text) are labeled as Node 251, 252, 253, and 254.

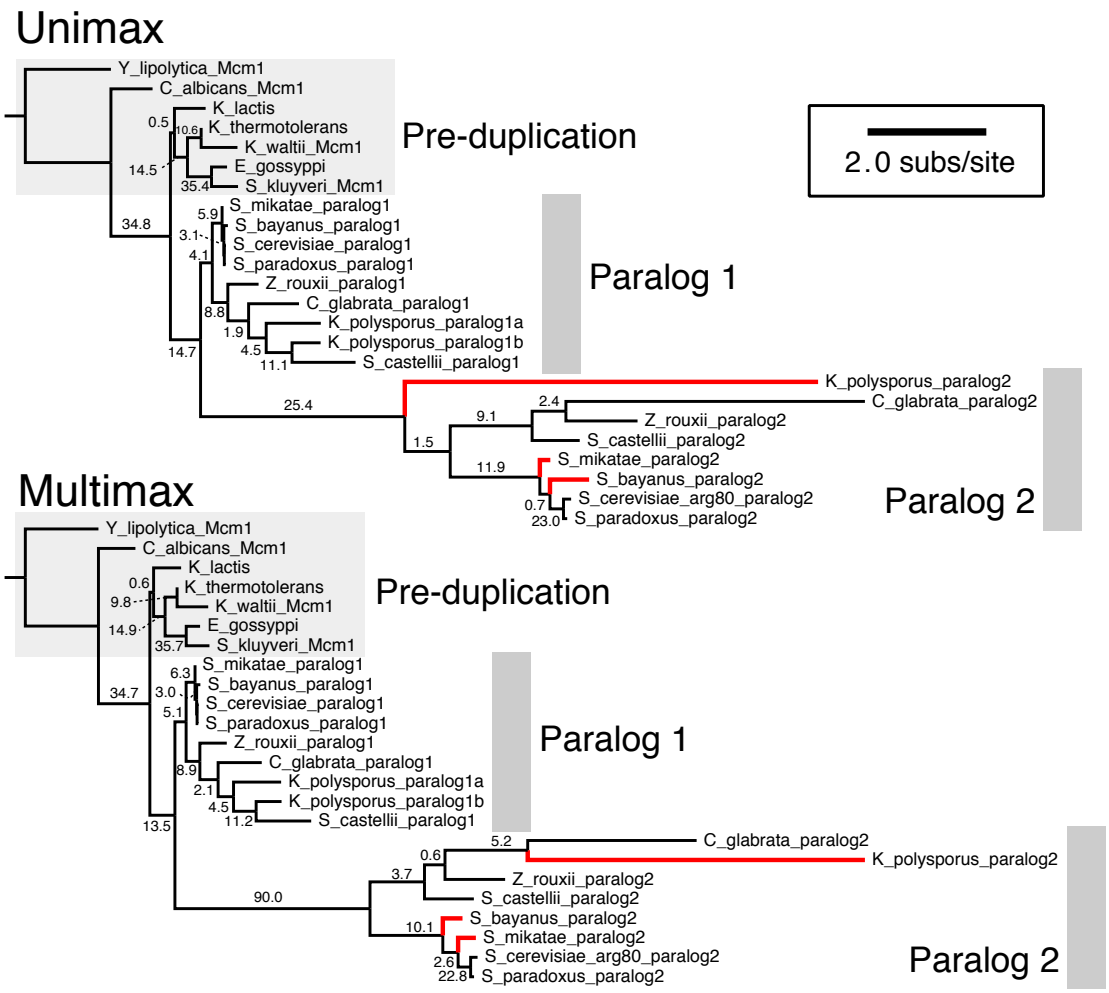


FIGURE 5. Phylogenies of Mcm1 protein family. Terminal amino acid sequences come from twelve species of Fungi Ascomycete Saccharomycotina. A gene duplication occurred sometime after the clade containing S. kluyveri, S. gossyppi, K. thermotolerans, K. walti, and K. lactis branched from the other species. The ML tree optimized using Unimax is shown on top, and the Multimax ML tree is shown on bottom. Red branches are discordant between the two trees. Support values on internal branches are approximate likelihood ratio test values.

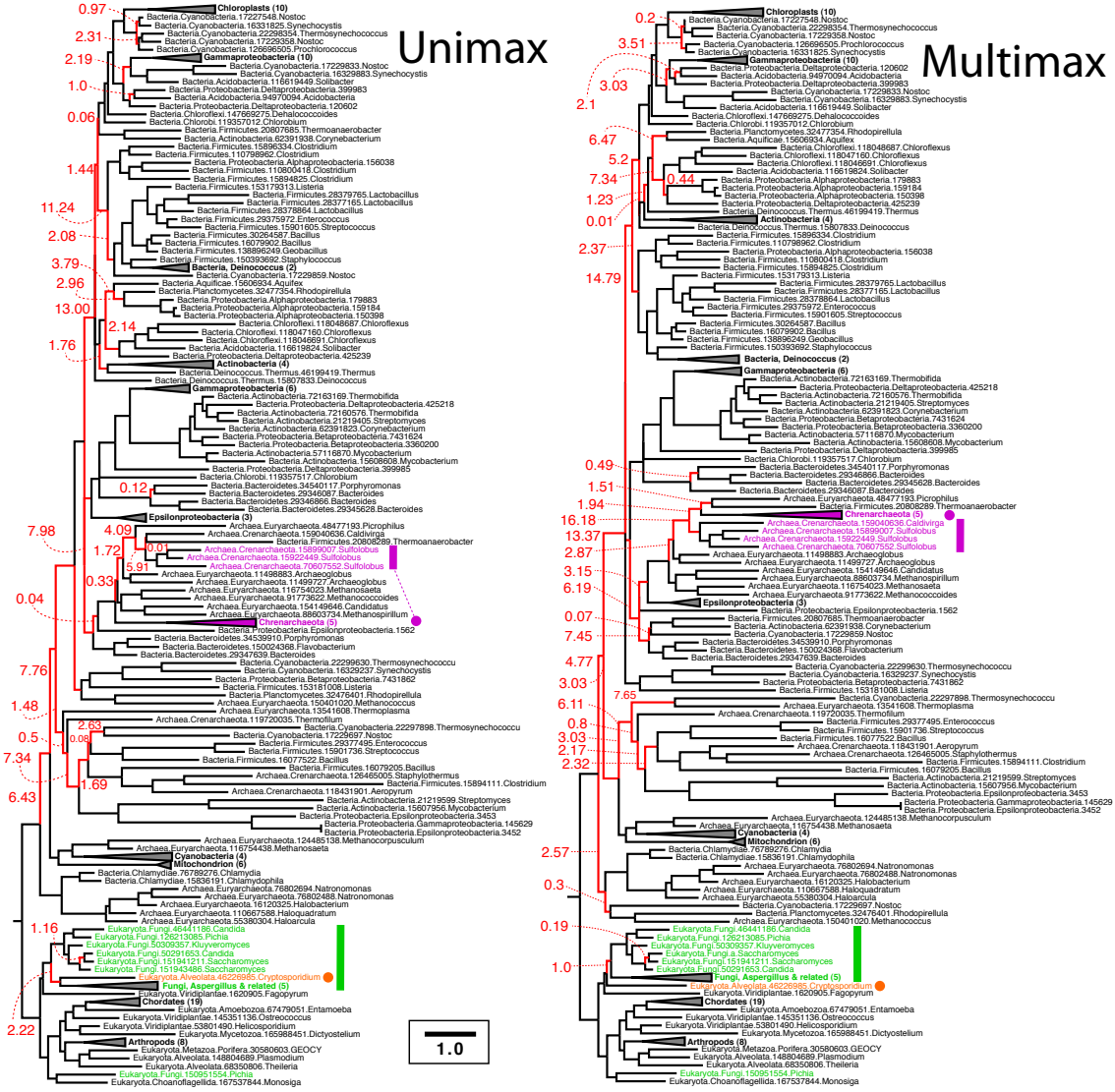


FIGURE 6. Phylogenies of Thioredoxin protein family. Terminal sequences are sampled from across the tree of life. The ML tree on the left was optimized using Unimax, and the tree on the right was optimized using Multimax. Red branches disagree between the two trees. Related Chrenarchaeota species are highlighted in purple, Fungi are highlighted in Green, and Cryptosporidium is highlighted in Orange. Support values on branches are approximate likelihood ratio test values.

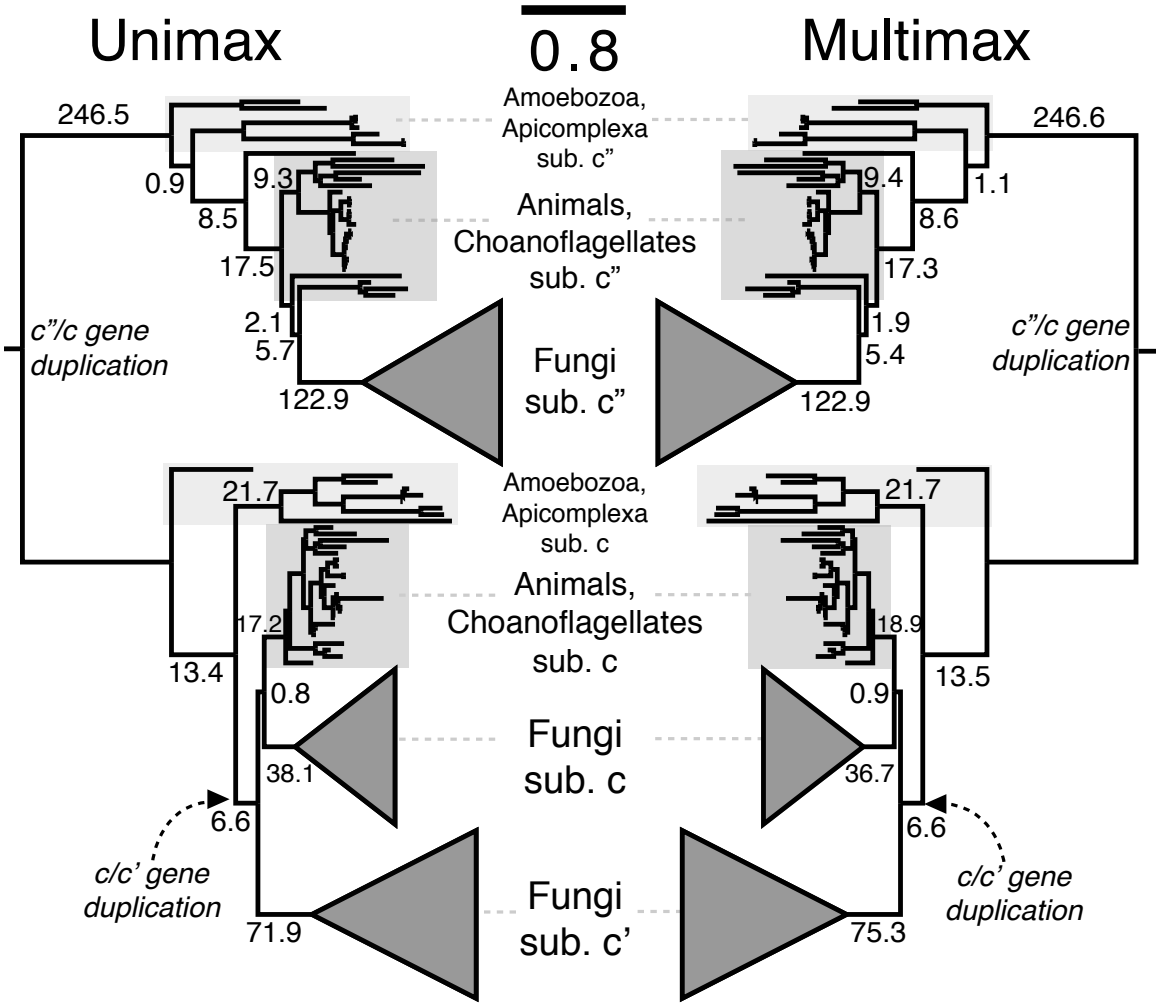


FIGURE 7. Phylogenies of V-ATPase subunits c, c', and c". Sequences are samples from across Opisthokonts. The Unimax ML tree is on the left, the Multimax ML tree is on the right. The two trees have identical topologies, but different branch lengths and approximate likelihood ratio test values.

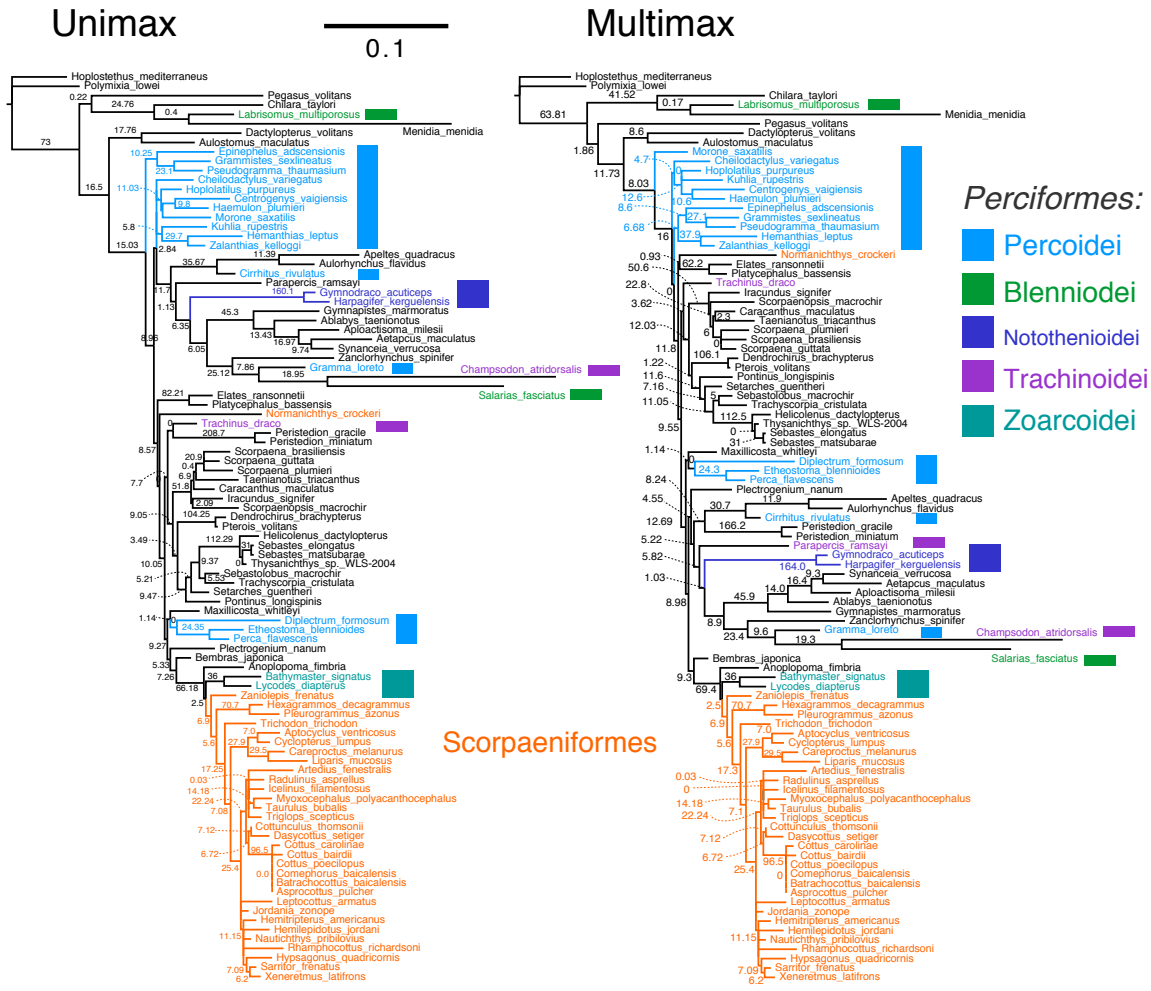


FIGURE 8. Phylogenies of the TMO-4c4 gene family in Actinopterygii (ray-finned fish). The Unimax ML tree is on the left; the Multimax ML tree is on the right. Perciformes and suborders are highlighted in blue, purple, and green. Scorpaeniformes are highlighted in orange. Support values on internal branches are approximate likelihood ratio test values.

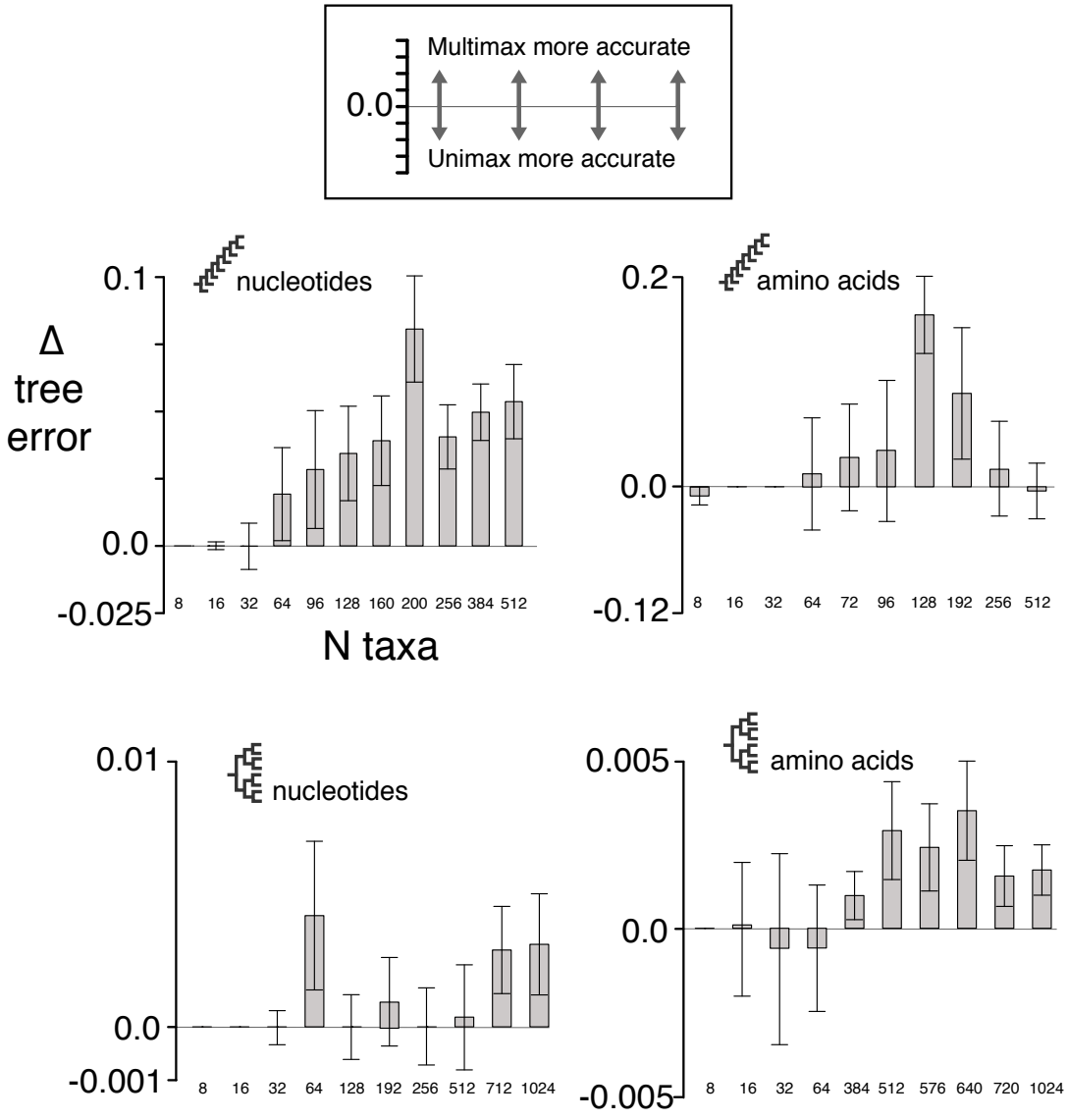


FIGURE 9. Tree error for Unimax and Multimax ML trees. N taxa, the number of sequences in the simulated alignment. Δ tree error, the mean difference of UM relative tree error and MM relative tree error (see Methods). Values above 0.0 indicate that MM was more accurate; values below 0.0 indicate that UM was more accurate. Error bars are standard error of the mean.

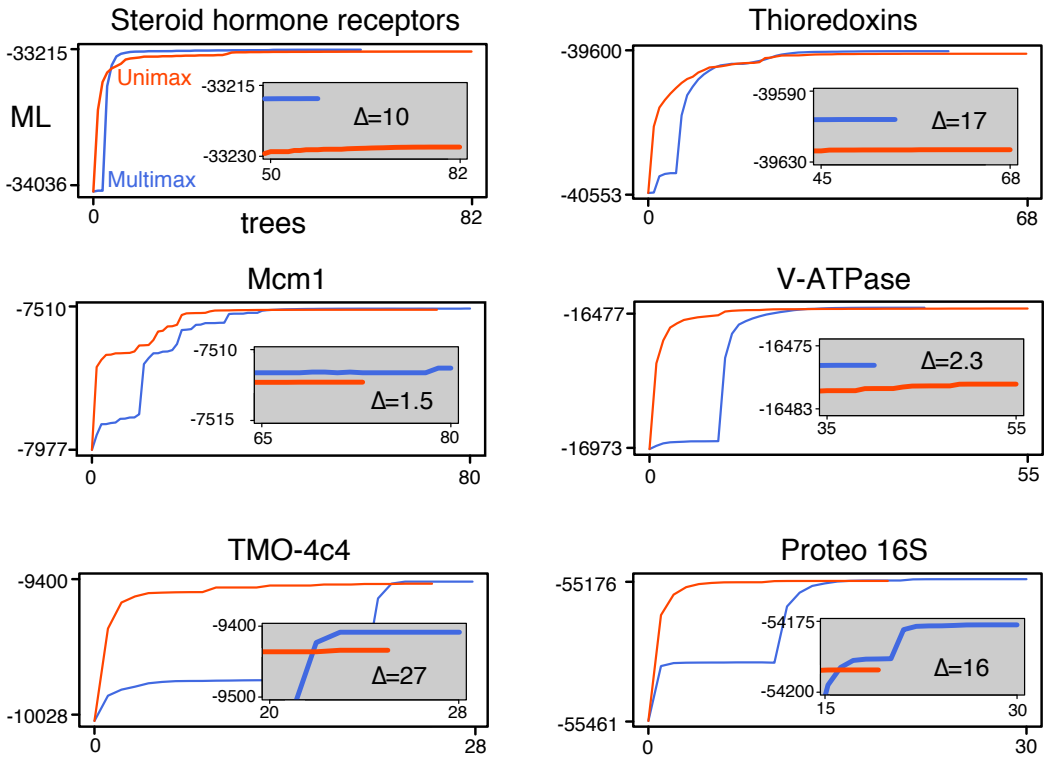


FIGURE 10. Maximum likelihood values of phylogenies for empirical alignments. **ML**, the maximum log-likelihood of the best tree during the tree search. **trees**, the unique topologies accepted by UM and MM during their search of tree space. Δ, the final difference in log(L) scores between the Unimax and Multimax solutions.

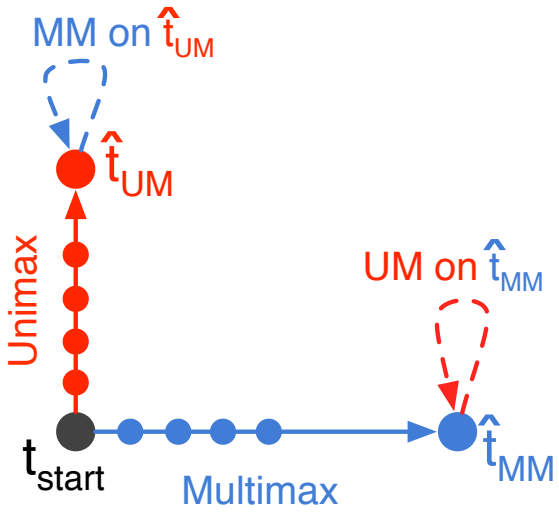
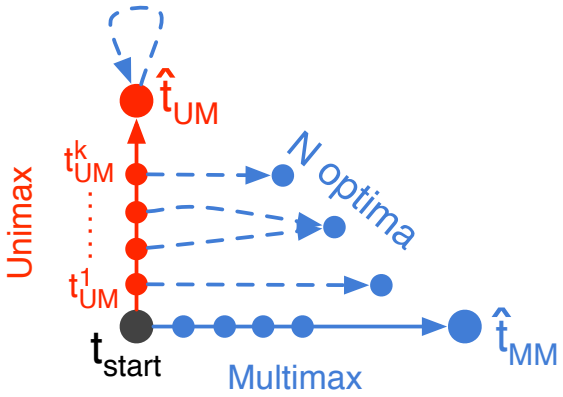


FIGURE 11. Schematic illustration of reciprocal restart analysis on ML trees. ML search begins from an initial tree (t_{start}), typically found using a neighbor-joining algorithm. UM and MM drive the search into different regions of tree space. UM's ML tree ($\hat{t}_{UM}$) is different from MM's ML tree ($\hat{t}_{MM}$). In this reciprocal restart analysis, MM-driven search was restarted from $\hat{t}_{UM}$ and UM-driven search was restarted from t_{MM}. In no examined cases, could UM or MM drive tree search to improve the other algorithm's ML tree.



alignment	k	N optima	$N > t_{UM}$
steroid-hormone receptors	31	12	4
Mcm1	20	7	4
thioredoxins	139	107	33
16S	46	33	5
TMO-4c4	30	29	7
V-ATpase subunits c, c', c"	31	0	0

FIGURE 12. Schematic illustration of path restart analysis on UM trees. MM was restarted from all trees along UM's path through tree space (t^1_{UM} through t^k_{UM}). The table shows path restart data from six sequence families evolved under real conditions (see Methods). k, the number of trees unique to the UM path, excluding $\hat{t}_{UM}$. N optima, the number of new optima found by MM. The fourth column, $N > \hat{t}_{UM}$, is the number of new optima found by MM that also had log-likelihood scores better than the UM ML tree ($\hat{t}_{UM}$).

sequence alignments in which UM and MM ML trees disagreed, I restarted UM from the MM-driven ML tree and restarted MM from the UM-driven ML tree. For all five alignments, UM and MM were unable to drive the tree search algorithm to find a better topology, indicating that UM-driven and MM-driven search indeed found topological optima (but the MM optimum was better). Further restart analysis revealed the second theory to be true: Unimax led to suboptimal trees. Specifically, I restarted MM from every tree along the UM search path; MM-driven search then found several additional optima with higher ln(L)s than the original UM optimum (Fig. 12.). This reveals that UM-driven search repeatedly chose poor topologies.

Separable optimization impairs branch length accuracy

In order to determine why UM-driven search chose suboptimal trees, I examined the choices UM and MM made immediately before they diverged in tree space (Fig. 13.). Tree search began from an initial tree (t_{start}) with initial branch lengths (bl_{NJ}), constructed using a neighbor-joining algorithm. UM and MM optimized the values of bl_{NJ} to arrive at different ML branch lengths $\hat{bl}_{UM}$ and $\hat{bl}_{MM}$, respectively. From the ML branch lengths on t_{start}, the tree search algorithm then evaluated a list of candidate tree swaps for the next tree (t^1). Each potential swap was scored according to its estimated likelihood improvement. PhyML calculated a swap score for each proposal by first making the swap, optimizing only the branch lengths directly affected by the swap, recording the likelihood of this partially-optimized tree, and then restoring the tree to its pre-swap condition. Because proposed trees are partially optimized – rather than fully optimized – swap scores are highly dependent on the ML branch lengths of the pre-swap tree. So if

one uses an optimization algorithm that systematically finds erred branches, then one might expect to make poor topology swaps.

In order to determine if UM branch lengths were erred, I measured the accuracy of the sum of all branch lengths on the ML trees. I found UM ML branches to be less accurate than MM ML branches (Fig. 14.). Specifically, the sum of all branch lengths on the UM ML trees were too long, especially on pinnate-shaped trees. On the six empirical alignments, branch length error cannot be decisively tested because true branch lengths are unknown. Instead, I recorded the sum of all branch lengths on the UM and MM trees over the duration of tree search. Consistent with my observations from simulated alignments, the UM ML trees were systematically longer than the MM ML trees (Fig. 15.).

I next determined if UM branch lengths were less accurate simply because UM failed to climb peaks in the space of continuous parameters. I reciprocally restarted UM from $\hat{bl}_{MM}$ and MM from $\hat{bl}_{UM}$, disabling the tree search algorithm (Fig. 13.). For all six empirical alignments, UM and MM were unable to find better reciprocal ML branch lengths on t_{start}, indicating that $\hat{bl}_{UM}$ and $\hat{bl}_{MM}$ were indeed optima in the space of continuous parameters.

Taken together, these observations suggest that UM chose suboptimal trees because UM's ML branches were erred.

Separable optimization finds ML values that are order dependent

Unimax's ML branch lengths depend on the sequential order in which parameters are optimized. To illustrate this point, I implemented alternative versions of UM with different sequential orderings for branch lengths. I measured how these alternative optimization orders affected ML values for a branch labeled

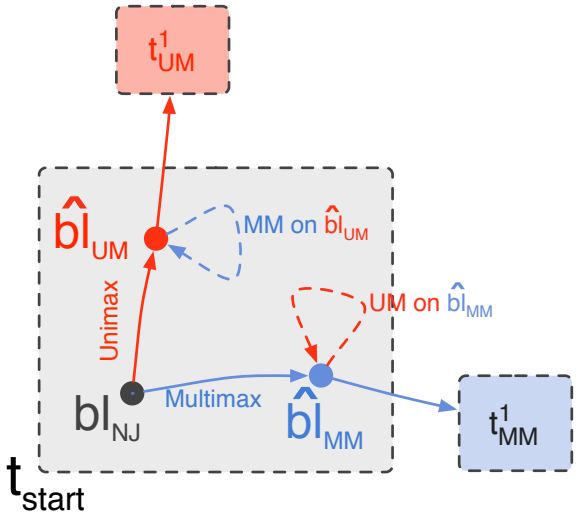


FIGURE 13. Schematic illustration of reciprocal restart analysis on initial trees. Tree search began on an initial tree t_{start}, with initial branch lengths bl_{NJ}. UM and MM found ML branch lengths on t_{start}, labeled $\hat{bl}_{UM}$ and $\hat{bl}_{MM}$, respectively. When MM was restarted from $\hat{bl}_{UM}$, MM was unable to improve the values of $\hat{bl}_{UM}$. Similarly, UM restarted from $\hat{bl}_{MM}$ was unable to improve the values of $\hat{bl}_{MM}$. UM-driven search and MM-driven search next chose different topology swaps, labeled t^1_{UM} and t^1_{MM}.

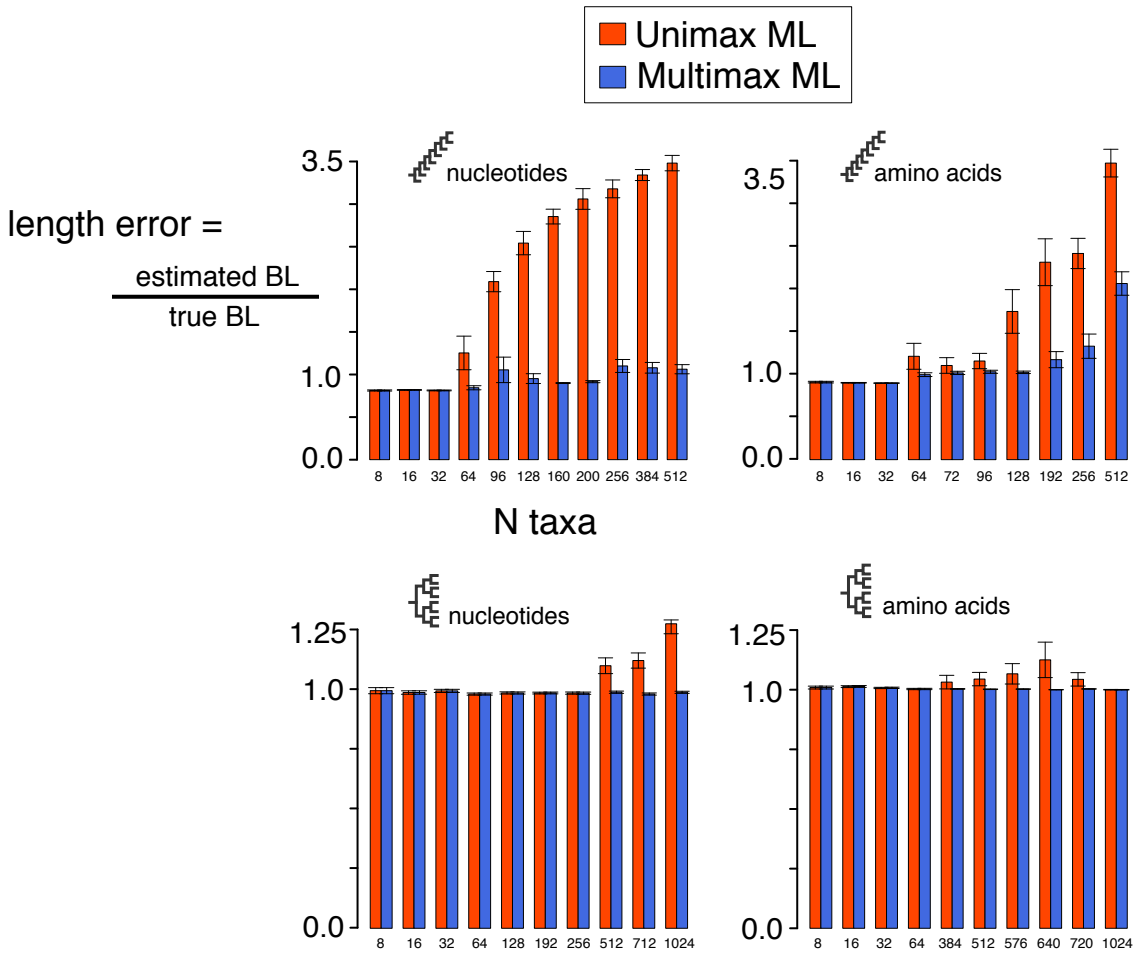


FIGURE 14. Tree length accuracy. Length error is ratio of the estimated (ML) branch lenghs to the true branch lengths. N taxa is the number of sequences in the simulated alignment. Error bars are standard error of the mean.

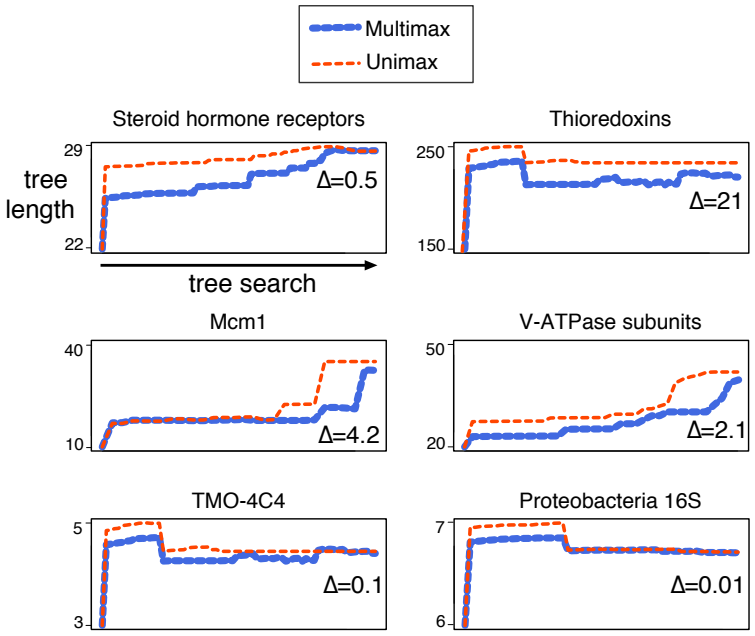


FIGURE 15. The length of ML trees over the duration of tree search. Tree search on the horizontal axis corresponds to the set of unique topologies explored by Unimax and Multimax. Tree length on the vertical axis is the sum of all branch lengths.

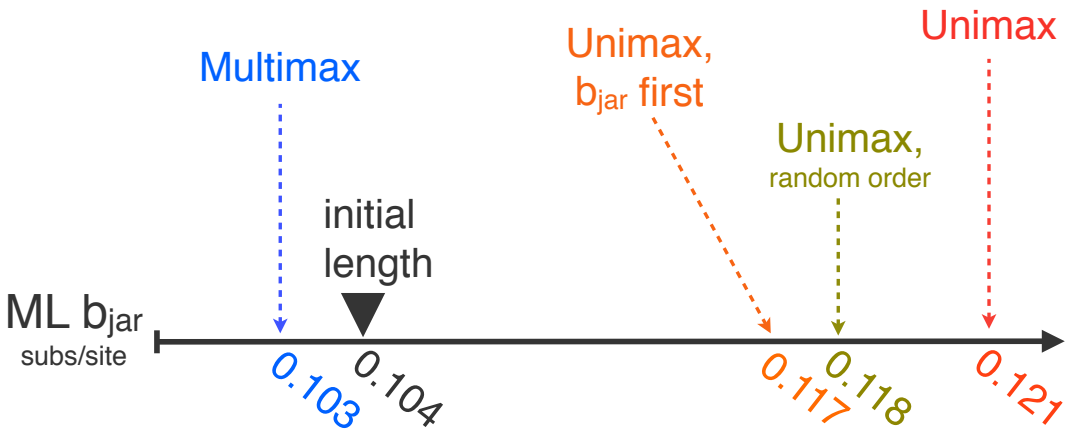


FIGURE 16. Unimax suffers from an order-dependence problem. Multimax and three versions of Unimax arrived at different ML lengths for the branch b_{jar}, attached to the sequence AngJapARb in Fig. 4.. The initial length of b_{jar} was determined by the neighbor-joining algorithm.

b_{jar}, attached to the protein sequence for the *Anguilla japonica* androgen receptor beta (AngJapARb) in the tree of steroid hormone receptor proteins (Fig. 4.). The branch b_{jar} began with the length 0.104 substitutions per site (subs/site) in the initial neighbor-joined tree (Fig. 16.). MM reduced b_{jar} to 0.13 subs/site. UM, in contrast, first optimized branches 1 through 444, then increased b_{jar} (numbered 445) to length 0.121 subs/site, and then optimized branches numbered 446 through 473. I modified UM to optimize b_{jar} first, and then optimize the other branches afterwards. This version of UM arrived at an ML length for b_{jar} of 0.117 subs/site. I next modified UM to optimize branch lengths in a random order. This version of UM optimized b_{jar} to 0.118 subs/site. The fact that three different UM orderings arrived at three different ML lengths for b_{jar} indicates that UM's ML solutions depend on the order in which parameters are sequentially optimized.

Effect of Multidimensional Optimization on Ancestral Sequence Reconstruction

In order to determine if UM's assumption of parameter separability has downstream consequences for ancestral sequence reconstruction, I inferred ancestral sequences on the UM and MM ML trees for the steroid hormone receptor alignment. For several ancestors of interest, I compared their ML states and their statistical support on the UM and MM ML trees.

I observed that UM and MM ML trees yielded significantly different ancestral sequences. For the last-shared ancestor of all Androgen receptors (AncAR, node251), the ancestral sequences disagreed at 5% (17 of 359) of the sites (Fig. 17.). Twelve of these sites form a motif at the beginning of the AR sequence; this motif is missing in sharks, skates, gar, and other early-branching fish. The inferred presence or absence of this motif in ancestral sequences is primarily contingent on

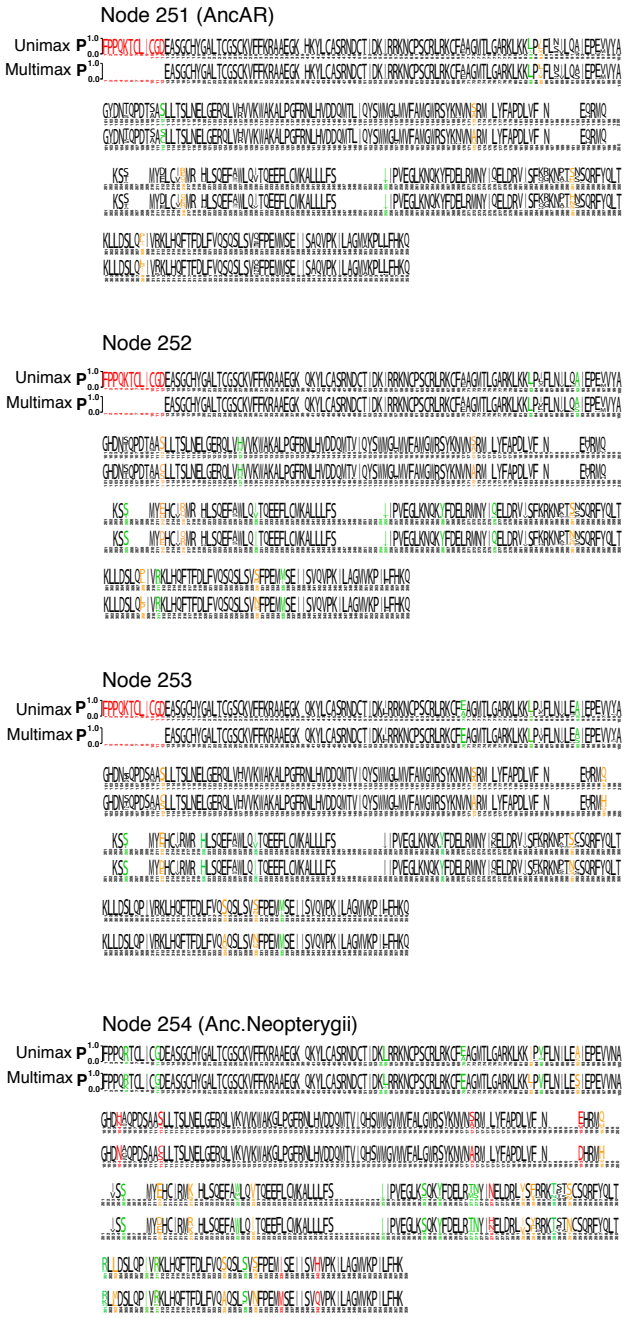


FIGURE 17. Reconstructed ancestral androgen receptor sequences. Node names correspond to labels in Fig. 4.. The height of each character expresses its posterior probability at the indicated site. Empty sites indicate insertions or deletions. **Red** indicates sites where the ML states on the MM tree and UM tree disagree. **Orange** indicates sites where the ML states disagree between trees, but both sequences infer complementary low support ($\leq$0.5%) for alternative state. **Green** indicates sites where the ML states agree, and one state vector – but not the other – introduces significant uncertainty about the ML state.

the phylogenetic placement of gar fish. The UM tree placed gar deep inside the Neopterygii clade, and AncAR was therefore reconstructed to contain the motif. On the MM tree, however, gar was placed in a more basal position and AncAR was reconstructed to not contain this motif. The UM and MM trees also disagreed about the presence of this motif at two nodes descendant from AncAR (nodes 252 and 253). Aside from this motif, the last-shared ancestor of Neopterygii (node 254) was reconstructed with nineteen disagreeing sites between the UM and MM trees. At seven of these sites, the dissenting state was not significantly supported in the reconstruction on the other tree.

The role of line search, versus quadratic interpolation

My particular implementation of MM using BFGS differs from UM using Brent's method not only in their treatment of parameter separability. Specifically, BFGS and Brent's method use different mechanisms to move forward in parameter space. BFGS uses line search to move in a direction informed by the functional gradient, whereas Brent's method uses quadratic interpolation to move in single dimensions. I determined that this difference is not the primary reason that UM and MM drove the tree search to different ML trees. I implemented an alternative version of UM that uses line search, such that UM (with line search) and MM would be identical except in their assumption of separability. I used this alternative version of UM to infer ML phylogenies for the six empirical alignments in this study. I observed that UM using line search and UM using quadratic interpolation led to the same ML topologies for all six empirical alignments. Further, the final ML values between the two methods differed by less than 0.0001 log likelihood units. This means that the use of line search, rather than quadratic interpolation, is

not the reason that UM and MM led to different ML trees. This result indicates that UM's assumption of parameter separability is the primary cause of UM's phylogenetic impairment.

Discussion

My results demonstrate that UM's assumption of parameter separability impairs its ability to find accurate ML phylogenies. By using an alternative ML algorithm that does not assume separability, I was able to find more accurate and higher-likelihood trees. UM repeatedly made suboptimal topology choices during its search of tree space. UM ML branch lengths are less accurate than MM ML branch lengths, and this difference is likely to be a mechanism by which UM-driven tree search choose poor topologies. The assumption of parameter separability not only affects phylogenetic accuracy, it also affects the downstream inference of ancestral states.

Prior work has shown that ML phylogenetic inference is NP-hard (Chor and Tuller (2005); Roch (2006)). Consequently, I know with certainty that no ML search heuristic running in polynomial time – including UM and MM – can be guaranteed to find the true ML phylogeny for every alignment. Indeed, I observed that UM and MM were not accurate all the time, but MM was more accurate more often.

An open question in phylogenetics is whether likelihood landscapes are simple or complex (Fukami and Tateno (1989); Steel (1994); Rogers and Swofford (1999); Yang (2000); Chor et al. (2000); Billera et al. (2001)). I observed that UM and MM arrived at unique ML trees, and at unique ML branch lengths for fixed trees. These results indicate that the phylogenetic likelihood landscape is complex on both the

space of topologies and also on the subspace of continuous model parameters, at least for the conditions studied here.

Although MM has been previously implemented in software for ML phylogenetics, its performance heretofore has not been rigorously investigated. The software suite PAML implements MM using BFGS as an option called "simultaneous update." PAML provides this option as a sort of algorithmic curio, without any deeper analysis of the algorithm's efficacy and accuracy. Because BFGS – and MM in general – has not been deeply studied as a tool for phylogenetics, there has not been any real motivation for evolutionary biologists to use PAML's implementation of BFGS. Rather, the scientific community has continued to use the default UM implementations provided by PhyML, RaxML, and GARLI (Guindon et al. (2010); Stamatakis (2006); Zwickl (2006)); perhaps this is because PhyML, RaxML, and GARLI provide other useful features, including support for large sequence datasets and command-line facilities for high-throughput batch analysis. I implemented our own version of BFGS, rather than using PAML's code, in order to ensure computational efficiency and in order to instrument the algorithm to report useful metrics, including likelihood gradients and neighbor-swap tree scores. Our MM implementation is built within the open source code for PhyML. I encourage you to use our software. It is available at the following URL: http://markov.uoregon.edu/software/m3l.

CHAPTER III

ANCESTRAL RECONSTRUCTION AND TREE UNCERTAINTY

In this chapter, I show that ancestral sequence reconstruction is robust to phylogenenetic uncertainty. Specifically, I discuss the relationship between poorly-supported phylogenies and the downstream accuracy of ancestral reconstruction. This work was previously published in the Oxford Journal Molecular Biology and Evolution (Hanson-Smith et al. (2010)).

Ancestral sequence reconstruction (ASR) is widely used to formulate and test hypotheses about the sequences, functions, and structures of ancient genes. Ancestral sequences are usually inferred from an alignment of extant sequences using a maximum likelihood (ML) phylogenetic algorithm, which calculates the most likely ancestral sequence assuming a probabilistic model of sequence evolution and a specific phylogenytypically the tree with the ML. The true phylogeny is seldom known with certainty, however. ML methods ignore this uncertainty, whereas Bayesian methods incorporate it by integrating the likelihood of each ancestral state over a distribution of possible trees. It is not known whether Bayesian approaches to phylogenetic uncertainty improve the accuracy of inferred ancestral sequences. Here, I use simulation-based experiments under both simplified and empirically derived conditions to compare the accuracy of ASR carried out using ML and Bayesian approaches. I show that incorporating phylogenetic uncertainty by integrating over topologies very rarely changes the inferred ancestral state and does not improve the accuracy of the reconstructed ancestral sequence. Ancestral state reconstructions are robust to uncertainty about the underlying tree because the conditions that produce phylogenetic uncertainty also make the

ancestral state identical across plausible trees; conversely, the conditions under which different phylogenies yield different inferred ancestral states produce little or no ambiguity about the true phylogeny. These results suggest that ML can produce accurate ASRs, even in the face of phylogenetic uncertainty. Using Bayesian integration to incorporate this uncertainty is neither necessary nor beneficial.

The properties and evolution of ancient genes and proteins can seldom be directly studied, because such molecules are rarely preserved intact over very long periods of time. In 1963, Pauling and Zuckerkandl proposed that ancestral molecules could one day be "resurrected" by inferring their sequences and then synthesizing them (Pauling and Zuckerkandl (1963)). Decades later, the methods of ancestral sequence reconstruction (ASR) have emerged as important tools for examining the trajectory of molecular sequence evolution and testing hypotheses about the functional evolution of ancient genes (Thornton (2004); Liberles (2007); Dean and Thornton (2007)). Among numerous examples, ASR has been used in the last decade to investigate the evolution of elongation-factor proteins (Gaucher et al. (2003), Gaucher et al. (2007)), steroid hormone receptors (Thornton et al. (2003), Bridgham et al. (2006), Ortlund et al. (2007)), visual pigments (Shi and Yokoyama (2004); Chang et al. (2002)), fluorescent proteins (Ugalde et al. (2004)), and alcohol dehydrogenases (Thomson et al. (2005)).

Although the first ASR practitioners used parsimony methods (e.g., Jermann et al. (1995)), most modern studies use maximum likelihood (ML) (Yang et al. (1995); Koshi and Goldstein (1996); Pupko et al. (2000)). ML begins with an alignment of extant gene sequences, a phylogeny relating those sequences, and a statistical model of evolution. For each internal node in the phylogeny and each site in the sequence, the likelihood of each possible ancestral state—defined as the

probability of observing all the extant states given that ancestral state, the tree, and the model—is calculated. The ML ancestral state is the state with the highest likelihood. Confidence in any ancestral state inference is typically expressed as its posterior probability, defined as the likelihood of the state (weighted by its prior probability) divided by the sum of the prior-weighted likelihoods for all states.

The ML approach to ancestral reconstruction assumes that the alignment, tree, model, and model parameters are known *a priori* to be correct. In practice, this assumption is often not valid; for many real-world datasets, alternatives to the ML tree and parameter values cannot be ruled out. To accommodate these sources of uncertainty, Bayesian methods have been proposed. Whereas ML assumes the most likely estimate of the tree and model parameters, Bayesian approaches incorporate uncertainty by summing likelihoods over a distribution of possible trees or parameter values, each weighted by its posterior probability. Pagel et al. proposed a Bayesian method for integrating topological uncertainty into inference of ancestral states for binary and other discrete characters (Pagel et al. (2004)). Schultz and Churchill proposed a Bayesian method to integrate uncertainty about the parameters of the evolutionary model into discrete character reconstructions (Schultz and Churchill (1999)). For inference of ancestral DNA and protein sequences, Huelsenbeck and Bollback developed a Bayesian method to integrate uncertainty about the tree topology, branch lengths, and model parameters (Huelsenbeck and Bollback (2001)).

It is not known how Bayesian approaches affect the accuracy of reconstructed ancestral sequences. Here I focus on the specific effects of one source of uncertainty—the phylogeny. There have been a few attempts to characterize the robustness of reconstructed ancestral sequences with respect to phylogenetic

uncertainty in specific cases: Gaucher et al. reconstructed ancestral elongation factor proteins on two plausible phylogenies (Gaucher et al. (2003)), and Bridgham et al. reconstructed the ancestral corticosteroid receptor on all trees within the 95% confidence interval from a Bayesian phylogenetic analysis (Bridgham et al. (2006)). In both cases, the maximum a posteriori ancestral sequences changed very little when different phylogenies were assumed, and the functions of the reconstructed proteins in experimental assays were also unchanged. Huelsenbeck and Bollback used simulations to show that integrating uncertainty about the phylogeny, branch lengths, and model parameters can affect the posterior probabilities of ancestral states (Huelsenbeck and Bollback (2001)), but they did not study the effect of integration on the inferred maximum a posteriori state or the accuracy of those inferences.

To determine the causal effects of integrating over phylogenetic uncertainty on ASR accuracy, I implement a topological empirical Bayesian method for ancestral reconstruction that is identical to the ML algorithm, except that it integrates over topologies. This approach allows us to directly infer the effects of incorporating phylogenetic uncertainty on ASR accuracy. I simulate and record the evolution of sequences under a variety of simplified and empirically derived conditions and infer ancestral states from the evolved alignments, allowing us to characterize the accuracy of each approach to ASR by comparing inferred ancestral sequences to the "true" ancestors recorded during the simulation.

Materials and Methods

Ancestral State Reconstruction Algorithms

The ML method for ancestral sequence reconstruction, also called the empirical Bayes method (Yang et al. (1995)), calculates the posterior probability that some ancestral node contained state a at a sequence site of interest, given the observed sequence data d, an evolutionary model m, a topology $\hat{t}$, and a set of branch lengths and other model parameters $\hat{\theta}$; the topology and parameters are those that maximize the likelihood over all data columns in the alignment. The conditional likelihood of a equals the probability of observing d given a, m, $\hat{t}$, and $\hat{\theta}$. The prior-weighted conditional likelihood of a is the conditional likelihood of a multiplied by the prior probability of observing a, which is given by π_a, the equilibrium state frequency of a. The posterior probability of a equals the prior-weighted conditional likelihood of a divided by the sum of the prior-weighted conditional likelihoods for all possible ancestral state assignments (4 for nucleotides or 20 for amino acids) (Equation 3.1).

$$P(a|d,m,\hat{t},\hat{\theta}) = \frac{P(d|a,m,\hat{t},\hat{\theta})\pi_a}{\sum_a P(d|a,m,\hat{t},\hat{\theta})\pi_a} \tag{3.1}$$

The ML state assignment is the state with the highest prior-weighted likelihood (and necessarily the highest posterior probability, as well). The ML sequence is the string of ML states. To reconstruct ML ancestral sequences, I used PAML v.4.1 (Yang (1997, 2007)).

The Topological Empirical Bayes (TEB) approach to ASR differs from ML only by integrating ancestral reconstructions over a distribution of trees (Equation 2). The TEB posterior probability of ancestral state a is the weighted average of

the posterior probability of a over all possible trees, where the weights are given by the empirical Bayes posterior probability of each tree t. The empirical Bayes posterior probability P_{EB} of a tree assumes the maximum likelihood estimate of branch lengths and other model parameters $\hat{\theta}_t$ on each tree (Kolaczkowski and Thornton (2008), Kolaczkowski and Thornton (2009)):

$$P_{TEB}(a|d,m) = \sum_t P(a|d,t,m,\hat{\theta}_t) \times P_{EB}(t|d,m,\hat{\theta}_t) \tag{3.2}$$

Equation 3.2 takes a different form from but is equivalent to (see Supplemental Note 1) the expression used by others (Pagel et al. (2004), Huelsenbeck and Bollback (2001)) for ancestral state reconstructions integrated over topologies:

$$P_{TEB}(a|d,m) = \frac{\sum_t P(d|a,t,m,\hat{\theta}_t)\pi_a P(t)}{\sum_t \sum_a P(d|a,t,m,\hat{\theta}_t)\pi_a P(t)} \tag{3.3}$$

The ML method also has an empirical Bayesian interpretation, because Equation 3.1 calculates a posterior probability and uses priors on ancestral states. For simplicity, I will refer to the approach which uses only the ML tree as the "ML method" and the approach which integrates over trees as the "TEB method."

One issue with estimating ancestral states from a distribution of trees is that every topology contains different ancestral nodes. I accommodate this problem by defining an ancestral node to be reconstructed as the most recent common ancestor (MRCA) of a specified set of descendants (Pagel et al. (2004)). On any rooted tree, the clade descending from the specified ancestor will contain all members of this set; additional sequences may also be included in that clade, depending on the topology. A similar approach can be used to describe internal nodes on unrooted

trees in relation to the split that places a specified set of terminal sequences into the smallest possible partition of the tree.

I implemented both the TEB and ML method in a new software package called *Lazarus*. This package spawns, manages, and then parses large batches of parallelized PAML jobs, one for each of a set of user-specified topologies. For each topology, branch lengths and model parameters are optimized by ML, the maximum likelihood of the tree is calculated, and the posterior probability of each ancestral state is calculated on that topology. *Lazarus* then parses these results to calculate the posterior probability of each ancestral state integrated over topologies. Lazarus includes a modular Python API with object classes for quickly abstracting ancestral reconstruction data and is available at

Simulations

I compared the ancestral states reconstructed by the ML and TEB methods on data simulated under both controlled and empirically-derived conditions. The correct evolutionary model was assumed for all ancestral reconstructions.

Four-Taxon Phylogenetic Uncertainty

I simulated sequence evolution on four-taxon ultrametric trees of variable height and internal branch length (Fig. 18.A) and on four-taxon trees with randomly generated branch lengths. I examined ultrametric trees because they can be described by specifying only the total height of the tree and the lengths of the internal branches; the limited number of free parameters allows a detailed investigation of ancestral reconstruction methods as phylogenetic signal varies. Further, ultrametric trees represent the most difficult conditions for ancestral

sequence reconstruction. For a pair of terminal branches with any given sum of lengths descending from an internal node, the ultrametric case represents the greatest total loss of character information about the ancestor; conversely, as some branches descending from an ancestral node become longer and others shorter, the information in the short branch has a more determinative effect on the inferred ancestral state. In the limit as one descendant branch length approaches zero, the ancestral state is inferred without ambiguity or error as the state in the sequence at the end of that branch.

On ultrametric trees, the internal branch length (labeled 'r' in Fig. 18.A) was varied from (0.01, 0.02, 0.03, 0.05, 0.1, 0.2), and the overall height of the descendant clade (labeled 'h' in Fig. 18.A) varied from 0.25 to 0.75 substitutions per site in intervals of 0.125. For each combination of 'r' and 'h,' I used Seq-Gen (Rambaut and Grassly (1997)) to generate 100 sets of replicate descendant amino acid sequences of length 400 sites, using the JTT evolutionary model (Jones et al. (1991)). For the non-ultrametric simulations, 1000 four-taxon trees were generated by randomly drawing an internal branch length from the uniform distribution $U[0.01, 0.1]$ and drawing four terminal branches from the uniform distribution $U[0.25, 0.75]$. Seq-Gen was then used to simulate the evolution of sequences 400 amino acids long on each tree (Fig. 18.C).

For each replicate, I used ML and TEB ASR to infer the posterior probability of reconstructed ancestral states in the most-recent-common ancestor of taxa $\{A, B, C\}$, of $\{A, B\}$, of $\{A, C\}$, and of $\{B, C\}$. Depending on the tree, some of these ancestors are the same. For example, on the tree $(((A, B), C, D)$, the ancestor of $\{A, C\}$ is the same node as the ancestor of $\{A, B, C\}$. However, on tree $(((A, C), B, D)$, the ancestors for $\{A, B, C\}$ and $\{A, C\}$ are unique. I compared the maximum a posteriori ancestral state from TEB and ML to each other and to the true state, which was recorded at all nodes during the simulation. I analyzed the concordance and accuracy of TEB and ML ancestral states across all replicates and in relation to the values of 'r' and 'h,' the state pattern in descendant taxa, and whether the set of taxa in the clade descending from the ancestral node of interest in the ML tree is identical to that set in the true tree. With respect to the last criterion, the membership may be correct, a spurious taxon may be included as a

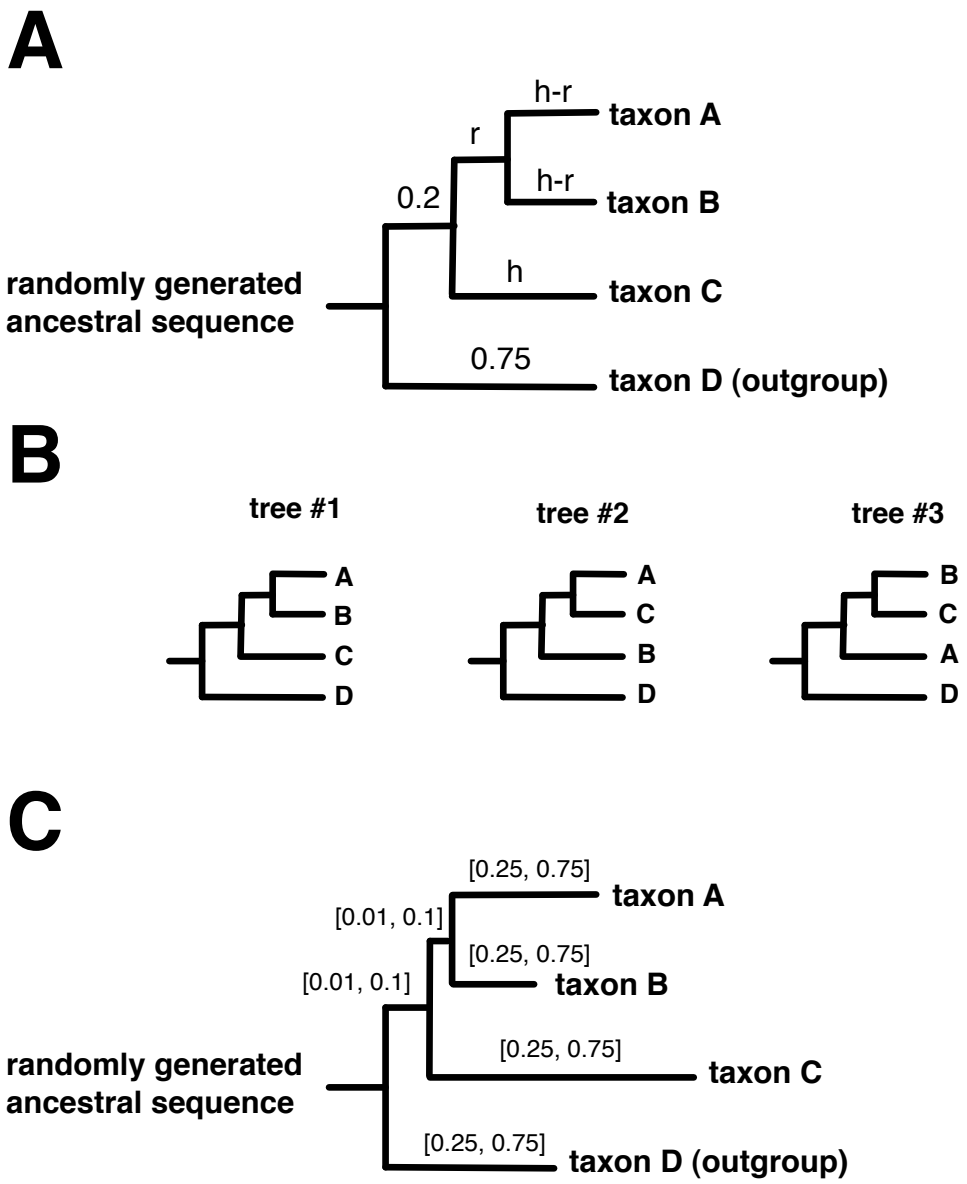


FIGURE 18. Four-taxon simulation conditions. **(A)** I seeded randomly-generated amino acid sequences at the root of an ultrametric tree with four terminal branches. I simulated the ancestral sequences evolving across the branches to produce four descendant sequences (including one outgroup descendant). Simulations were performed under a variety of conditions by adjusting the internal branch length r and the overall height of the descendant clade h. **(B)** For each set of replicate sequences, I estimated the ML branch lengths and calculated the posterior probability of all three possible topologies. **(C)** Sequences were also simulated using non-ultrametric four-taxon trees with terminal branch lengths drawn from the uniform interval $[0.25, 0.75]$ and internal branch lengths from the interval $[0.01, 0.1]$.

descendant (mem+), or a taxon may be incorrectly excluded from the clade (mem-).

Empirically Derived Phylogenetic Uncertainty

I also compared the accuracy of ML and TEB reconstructions inferred from sequences simulated on empirically-derived trees. I used phylogenies inferred from the extant sequences of alcohol dehydrogenase (ADH) proteins (Thomson et al. (2005)), steroid hormone-receptors (Bridgham et al. (2006)), green fluorescent-like proteins (GFP) (Kelmanson and Matz (2003); Ugalde et al. (2004)), and Tu family elongation factor (EF-Tu) proteins (Gaucher et al. (2003)). For each gene family, the phylogeny and branch lengths were calculated by ML using Phyml version 2.4.4 (Guindon and Gascuel (2003)). The posterior probabilities of phylogenies in the 95% credible set (1,195 trees for ADH, 3,335 for steroid hormone receptors, 655 for GFP, and 544 for EF-Tu) were inferred using empirical Bayes Markov Chain Monte Carlo (BMCMC), which integrates over topologies, each of which is assigned its maximum likelihood branch lengths (Kolaczkowski and Thornton (2007)). The ML phylogenies for AFH, GFP, and EF-Tu (Fig. 19.) differ only slightly from the original ML phylogenies shown in those datasets' corresponding publications. On each ML phylogeny, 100 replicates of protein sequences 400 amino acids long were then evolved by simulation, using the JTT model of evolution, to yield terminal descendant sequences. For each replicate, ancestral sequences at all internal nodes were then reconstructed using ML and TEB. I examined only the uncertain nodes (with Bayesian posterior probability less than 1.0) and their immediate neighboring nodes; nodes with $PP = 1.0$ have no uncertainty over which to integrate, and therefore the TEB and ML reconstructions are identical.

State Pattern Analysis

To illustrate how integrating over topologies affects ancestral reconstruction for different data patterns under specific conditions, I performed ASR using ML and TEB and calculated the probability of each ancestral state for each of the possible state patterns of four nucleotides. I simulated DNA sequences 50,000 nucleotides long using the JC69 model on four taxon ultrametric trees with high phylogenetic uncertainty (h=0.3, r=0.01) or virtually no phylogenetic uncertainty (h=0.3, r=0.2). I then examined the posterior probability of each ancestral state inferred using ML and TEB for each of the possible state patterns for four-state data. Character state patterns are indicated using variables representing nucleotides of the same type: for example, pattern xyxy for the four-taxon case stands for the realizations ACAC, AGAG, ATAT, CACA, ...TGTG at that site in the four leaves, respectively.

Statistical Analysis

The correspondence between posterior probabilities (PPs) and the frequency of correct inferences for TEB and for ML were analyzed by binning inferences according to their PPs and calculating the mean PP (x) and the fraction of correct reconstructions (y) in each bin. The fit of the resulting points to the function y=x was evaluated using a chi-square distribution with degrees of freedom equal to the number of bins. The significance of the difference between ML and TEB in fit to the function y=x was assessed by evaluating the ratio of the chi-square statistics for the two methods using an F-distribution with degrees of freedom equal to the number of bins. To compare the differences in mean accuracy of the ML and TEB

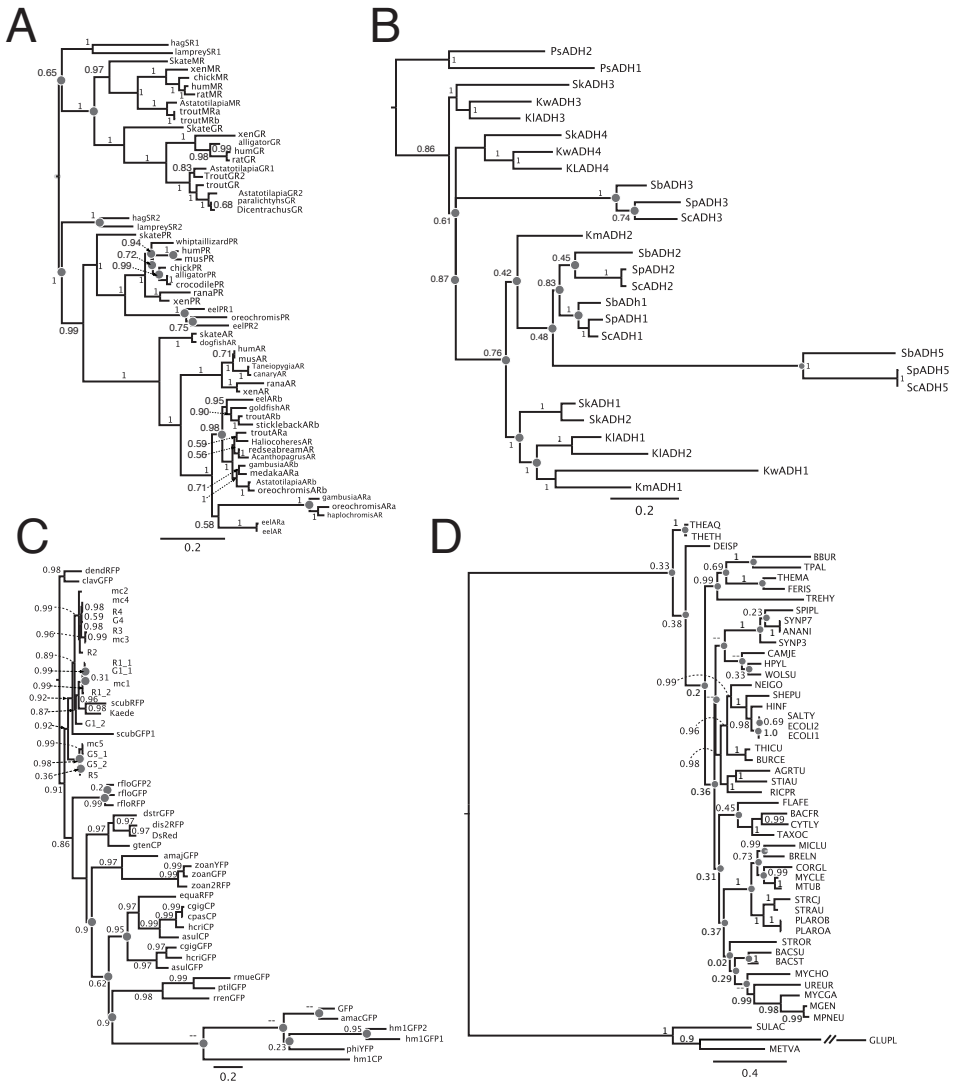


FIGURE 19. Empirical phylogenies used for simulations. Internal nodes are labeled with their empirical Bayes posterior probability; circles indicate nodes at which ancestral sequences were reconstructed. A) Steroid hormone receptors (Bridgham et al. (2006)). The tree and branch lengths were inferred from empirical protein sequences using the JTT+G model. B) Alcohol dehydrogenases (Thomson et al. (2005)). The tree and branch lengths were inferred from empirical DNA sequences using GTR+G. C) Green fluorescent-like proteins Kelmanson and Matz (2003). The tree and branch lengths were inferred from empirical DNA sequences using GTR+G. D) EF-Tu family elongation factors (Gaucher et al. (2003)). The tree and branch lengths were inferred from empirical protein sequences using JTT+G.

reconstructions, I conducted a paired two-sample t-test against the null hypothesis of no significant difference in accuracy between the two methods.

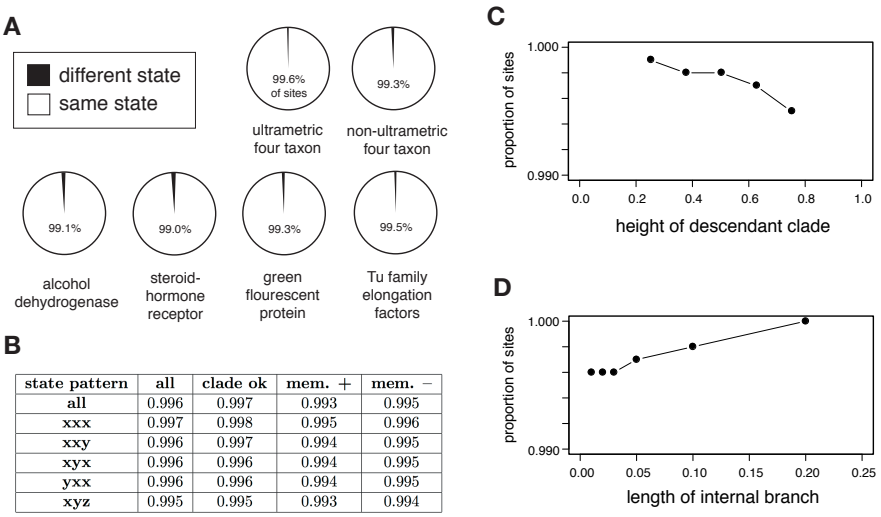
Results

Effect of Incorporating Phylogenetic Uncertainty

To determine how incorporating topological uncertainty affects ancestral sequence reconstruction, I first examined the extent to which ancestors inferred using ML and TEB differ from each other under a range of conditions. I found that integrating over trees only rarely affected the inferred state at ancestral nodes (Fig. 20.A). In simulations on ultrametric four-taxon trees with varying levels of phylogenetic noise, the ancestral states inferred by ML and TEB differed at only 0.4% of sites. On non-ultrametric trees, they differed at 0.7% of sites. On larger trees derived from empirical datasets of four gene families previously analyzed using ASR—steroid hormone receptors, alcohol dehydrogenases, green fluorescent-like proteins, and Tu family elongation factors—ML and TEB reconstructions differed by one percent or less (Fig. 20.A).

To determine whether certain phylogenetic conditions cause integrating over topological uncertainty to have a stronger effect on inferred ancestral states, I decomposed the results of the ultrametric four-taxon simulations according to the state patterns in the terminal sequences that descend from the reconstructed ancestor, the length of the branches on the tree, and the ways (if any) that the ML tree differs from the true tree (Supplemental Table 2). There were no state patterns that resulted in differences between ML and TEB ancestors greater than 0.5 percent. The effect of integrating over uncertainty was slightly greater for divergent state patterns in which all ingroup descendants have different states (pattern *xyz*) than for patterns that contain phylogenetic signal (*xxx* or *xxy*, Fig. 20.B). Similarly, no branch length conditions examined caused ML and TEB to differ by more than 0.5 percent; ML and TEB ancestors differed least when the total root-to-tip branch length was short, and they differed to a slightly greater extent as the terminal branches became very long (Fig. 20.C). When the ML tree was correct (as it was in the majority of cases), integrating over uncertainty had a particularly weak effect on the inferred ancestor; however, even when the ML phylogeny erroneously inferred a spurious sequence as a descendant of the ancestor of interest or excluded a true descendant, the two methods still produced identical inferences at $> 99\%$ of sites (Fig. 20.B). Together, these data indicate that integrating over topological uncertainty per se does not strongly affect ancestral reconstructions; the effects are weak under conditions that cause the traces of the ancestral state to be lost in descendant sequences and virtually non-existent under those that preserve phylogenetic signal about the ancestral state.

I next analyzed whether integrating over topological uncertainty tends to affect sites that are strongly or weakly supported by ML. Most ASR practitioners



state pattern	all	clade ok	mem. +	mem. −
all	0.996	0.997	0.993	0.995
xxx	0.997	0.998	0.995	0.996
xxy	0.996	0.997	0.994	0.995
xyx	0.996	0.996	0.994	0.995
yxx	0.996	0.996	0.994	0.995
xyz	0.995	0.995	0.993	0.994

FIGURE 20. Integrating over phylogenetic uncertainty rarely changes ancestors. **(A)** Proportion of sites simulated under a variety of conditions at which ML and TEB methods inferred the same or different states. **(B, C, D)** Details of similarity between ML and TEB reconstructions for the ultrametric four-taxon simulations. **(B)** Proportion of sites at which ML and TEB infer identical states is shown in terms of descendant state patterns and types of phylogenetic error. Each row presents results for sites in which the descendant taxa A, B, and C have the specified state pattern (where pattern *xxx* corresponds to AAA, CCC, GGG, or TTT; *xxy* corresponds to AAC, AAG, AAT, ... or TTG). Columns indicate whether the set of taxa descending from the reconstructed node in the ML tree corresponds to those in the true tree: *clade ok* means the descendant membership is correct, *mem. +* means the ML descendant set spuriously includes an extra taxon, and *mem.* – means the ML descendant set incorrectly excludes a taxon. **(C)** Similarity between ML and TEB reconstructions is plotted against the height of the descendant clade ("h" in Fig. 18.). **(D)** Similarity between ML and TEB reconstructions is shown versus the length of the internal branch ("r" in Figure 1).

examine the support for ancestral state inferences and experimentally characterize the robustness of their inferences to alternate reconstructions that have posterior probability above some defined plausibility cutoff (Bridgham et al. (2006); Ortlund et al. (2007); Thomson et al. (2005); Chang et al. (2002); Ugalde et al. (2004)). I found that ML and TEB reconstructions disagreed only at sites that were already ambiguous in the ML reconstruction (Fig. 21.). In both ultrametric and non-ultrametric four-taxon simulations, the ML and TEB reconstructions agreed at all sites at which the ML reconstruction had posterior probability (PP) greater than 0.70. In the ADH, GFP, and EF-Tu simulations, the two methods agreed at all sites with PP greater than 0.76, 0.63, and 0.71, respectively. In the steroid hormone simulation, the methods agreed at all sites with PP greater than 0.87, and they disagreed at only 0.003% of all sites reconstructed with posterior probability > 0.80. Over all four-taxon reconstructions, the maximum a posteriori ancestral state from TEB was different from the first- or second-best state using the ML method at only 0.001625% of sites. These data indicate that integrating over topological uncertainty never causes inferred ancestral states that are strongly supported by ML to be revised. Rather, TEB inferred a state different from the ML state only when that state was ambiguously reconstructed anyway, switching the favored state from one weakly supported possibility to another.

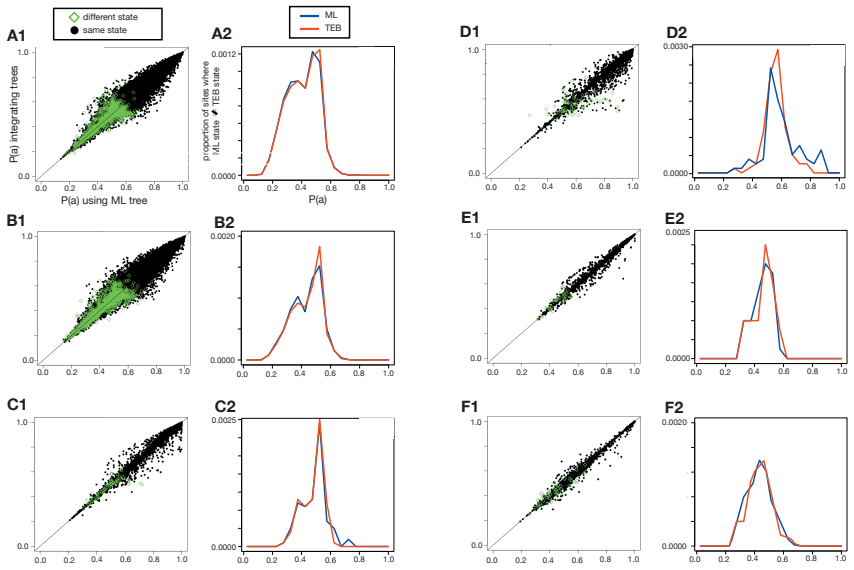


FIGURE 21. ML and TEB differences versus phylogenetic support. ML and TEB infer different ancestral states only when posterior probabilities are low. In each pair of plots, the left plot (A1, B1, etc.) compares the posterior probability of the maximum a posteriori state inferred by ML to that inferred by TEB. Black points show sites at which ML and TEB methods inferred the same state; green diamonds indicate that the two methods inferred different states. The right plots (A2, B2, etc.) are histograms of the green points in the left plot: I grouped all ASR inferences into 5%-sized bins based on their posterior probability and counted the proportion of sites at which ML and TEB inferred different states. Results are shown for simulations on ultrametric four-taxon trees **(A1,A2)**, non-ultrametric four-taxon trees **(B1,B2)**, and the steroid-hormone receptor **(C1,C2)**, ADH **(D1,D2)**, GFP **(E1,E2)**, and EF-Tu phylogenies **(F1,F2)**.

Effect of Incorporating Phylogenetic Uncertainty on ASR Accuracy

Although the ML and TEB methods inferred the same state at most sites, it is possible that TEB might produce more accurate reconstructions at the rare sites where the two methods differ. I measured accuracy as the proportion of sites at which the reconstructed state was identical to that of the true ancestor, which I recorded during each simulation. In the four-taxon and GFP simulations, ML was

slightly, but not significantly, more accurate than TEB (Fig. 22.A, Supplemental Table 7). In the ADH, steroid hormone receptor, and EF-Tu simulations, there was no difference in accuracy between the methods. The accuracy of both ML and TEB declined as terminal branch lengths grew longer, causing multiple substitutions to occur (Fig. 22.B). ML's superiority to TEB was greatest when the membership of the descendant clade was correct (Fig. 22.C), presumably because when the ML topology is the true tree, integrating phylogenetic uncertainty serves only to introduce error. Even when the ML tree was incorrect, however, TEB generally decreased accuracy; integrating over uncertainty improved accuracy only under the rare condition that the descendant state pattern was textitxyz and a spurious taxon had been included as a descendant of the node of interest. Under these conditions, both methods performed poorly, because little or no phylogenetic signal of the ancestral state was retained in the descendants. For all other state patterns and forms of phylogenetic error, ML had accuracy equal to or slightly greater than that of TEB.

Effect of Incorporating Phylogenetic Uncertainty on ASR Posterior Probabilities

I next examined whether TEB or ML yielded more accurate estimates of statistical confidence in inferred ancestral states. For all simulations, I binned reconstructed ancestral sites by their posterior probability and counted the proportion of accurate inferences in each bin (Fig. 23.). If posterior probability is an accurate predictor of the probability that an inferred state is correct, the mean PP in that bin should equal the proportion of correct ancestral state inferences. I observed that the ML and TEB methods generally produced similar PP values, and both types of PP were good predictors of mean accuracy. The major exception

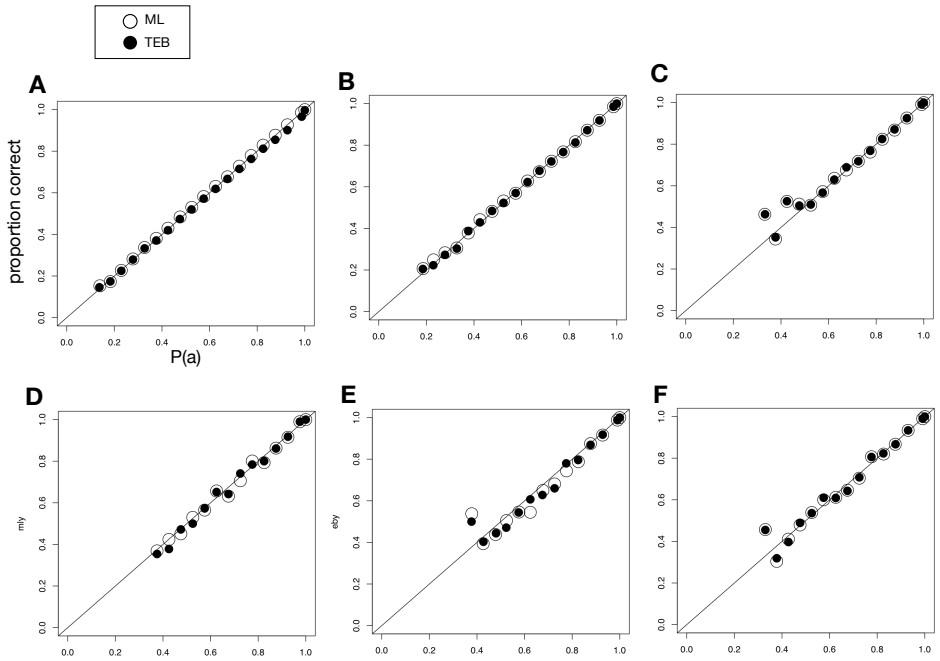
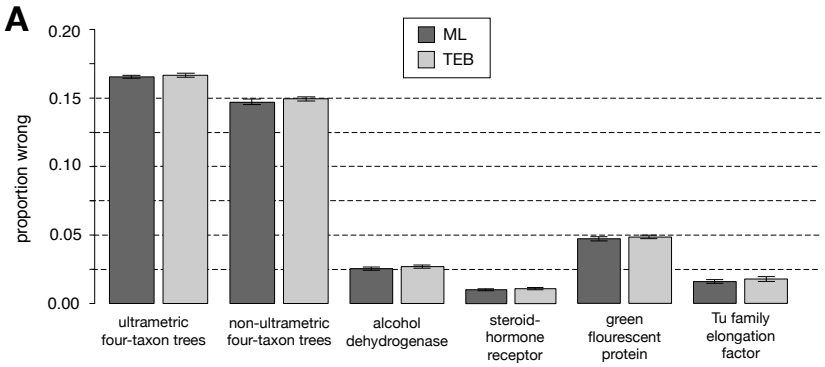


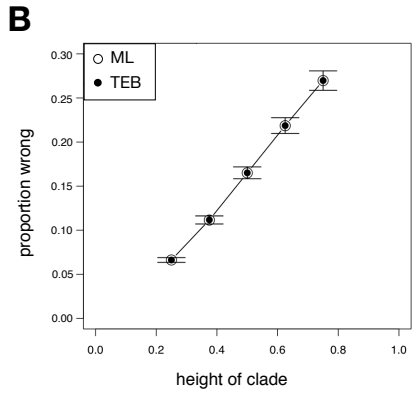
FIGURE 22. ASR error rates, measured as the proportion of sites at which the maximum a posteriori reconstructions differ from the true ancestral state. **(A)** Results from the four-taxon and empirically-derived conditions are averaged over all replicates. None of the differences between ML and TEB are statistically significant. **(B)** Results from the ultrametric four-taxon simulation are shown versus the height of the descendant clade (where height equals "h" in Fig. 18.. Error bars for ML and TEB are nearly identical. **(C)** Detailed results from the ultrametric four-taxon simulation. Each cell reports two values: the proportion of sites incorrectly reconstructed by ML (top) and TEB (bottom). Bold values indicate the method with higher accuracy. Data are sorted according to the same criteria in Fig. 20.B.

to this pattern was the four-taxon simulation on ultrametric trees, in which integrating over trees slightly inflated support for reconstructions with PP > 0.5 (Fig. 23.A); a chi-square test indicates that ML's posterior probabilities fit the ideal better than TEB's PPs do, but the difference is small and does not reach statistical significance ($P = 0.16$, Supplemental Table 1). When the ML tree was correct, ML's PPs were more accurate than TEB, but TEB was more accurate when the ML tree was wrong; because the former conditions are more frequent than the latter, however, ML's accuracy was higher overall. For the empirically derived conditions, ML's PPs were slightly more accurate, but the difference was again small and not statistically significant (Supplemental Tables 3, 4, 5, 6).

An Intrinsic Tradeoff Explains Why Incorporating Uncertainty Does Not Affect ASR

In order to understand why integrating over phylogenies has such a weak effect on ancestral reconstruction, I examined the relationship between the plausibility of alternate phylogenies and the dependence of the reconstructed state on the assumed phylogeny. I conjectured that as phylogenetic uncertainty increases, the same state will be reconstructed on the plausible trees. To test this hypothesis, I grouped all the replicates from the ultrametric four-taxon simulations according to the posterior probability of their ML tree. For each replicate, I counted the proportion of sites at which the inferred ancestral state differs between the ML tree and the tree with the next-highest PP (Fig. 24.A). I observed that when the ML tree was uncertain (PP <1.0), the ancestral states among trees rarely disagreed. In contrast, when the ML tree was absolutely certain (PP = 1.0), the ancestral states on the ML tree and the second-best tree disagreed at up to 25%





C

state pattern	all	clade ok	mem. +	mem. −
all: ML	**0.158**	**0.143**	0.178	0.182
all: TEB	0.160	0.145	0.178	0.182
xxx: ML	0.009	0.010	0.004	0.004
xxx: TEB	0.009	0.010	0.004	0.004
xxy: ML	**0.148**	**0.148**	0.133	0.177
xxy: TEB	0.162	0.161	0.133	0.177
xyx: ML	0.108	0.111	0.102	0.102
xyx: TEB	0.108	0.111	0.102	0.102
yxx: ML	**0.114**	**0.098**	0.160	0.151
yxx: TEB	0.115	0.100	0.160	0.151
xyz: ML	**0.398**	**0.381**	0.437	**0.426**
xyz: TEB	0.412	0.404	**0.426**	0.433

FIGURE 23. Relationship of ancestral PP to accuracy. The posterior probability (PP) of inferred ancestral states is plotted against the probability that those states are correct. For both ML and TEB, I grouped all ancestral state inferences by their PP into 5%-sized bins. Within each bin, I calculated the proportion of inferred states that match the true state. Bins with fewer than 50 members were excluded. Data are shown for simulations on **(A)** ultrametric four-taxon, **(B)** non-ultrametric four-taxon, **(C)** ADH, **(D)** steroid hormone receptors, **(E)** GFP, and **(F)** EF-Tu phylogenies.

of sites; however, because the posterior probability of the second tree was so low, it contributed virtually zero weight to the TEB reconstruction. Support measures showed a similar trade-off: only when there was little or no uncertainty about the tree did the PP of an ancestral reconstruction differ among phylogenies. These results indicate that there is a trade-off between phylogenetic uncertainty and the extent to which ancestral state reconstruction depends on the phylogeny assumed.

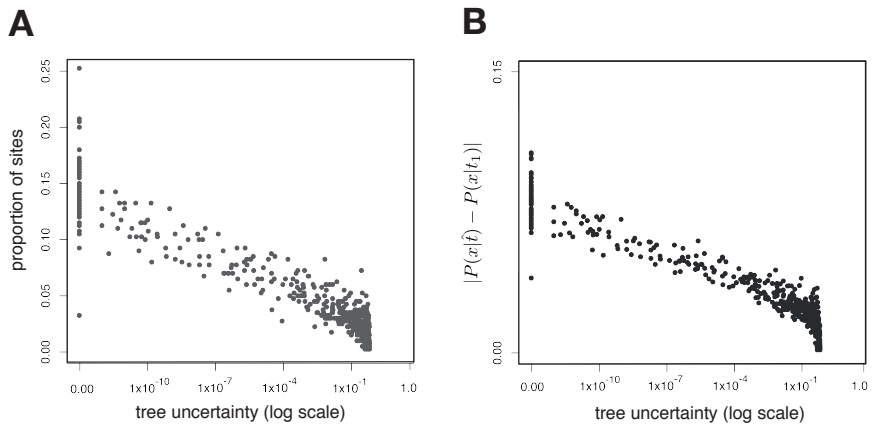


FIGURE 24. Phylogenetic uncertainty versus alternate ancestral reconstructions. Each point corresponds to one set of replicate descendants in the ultrametric four-taxon simulation. Tree uncertainty for each replicate is measured as 1.0 minus the posterior probability of the ML tree. **(A)** Tree uncertainty is plotted versus the proportion of sites at which the most likely ancestral state on the ML tree disagrees with the most likely ancestral state on the second-best tree. **(B)** Tree uncertainty is plotted versus the average absolute difference between the posterior probability of the most likely state on the ML tree minus the posterior probability of this same state on the second-best tree.

To understand this trade-off in detail, I examined ancestral reconstructions under two contrasting four-taxon conditions with different degrees of phylogenetic uncertainty (Fig. 25.). In one condition, the true phylogeny had a long internal branch, so the ML tree was inferred with no uncertainty (PP = 1.0); in the other,

the true phylogeny had a very short internal branch, so the ML tree was inferred with considerable uncertainty (PP = 0.384). For each state pattern, I reconstructed the ancestral state on all three possible topologies. I found that when there was no phylogenetic uncertainty, the probability of an ancestral state can differ radically given different trees; for three of the state patterns, the maximum a posteriori ancestral state inferred on the ML tree differed from that inferred on alternate trees. Because the internal branch was long, however, these alternate trees had zero posterior probability, so incorporating them into TEB reconstruction produces ancestral state inferences and posterior probabilities identical to the ML inference. In contrast, when the internal branch was short and the phylogeny was uncertain, all three topologies were close to being star trees. In this case, the probability of the ancestral state inferred on the ML tree was almost identical to the probability of that state given any other tree. Because the inferred ancestral state did not differ among phylogenies, TEB and ML again yielded the same reconstruction.

Discussion

My results demonstrate that a Bayesian approach to incorporating uncertainty about the underlying phylogeny is not necessary for ancestral state reconstruction. By comparing two methods of ancestral sequence reconstruction that differ only in that one assumes the ML phylogeny while the other integrates over phylogenies, I were able to determine the specific effect of incorporating phylogenetic uncertainty on ancestral state inferences, their statistical support, and their accuracy. I found that using TEB virtually never changes the inferred ancestral state; when it does, the reconstruction was already ambiguous using ML.

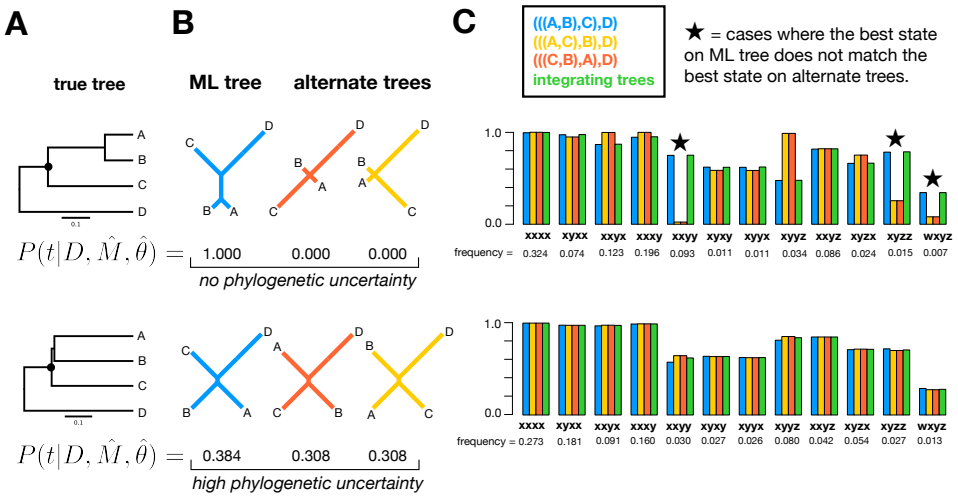


FIGURE 25. Inferred ancestral states are the same across uncertain trees. The conditions that produce phylogenetic uncertainty cause ancestral state inferences to be identical across trees. **(A)** I simulated sequences on trees with long (top) and short (bottom) internal branches. On each, I randomly generated an ancestral sequence 50,000 nucleotides long and simulated sequence evolution. **(B)** From the descendant sequences, I inferred the empirical Bayes posterior distribution of the three trees, each with its maximum likelihood branch lengths. **(C)** On each tree, I used the true model to reconstruct the common ancestor of descendants A, B, and C for all possible descendant state patterns (*xxxx*, *xyxx*, *wxyz*, etc.). Each bar corresponds to the posterior probability of the best ancestral state on the ML tree (blue), on the alternate trees (yellow and red), and integrated over all trees (green). Stars indicate state patterns for which the maximum a posteriori ancestral state on one of the alternate trees is different from that on the ML tree.

ML has slightly higher accuracy, and its posterior probabilities provide a slightly better predictor of the probability that an ancestral state inference is correct.

These analyses show that incorporating phylogenetic uncertainty only weakly affects ASR because the conditions that cause phylogenetic uncertainty also make the ancestral state the same across trees. This phenomenon occurs because when internal branches are short, the distance in tree-space is small between the ancestor on the ML tree and the ancestor on the second-best tree (Felsenstein (2004)). At the limit, the true tree is a star tree with a zero-length internal branch, and all resolved topologies have equal posterior probability, leading to maximal phylogenetic uncertainty; however, the ancestral nodes on the different topologies are identically located in tree space. In contrast, under the conditions that cause inferred ancestral states to differ among trees, there is typically no phylogenetic uncertainty to integrate over.

Prior work has shown that ASR is generally most accurate on star-like trees, because the descendant sequences contain maximum mutual information about the ancestral state when those descendants are completely independent phylogenetically (Blanchette et al. (2004); Lucena and Haussler (2005)). Those studies, however, assumed that the true phylogeny was known a priori, which is particularly unlikely for star-like trees with short internal branches. My work shows that phylogenetic uncertainty, which is inevitable under these conditions, is not expected to undermine the accuracy of ancestral state reconstruction on star-like trees. These results underscore the potential to accurately reconstruct ancestral sequences at the base of rapid phylogenetic radiations despite phylogenetic uncertainty, such as the ancestors of all mammals (Blanchette et al. (2004)) or all metazoans (Rokas et al. (2005)).

Previous work by Huelsenbeck and Bollback, like ours, showed a close relationship between ancestral posterior probabilities estimated using the ML tree and integrating over trees (Huelsenbeck and Bollback (2001)). Those authors did suggest, however, that uncertainty in the phylogeny might lead to significantly different interpretations of the ancestral state. This suggestion was illustrated using trees with arbitrarily assigned branch lengths and posterior probabilities; for all topologies in the illustration, the internal branch lengths were of significant length and the posterior probabilities were substantial. In reality, it is unlikely that any data set would support such a distribution of posterior probabilities over this set of tree/branch length combinations, because non-trivial posterior probabilities on "next-best" trees typically arise only when internal branches are short. My results show that when the posterior probabilities on trees are derived from sequence data rather than arbitrarily assigned, integrating over uncertainty has a negligible effect on ancestral sequence inference and a negative impact, if any, on accuracy.

My results should not be interpreted as an endorsement for sloppy analysis. Although incorporating phylogenetic uncertainty does not improve the accuracy of ancestral reconstruction, this does not mean the phylogeny is unimportant. Because ancestral reconstructions can vary across trees under some conditions, arbitrarily choosing an incorrect and implausible phylogeny could yield inaccurate reconstructions.

These findings should not be taken as evidence that ancestral reconstruction never errs. There are numerous potential sources of error that I did not evaluate, including use of incorrectly parameterized evolutionary models, which could yield incorrect (and strongly supported) inferences of phylogeny (Kolaczkowski and Thornton (2004)) or incorrect ancestral state reconstructions even when

the true tree is assumed. ASR practitioners should continue to use rigorous statistical practices, such as formal evaluation of a wide range of models that incorporate evolutionary heterogeneity (Posada (2001); Lartillot and Philippe (2004); Kolaczkowski and Thornton (2008)) and dense, targeted taxon sampling (Hillis (1998); Pollock et al. (2002); Heath et al. (2008)). My analyses were specific to Bayesian integration over uncertainty about the underlying phylogeny: I did not address the effect on ancestral reconstructions of integrating over uncertainty about branch lengths, the substitution model, or its parameters. Whether a Bayesian approach to these sources of uncertainty would improve or degrade ASR accuracy warrants further research.

In summary, incorporating phylogenetic uncertainty by integrating over topologies does not improve the accuracy of ancestral sequence reconstruction, because the conditions that cause phylogenetic uncertainty make the ancestral state the same across trees. Using the ML tree will typically yield the best ancestral reconstruction, even when the ML tree is uncertain. A Bayesian approach to phylogenetic uncertainty is intuitively appealing but computationally demanding and, in this case, unnecessary.

CHAPTER IV

CASE STUDY: EVOLUTION OF INCREASED COMPLEXITY

In this chapter, I provide a case study demonstrating the analysis pipeline previously described (Fig. 1.). I applied the methods in this pipeline to investigate the evolution of increased complexity in a molecular machine with multiple interacting parts. This project is groundbreaking because this is the first time the evolution of an entire molecular complex has been traced through history, and this is the first application of phylogenetic ancestral resurrection to all the members of a complex. Collaborators in the the Stevens Lab performed the in vitro experimentation. This work has been accepted for publication in the journal Nature.

Many cellular processes are carried out by molecular machines – assemblies of differentiated proteins that physically interact to execute biological functions (Pallen and Matzke (2006); Mulkidjanian et al. (2007); Forgac (2007); Dolezal et al. (2006); Clements et al. (2009); Archibald et al. (2000)). Despite much speculation, evidence is lacking concerning the mechanisms by which their complexity evolved. Comparative genomic approaches suggest that the the components of many molecular machines appeared sequentially during evolution, implying gradual increases of complexity by incorporating new parts into simpler machines (Pallen and Matzke (2006); Mulkidjanian et al. (2007); Dolezal et al. (2006); Clements et al. (2009); Gabaldón et al. (2005); Liu and Ochman (2007)). These horizontal analyses, however, are unable to decisively test these hypotheses or reveal the mechanisms by which additional parts became obligate components of existing systems. Here we perform the first vertical evolutionary analysis of a molecular machine by using ancestral gene resurrection (Thornton (2004); Frattini et al.

(2000)) and manipulative genetic experiments to reconstruct all components of a complex – the hexameric transmembrane ring of the vacuolar H+-ATPase (V-ATPase) proton pump – and identify the specific genetic and functional changes that caused an increase in complexity hundreds of millions of years ago. We show that the transmembrane ring of Fungi which is composed of three paralogous proteins evolved from a two-paralog ancestral complex because of a very small number of degenerative mutations, without the evolution of apparent new functions by the parts. After a gene duplication, both descendant proteins lost some of the specific inter-subunit interfaces required for their interactions with other ring proteins; these losses were complementary, so both copies became obligate components with restricted spatial roles in the complex. Reintroducing a single historical mutation from each paralog lineage into the resurrected ancestral proteins is sufficient to recapitulate this asymmetric degeneration and trigger the requirement for the more complex three-component ring. Our experiments show that increased complexity in an essential molecular machine evolved by simple, high-probability evolutionary processes and suggest a plausible mechanism for the evolution of complexity in other multi-paralog machines whose parts function in specific spatial orientations.

The V-ATPase proton pump is a multi-subunit protein complex that pumps hydrogen ions across membranes to acidify subcellular compartments; this function is required for intracellular protein trafficking, coupled transport of small molecules, and receptor-mediated endocytosis (Forgac (2007)). V-ATPase dysfunction has been implicated in human osteopetrosis, acquired drug resistance in human tumors, and pathogen virulence (Frattini et al. (2000); Pérez-Sayáns et al. (2009); Xu et al. (2010)). A key component of the V-ATPase is the V0 subcomplex, a hexameric

protein ring that utilizes a rotary mechanism to move protons across organelle membranes (Fig. 26.A) (Hirata et al. (2003); Imamura et al. (2005)). Although the V-ATPase is found throughout Eukaryotes, the V0 ring varies in subunit composition among lineages. In animals and most other eukaryotes, the ring consists of one subunit of Vma16 protein and five copies of its paralog, Vma3 (Fig. 26.B) (Forgac (2007)). In Fungi, the ring consists of one Vma16 subunit, four copies of Vma3, and one Vma11 subunit, arranged in a specific orientation relative to each other (Powell et al. (2000)). All three proteins are required for V-ATPase to function in Fungi (Umemoto et al. (1990, 1991)), but the mechanisms by which both Vma3 and Vma11 became obligate components with specific positions in the complex are unknown.

To address this issue, we reconstructed the ancestral ring proteins from periods just before and after the increase in complexity (Harms and Thornton (2010)), synthesized and functionally characterized them (Thornton (2004); Liberles (2007)), and used manipulative methods to identify the genetic and molecular mechanisms by which their functions changed (20). We first inferred the phylogeny and best-fit evolutionary model of the Vma3/11/16 protein family from the sequences of all 139 extant family members available in Genbank. The maximum likelihood phylogeny (Fig. 27.) indicates that Vma11 and Vma3 are sister proteins produced by a duplication of an ancestral gene (Anc.3-11) deep in the Fungal lineage, before the last common ancestor of all Fungi (800 million years ago) but after the divergence of Fungi from other eukaryotes (1 billion years ago) (Taylor and Berbee (2006)). The Vma11/Vma3 and Vma16 lineages, in turn, descend from an even older gene duplication deep in the Eukaryotic lineage (Fig. 26.B). We then used a maximum likelihood algorithm (Yang et al. (1995)) to infer the ancestral

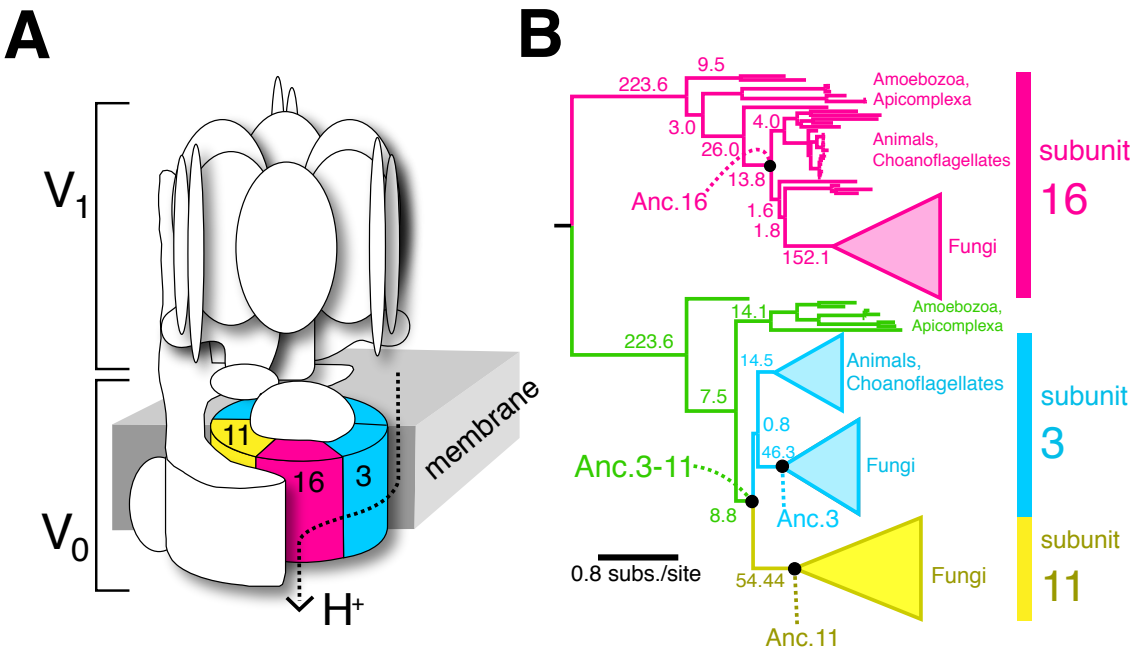


FIGURE 26. Structure and evolution of the V-ATPase complex. (A) In S. cerevisiae, the V-ATPase is assembled from fourteen subunits; the V1 subdomain is on the cytosolic side of the organelle membrane, and V0 is membrane-bound. (B) The maximum likelihood phylogeny of V-ATPase subunits Vma3, Vma11, and Vma16. Amoebozoa, Apicomplexa, Animals, and Choanoflagellates contain subunits 3 and 16, whereas Fungi contain 3, 11, and 16. Black circles show ancestral proteins reconstructed in this study. Colors correspond to those of the subunits in panel A; green lineages are the unduplicated orthologs of Vma3 and Vma11. Decimal values at internal branches express the approximate-likelihood ratio support for the monophyly of the descendant clade. A more detailed phylogeny is found in Fig. 27..

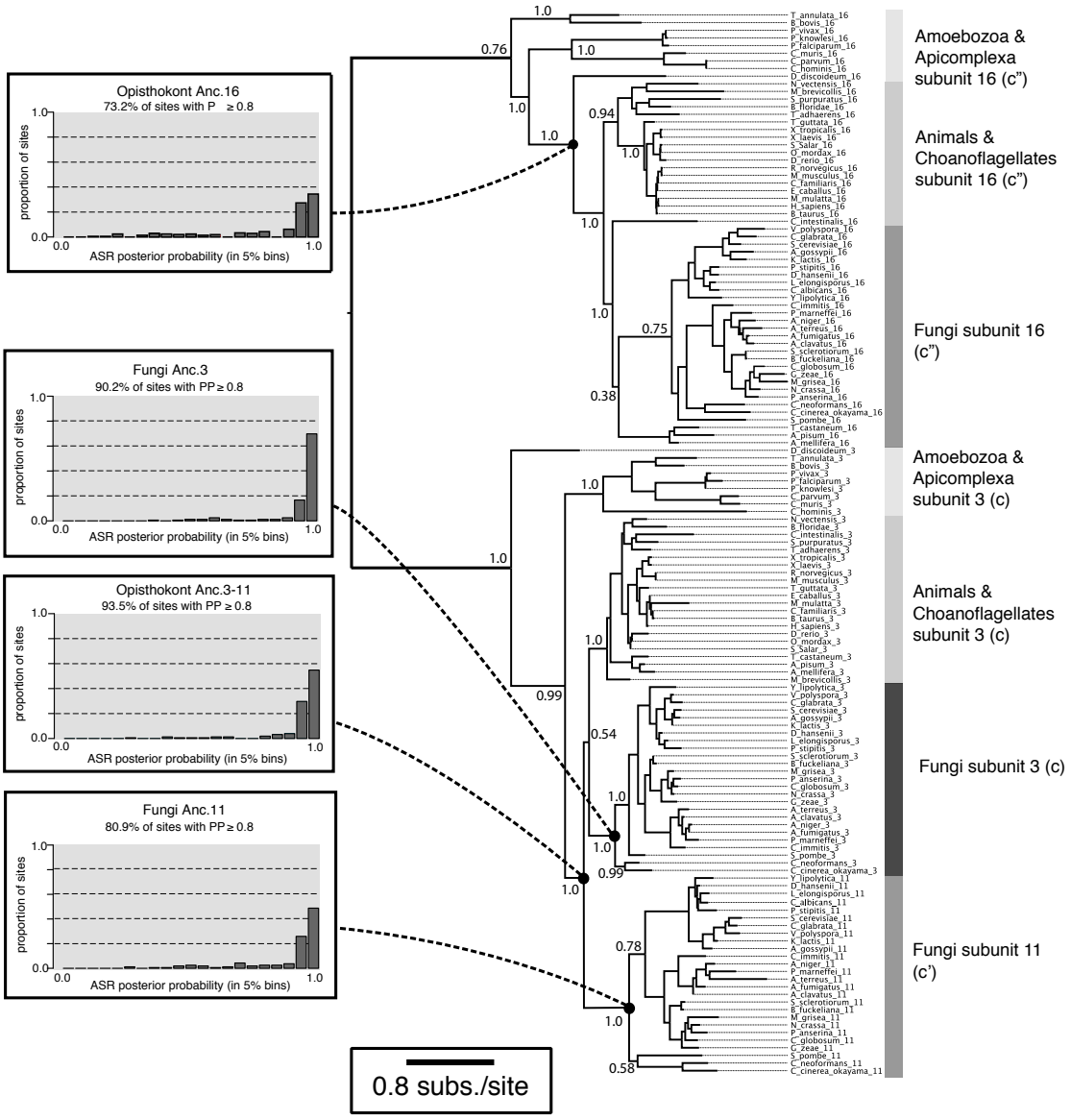


FIGURE 27. The maximum likelihood phylogeny of V-ATPase subunits Vma3, Vma11, and Vma16.

amino acid sequences with the highest probability of producing all the extant sequence data, given the best-fit phylogeny and model. We reconstructed the two paralogs that composed the ancient eukaryotic ring – Anc.3-11 (the last common ancestral protein from which the Vma3 and Vma11 lineages descend) and Anc.16 (the ancestral proteins from which all Vma16s descend); we also reconstructed the duplicated paralogs comprised by the three-member fungal ring in the ancestor of all Fungi – Anc.3 and Anc.11, the progenitors of the differentiated subunits Vma3 and Vma11, respectively, after Anc.3-11 was duplicated (Taylor and Berbee (2006)). We synthesized DNAs that code for these reconstructed proteins and assayed their functions in vivo using extant yeast S. cerevisiae.

Materials and Methods

In Silico Reconstruction of Ancestral Protein Sequences

V0 complex subunits Vma3, Vma11, and Vma16 are sometimes referred to as subunits c, c, and c in the specialty literature. We queried GenBank for all Eukaryote V-ATPase V0 ring sequences. Our query returned subunit 3, 11, and 16 protein sequences for twenty-six species in Fungi, and subunit 3 and 11 sequences for thirty-five species in Metazoa, Amoebozoa, and Apicomplexa. We aligned the sequences using PRANK v0.081202 (Loytynoja and Goldman (2005, 2008)). This alignment is best-fit by the Whelan-Goldman matrix (WAG) with gamma-distributed rate variation (+G) and proportion of invariant sites (+I), according to the Akaike Information Criterion as implemented in PROTTEST (Abascal et al. (2005)). Using WAG+G+I, we used PhyML v3.0 to infer the maximum likelihood (ML) topology, branch lengths, and model parameters (Guindon and Gascuel (2003)). We optimized the topology using the best result from Nearest-

Neighbor-Interchange and Subtree Pruning and Regrafting; we optimized all other free parameters using the default hill-climbing algorithm in PhyML. Phylogenetic support was calculated as the approximate likelihood ratio (36). Our ML analysis inferred the Nematoda 3 and 11 sequences to be connected by a very long branch basal to the Chromalveolata lineages; this result is inconsistent with our expectation that Nematoda are animals (Aguinaldo et al. (1997)) and we therefore excluded Nematoda data from further downstream analysis.

We reconstructed ML ancestral states at each site for all ancestral nodes in our ML phylogeny using our own set of Python scripts, called Lazarus, which wraps PAML version 4.1 (Yang (2007)). Lazarus parsimoniously places ancestral gap characters according to Fitchs algorithm (39). We characterized the overall support for Anc.3-11, Anc.16, Anc.3, and Anc.11 by binning the posterior probability of the ML state at each site into 5%-sized bins and then counting the proportion of total sites within each bin (Supplement S2).

Robustness to Alignment Uncertainty

In order to assess if our ancestral reconstructions are robust to alignment uncertainty, we aligned our protein sequences using four different alignment algorithms: CLUSTAL version 2.0.10 (Thompson et al. (1994)), MUSCLE v3.7 (Edgar (2004)), AMAP v2.2 (Do et al. (2005)), and PRANK v0.081202 (Loytynoja and Goldman (2005, 2008)). We then inferred the ML phylogeny and branch lengths for each alignment, using the methods described above. The resultant alignments varied in length from 347 sites (using CLUSTAL) to 683 sites (using PRANK), but all four alignments yielded the same ML topology with nearly identical ML branch lengths.

In order to determine which alignment algorithm yields the most accurate ancestral inferences under V-ATPase phylogenetic conditions, we simulated sequences across the V-ATPase ML phylogeny using insertion and deletion rates ranging from 0.0 to 0.1 indels per site. For each indel rate, we generated ten random unique indel-free ancestral sequences 400 amino acids in length and then used INdelible (Fletcher and Yang (2009)) to simulate the ancestral sequence evolving along the branches of our ML phylogeny under the conditions of WAG+I+G model with indel events randomly injected according to the specified indel rate. The size of each indel event was drawn from a Zipfian distribution with coefficient equal to 1.1 and the maximum length limited to 10 amino acids. We aligned each replicates descendant sequences using AMAP, CLUSTAL, MUSCLEs, and PRANK; for each alignment, we inferred the ML topology, branch lengths, and model parameters using the methods described above. We used Lazarus to reconstruct all ancestral states, and queried Lazarus for the most-recent shared ancestor for Opisthokont subunit 3/11 and Opisthokont subunit 16 sequences. We measured the error of ancestral reconstructions as the proportion of ancestral sites that incorrectly contained an indel character (Fig. **??**).

Plasmids and Yeast Strains

Bacterial and yeast manipulations were performed using standard laboratory protocols for molecular biology (Sambrook and Russel (2001)). Plasmids used can be found in Supplement S5. Ancestral sequences (pGF140, pGF139, pGF506, and pGF508) were synthesized by GenScript (Piscataway, NJ) with a yeast codon bias. Triple hemagglutinin (HA) epitope tags were included prior to each stop codon. The Anc.3-11, Anc.16, Anc.3, and Anc.11 genes were subcloned to single-copy,

CEN-based yeast vectors. The ADH terminator sequence (247 base pairs) and NatR drug resistance marker (Goldstein and McCusker (1999)) were polymerase chain reaction (PCR) amplified with 40 bp tails homologous to the 3' end of each coding region and vector sequence. Vectors were gapped, co-transformed into SF838-1D yeast with PCR fragments, and cells were selected for NatR. A second round of in vivo ligation was used to place the ancestral genes under 500 bps of the VMA3 or VMA16 promoters to create pGF140 and pGF139, respectively. The following vectors all used a similar cloning strategy: pGF240 - pGF41, pGF252, pGF253, pGF503 - pGF508, pGF510, pGF512 - pGF515, pGF517 - pGF519, pGF521, pGF523, pGF528, pGF529, pGF531, pGF534 - pGF537, and pGF542. Briefly, the relevant locus (Anc.3-11, Anc.16, or Anc.3) was PCR amplified with 5' and 3' untranslated flanking sequence and cloned into pCR4Blunt-TOPO (Invitrogen, Carlsbad, CA). If necessary, a modified Quikchange protocol (Zheng et al. (2004)) was used to introduce point mutations before the gene was subcloned into a yeast vector (pRS316 or pRS415). To generate pGF502, codon 31 through the stop codon of Anc.16 were amplified with the ADH::NatR cassette from pGF139, cloned into TOPO, and in vivo ligated downstream of the VMA16 promoter (including a start codon) in pRS415.

A triple-fragment in vivo ligation was used to generate pGF646 - pGF651. Gapped vector containing the VMA16 promoter was transformed into yeast with two PCR fragments of the ring genes to be fused. For pGF646, the coding region of (i) VMA16 (without codons 2-41) and (ii) the coding region of Anc.11 (without codons 2-5) were amplified by PCR. The proteolipid on the C-terminal portion of the gene fusion also contained the ADH terminator and NatR cassette; the amplified products contained PCR tails with homology to link the genes to

both the gapped vector and to each other. Gene fusions were modeled after the experimental design of Wang et al. (2007) where the lumenal protein sequence linking the two proteolipids was designed to be exactly 14 amino acids. To meet these criteria, additional amino acids were inserted into the following vectors linking the two subunits: pGF646 (TRVD), pGF648, pGF650 (TR), pGF649, pGF651 (GS).

Yeast strains used can be found in Supplement S5. Strains containing deletion cassettes other than KanR (Goldstein and McCusker (1999)) were constructed by PCR amplifying the HygR or NatR cassette from pAG32 or pAG25, respectively, with primer tails with homology to flanking sequences to the VMA11 or VMA16 loci. 11::KanR and 16::KanR strains (SF838-1D?) were transformed with the HygR and NatR PCR fragments, respectively, and selected for drug resistance. The 11::HygR locus was amplified and transformed into LGY113 (to create LGY125) and LGY115 (to create LGY124). This was repeated with the 16::NatR locus to create LGY139 and LGY143.

Yeast Growth Assays

Yeast were grown in liquid culture, diluted five-fold, and spotted onto YEPD media buffered to pH 5.0 or YEPD media containing 25 mM (Figs. 2, 3, 4) or 30 mM CaCl2 (Fig. 28.F).

Whole Cell Extract Preparation and Immunoblotting

Yeast extracts and Western blotting were performed as previously described (47). Antibodies used in this study included monoclonal primary anti-HA (Sigma-

Aldrich), anti-Dpm1 (5C5; Invitrogen), and secondary horseradish-conjugated anti-mouse antibody (Jackson ImmunoResearch Laboratory, West Grove, PA).

Fluorescence Microscopy

Staining with quinacrine was performed as previously described (Ryan et al. (2008)). The cell wall (shown in red) was visualized using concanavalin A tetramethylrhodamine (Invitrogen). Microscopy images were obtained using an Axioplan 2 fluorescence microscope (Carl Zeiss, Thornwood, NY). A 100x objective, AxioVision software (Carl Zeiss), and Adobe Photoshop CS (v. 8.0) were used.

Results

To functionally characterize the resurrected proteins, we transformed them into S. cerevisiae deficient for various ring components and therefore incapable of growth in the presence of elevated CaCl2 (Kane (2006)). We found that the ancestral two-component ring can functionally replace the three-component ring of extant yeast; this result indicates that neither the complex nor its parts evolved new functions required for growth under the conditions in which the ring is known to be important. When the resurrected Anc.3-11 was transformed into yeast deficient for Vma3 (3) or Vma11(11), growth in the presence of elevated CaCl2 was rescued, indicating that all the functions of the present-day Vma3 and Vma11 proteins were already present before their birth by gene duplication (Fig. 28.A). Further, Anc.3-11 unlike either of its present-day descendants can partially rescue growth in yeast that are doubly deficient for both Vma3 and Vma11 (311). In addition, the reconstructed Anc.16 rescued growth in Vma16-deficient S. cerevisiae (16) (Fig. 28.B), and co-expression of Anc.3-11 and Anc.16 together rescued

cell growth in 31116 yeast, which lack all three ring subunits (Fig. 28.C). The ancestral genes specifically restore proper V-ATPase function in acidification of the vacuole lumen (Fig. 28.G). Further, mutation of the ancestral subunits to remove glutamic acids residues known to be essential for V-ATPase enzyme function (16,24) abolished their ability to rescue growth on CaCl2 (Supplement S7). These inferences about the functions of Anc.3-11 and Anc.16 are robust to uncertainty about ancestral amino acid states. We reconstructed alternate versions of Anc.3-11 and Anc.16 by introducing amino acid states with posterior probability $\geq$0.2, but none of these abolished the ability of the ancestral genes to functionally substitute for the extant subunits (Supplement S8).

Similar experiments with the components of the ancestral three-component ring show that after Anc.3-11 duplicated, both Vma3 and Vma11 became necessary for a functional complex, because each lost specific ancestral functions that were maintained in the other. Unlike Anc.3-11, expression of Anc.3 can rescue growth and vacuole acidification in 3 but not 11Δ yeast, and Anc.11 can rescue growth in 11Δ but not 3Δ yeast (Fig. 28.D,E,G). Further, both Anc.3 and Anc.11 are required to fully rescue growth in 311 yeast (Fig. 28.F). These data indicate that after their birth by gene duplication, Anc.11 lost the ancestral proteins ability to carry out at least some functions of Vma3, and Anc.3 lost the ancestral capacity to carry out those of Vma11.

We conjectured that the evolution of the specialized roles of Vma3 and Vma11 reflected the loss of specific interaction interfaces required for ring assembly that were present in the ancestral protein. Previous experiments with fusions of extant yeast proteins have shown that the arrangement of subunits in the ring is constrained by the capacity of each subunit to partcipate in specific interfaces

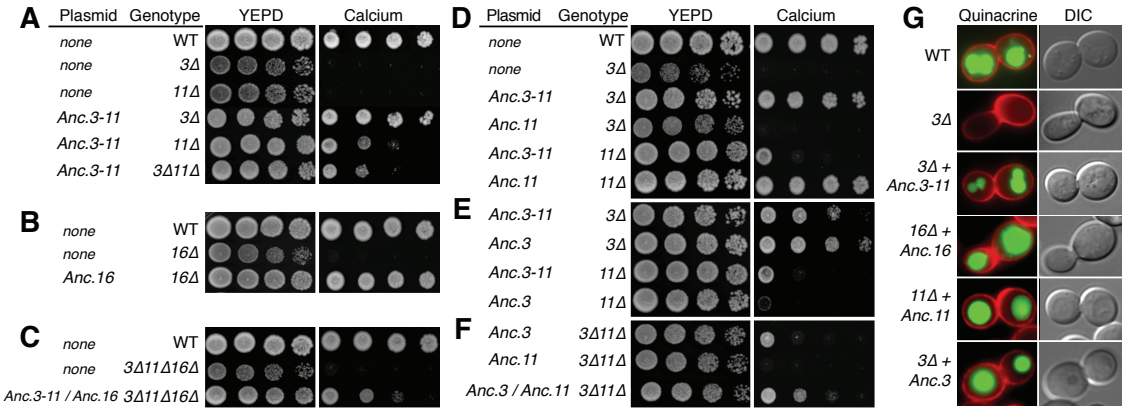


FIGURE 28. Reconstructed V-ATPase ancestors replace extant versions. The ancestral V0 subunits functionally replace the three-component ring in extant yeast. S. cerevisiae were plated on permissive media (YEPD) and media buffered with elevated CaCl2. (A) Anc.3-11 rescues growth in yeast deficient for endogenous subunit 3 (3?), subunit 11 (11?), or both (3?11?). Wild-type (WT) yeast growth is shown for comparison. (B) Anc.16 rescues growth in yeast deficient for subunit 16 (16?). (C) Expression of Anc.3-11 and Anc.16 together rescues growth in yeast deficient for subunits 3, 11, and 16. (D) Anc.11 rescues growth in 11? but not 3? yeast. (E) Anc.3 rescues growth in 3? but not 11? yeast. (F) Anc.3 and Anc.11 together rescue growth in 3 11 mutants. (G) Yeast expressing Anc.3-11, Anc.16, Anc.11, or Anc.3 properly acidified the vacuole lumen. Red signal labels yeast cell walls; green signal (quinacrine) labels acidified compartments. Yeast were visualized by differential interference contrast microscopy (DIC).

(which we labeled P, R, and Q) with the other subunits (Wang et al. (2007)). Specifically, Vma11 is restricted to a single position between Vma16 and Vma3, because its clockwise interface can participate only in interface R with Vma16, and its counterclockswise interface can participate only in interface P with clockwise side of Vma3 (Fig. 29.). Copies of Vma3, in contrast, occupy several positions in the ring, because they form interface P with other copies of Vma3 or Vma11, as well as interface Q with Vma16; Vma3 cannot, however, form interface R with Vma16. As a result, both Vma3 and Vma11 are required in extant yeast in order to form a complete ring with Vma16.

To test the hypothesis that specific interaction interfaces were lost during evolution, we engineered fusions using ancestral ring proteins to assay the capacity of each ancestral ring protein to form specific interfaces with the other subunits required for a functional complex. Because Anc.3-11 can complement the loss of both subunits 3 and 11, we hypothesized that Anc.3-11 subunit could participate in all three specific interaction interfaces, and that these capcities were then partitioned between Anc3 and Anc11 after the duplication of Anc.3-11 (Fig. 29.A,B). To test this hypothesis, we created six reciprocal gene fusions between yeast subunit 16 and subunits Anc.3-11, Anc.3, Anc.11 (Fig. 29.C). Each fusion constrains the structural position of subunits relative to extant subunit 16, allowing us to determine which arrangements yield a functional ring. As predicted, Anc.3-11 functioned on either side of subunit 16 (Fig. 29.D), indicating that it could form all three interfaces P, Q, and R. In contrast, Anc.3 functioned when constrained to participate in interface Q with Vma16 and interface P with subunit 3; however, ring function was lost when Anc.3 was constrained to form interface R with Vma16 (Fig. 29.E). Anc.11, in turn, functioned when constrained to participate in interface

R with Vma16 and interface P with Vma3, but ring function was lost when Anc.11 was constrained to participate in interface Q with Vma16 and interface P with Vma3. This result indicates that Anc.11 lost the capacity to form one or both of these interfaces during its post-duplication divergence from Anc.3-11 (Fig. 29.F).

Taken together, these data indicate that the specificity of the ring arrangement and the obligate roles of Vma3 and Vma11 evolved by complementary loss of asymmetric interactions with other members of the ring (Fig. 29.G,H). Before Anc.3-11 duplicated, the protein ring only contained an undifferentiated subunit 3/11 and a subunit 16. Immediately after Anc.3-11 duplicated, the two descendant subunits must have been functionally identical, so the protein ring could have assembled with many possible combinations of the two descendants, including copies of only one of the descendant proteins. This flexibility disappeared when Anc.3 lost the ancestral interface that allowed it to interact with the counterclockwise side of Vma16, and Anc.11 lost the ability to interact with Vma16s clockwise side and/or Vma3s counterclockwise side. These complementary losses are sufficient to explain the specific arrangement of contemporary subunits in reconstructed and present-day fungal rings.

To determine the genetic basis for the partitioning of Anc.3-11s functions between Vma3 and Vma11, we introduced historical mutations into Anc3.11 by directed mutagenesis and determined whether they recapitulated the shifts in function that occurred during the evolution of Anc.3 and Anc.11. The two phylogenetic branches leading from Anc.3-11 to Anc.3 and to Anc.11 contain 25 and 31 amino acid substitutions, respectively, but only a subset of these were strongly conserved in subunits 3 or 11 from extant Fungi (Fig. 30.A). We introduced each of these diagnostic substitutions into Anc.3-11 and determined

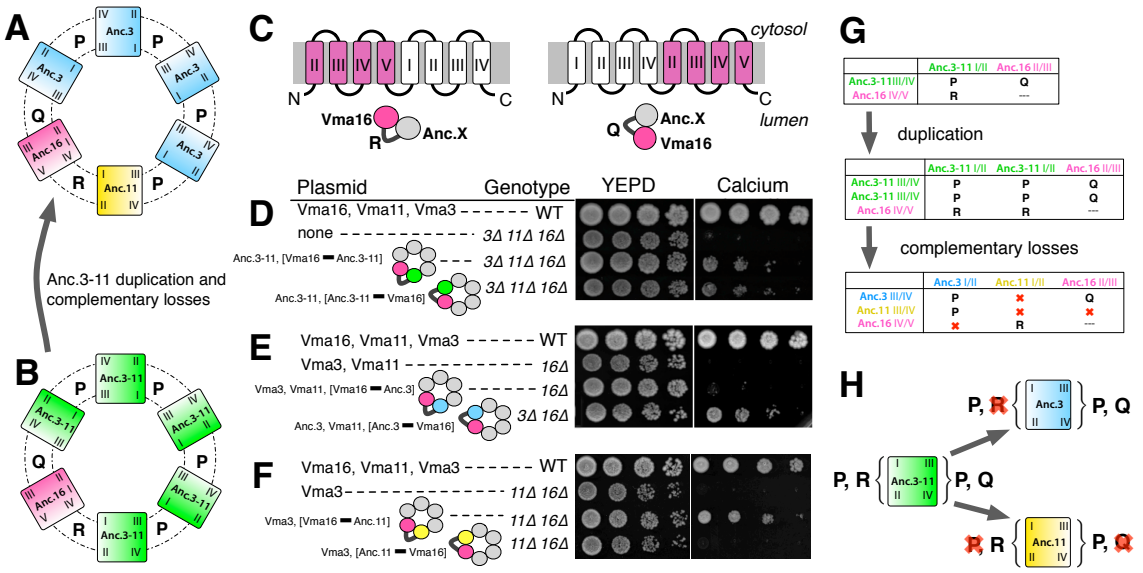


FIGURE 29. Increasing complexity by complementary loss of interactions in the fungal V0 ring. (A) Model of the protein ring composed of Anc.3, Anc.11, and Anc.16, arranged as in extant yeast (Wang et al. 2007). Unique intersubunit interfaces are labeled P, Q, and R. (B) Before duplication of Anc.3-11, the ring was assembled in similar fashion from two subunits. (C) To constrain the location of specific subunits, gene fusions were constructed by tethering an ancestral subunit to either the N or C-terminal side of yeast Vma16. Roman numerals indicate location of transmembrane helices (I through V) based on previous work (30). (D, E, F) Growth assays of yeast with fused V0 subunits to identify the interfaces in which ancestral subunits can participate. For each experiment, expressed V0 subunits are listed; tethered subunits are in brackets and connected by a thick line. Cartoons show the constrained location of the tethered subunit relative to Vma16. (D) Anc.3-11 can function on either side of Vma16. (E) Anc.3 can function only on the clockwise side of Vma16. (F) Anc.11 can function only on the counterclockwise side of Sc.16. (G) Interfaces formed by V0 subunits before and after duplication and complementary loss of interfaces, based on the data in panels D-F. Red Xs indicate lost interfaces. (H) Anc.3-11 participated in interfaces P and R on its clockwise side and interfaces P and Q on its counterclockwise side. Participation in R was lost along the lineage leading to Anc.3; participation in P and/or Q was lost along the lineage leading to Anc.11.

whether they recapitulated the loss by Anc.3 or Anc.11 of the capacity to complement Vma gene deletions. We found that a single amino acid replacement that occurred on the branch leading to Anc.11 (V15F) abolished the capacity of Anc.3-11 to function as subunit 3; it also enhanced the ability of Anc.3-11 as subunit 11 (Fig. 30.B). Moreover, a single historical replacement (M22I) from the branch leading to Anc.3 radically reduced the ancestral capacity to function as subunit 11 (Fig. 30.C); the Anc.3-11-M22I mutant retains some of the ancestral proteins capacity to rescue growth in the Vma11-deficient background, suggesting that other mutations also contributed to the functional evolution of Vma3. One other historical mutation (N88T) on this branch also impaired Anc.3-11s capacity to function as Vma11, but it also reduced the proteins capacity to function as Vma3, suggesting that epistatic interactions with other residues allow this mutation to be tolerated in Anc.3 and its descendants. Several of the replacements on the branch leading to Anc.11 display a similar pattern, reducing the proteins capacity to replace Vma3, suggesting that these historical replacements function better together than in isolation.

Discussion

How complexity and specific gene functions can evolve has long puzzled evolutionary biologists (Ohno (1970); Jacob (1977); Lynch (2007b)), because mutations that compromise function are far more frequent than those which generate functional novelty (Hietpas et al. (2011)). Our results indicate that increases in the architectural complexity of molecular assemblies can evolve due to a small number of simple, relatively high-probability mutations that degrade ancestral functions but leave other functions intact. The specific roles of subunits

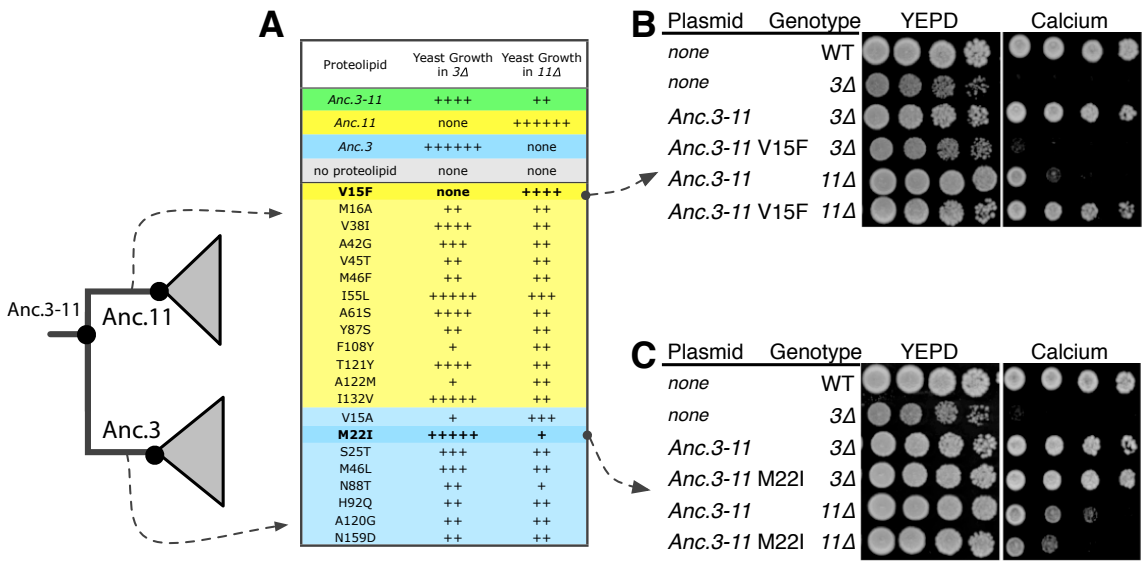


FIGURE 30. Genetic basis for functional differentiation of Anc.3 and Anc.11. (A) Experimental analysis of historical mutations. Strongly conserved historical amino acid replacements from the branches on which Anc.3-11s functions were partitioned are listed in the table. Yellow, replacements on the branch leading to Anc. 11; blue, replacements on the branch leading to Anc.3. Each mutation was introduced singly into Anc.3-11; the variant genes were transformed into S. cerevisiae, and growth was assayed on elevated CaCl2. The table shows growth in semiquantitative terms from zero (none) to wild-type (++++++). Bold mutations recapitulate in whole or part the functional evolution of Anc.11 and Anc. 3. (B) Replacement V15F abolishes Anc.3-11s capacity to function as subunit 3 and enhances its capacity to function as subunit 11. (C) Replacement M22I impairs the capacity of Anc.3-11 to function as subunit 11 without affecting its capacity to function as subunit 3.

Vma3 and Vma11 appear to have been acquired when duplicated genes lost some but not all of the ancestral proteins capacity to participate in interactions with copies of itself and another protein required for proper ring assembly. Because complementary losses occurred in both lineages, the two descendant subunits became obligate components, and the complexity of the ring increased. It is possible that specialization of the duplicated subunits allowed increases in fitness, but genome-wide interaction screens and the phenotype of vma11Δ yeast provide no evidence that Vma11 evolved novel functions in addition to those it inherited from Anc.3-11 in the V0 ring (Tong et al. (2004)).

Our results indicate that increases in architectural complexity can evolve due to a small number of simple, relatively high-probability mutations that degrade ancestral functions but leave other functions intact. The specific roles of subunits Vma3 and Vma11 appear to have been acquired when duplicated genes lost some but not all of the ancestral proteins capacity to participate in interactions with copies of itself and another protein required for proper ring assembly. Because complementary losses occurred in both lineages, the two descendant subunits became obligate components, and the architectural complexity of the ring increased.

Because ours is the first mechanistic analysis of the evolutionary trajectory of a molecular machine, the generality of our observations is unknown. By definition, however, all molecular machines involve differentiated parts in specific spatial orientations, and many are composed in whole or part of paralogous proteins. In any such complex, additional paralogs could become obligate components due to gene duplication (Pereira-Leal et al. (2007)) and subsequent mutations that cause specific interaction interfaces among them to degenerate. For example, –

the V1 subcomplex, which represents the rest of the V-ATPase proton-translocating complex (Mulkidjanian et al. (2007)), as well as chaperonin complexes (Archibald et al. (2000)), the NADH:ubiquinone oxidoreductase (Gabaldón et al. (2005)), the mitochondrial import machinery (Dolezal et al. (2006)), and the bacterial flagellums rod, hook, and filament (Pallen and Matzke (2006)) are all composed of paralogous subunits that can function only in specific spatial orientations; the mechanisms we observed that account for the increased complexity of the V0 ring could plausibly be involved in the evolution of these machines as well.

This view of the evolution of molecular machines is related to recent models that explain other biological phenomena such as the retention of large numbers of duplicate genes and mobile genetic elements within genomes as the product of degenerative processes acting upon biological systems with some degree of modularity (Lynch (2007a,b); Force et al. (1999)). Although mutations that enhanced the functions of individual ring components may also have occurred during evolution, our data indicate that simple degenerative mutations are sufficient to explain the historical increase in complexity of a crucial molecular machine. There is no need to invoke the acquisition of novel functions caused by low-probability mutational combinations.

CHAPTER V

CONCLUSION

In this dissertation I described a computational analysis pipeline for studying the evolutionary history of protein families. This pipeline begins with protein sequences that are evolutionarily related, and then proceeds to align the sequences, infer a phylogeny from the alignment, and then reconstruct ancestral sequences on the phylogeny. Reconstructed ancestral sequences can be physically synthesized and expressed *in vivo* in order to observe their ancient functions.

In chapter II, I showed that ML phylogenetic error can be partially ameliorated by using a multidimensional search heuristic. Virtually all implementations of ML phylogenetic inference use a simple heuristic that assumes parameters are separable. I implemented a multidimensional heuristic that does not assume parameter separability, and thus simultaneously optimizes all parameters. I observed that this heuristic found more accurate and higher-likelihood phylogenies more often than the simpler heuristic.

In chapter III, I showed that statistical uncertainty about ML phylogenies does not significantly impact the downstream accuracy of ancestral reconstruction. The conditions that cause phylogenetic uncertainty also create a situation in which ancestral sequences are the same across alternate phylogenies. Phylogenetic uncertainty is correlated with short tree branches, which eliminate opportunity for ancestral variance. This result is important because it allows experimentalists to avoid a very time-consuming computation that may require weeks or months to complete.

Finally, in chapter IV, I combined the pieces of this pipeline – in collaboration with molecular biologists in the Stevens Lab – to investigate the evolution of a molecular machine. Our work not only demonstrates the computational techniques I advocated in previous chapters, but also provides a novel biological result: molecular machines can evolve increased complexity through degenerative mechanisms.

Error versus Uncertainty

My results in chapter II may seem incongruous with my results in chapter III. I want to dissuade readers from the following specious line of thinking: *Phylogenetic uncertainty does not significantly affect the accuracy of ancestral sequence reconstruction, so who cares about the accuracy of my ML phylogeny?*

The key to this puzzle is that error and uncertainty are not the same thing. Error is inaccuracy, whereas uncertainty is ambiguity. In an ideal world, our metrics for phylogenetic uncertainty would be perfect predictors of phylogenetic accuracy – but this is not the case. This means that a correct inference can sometimes be very uncertain, and an incorrect inference can be strongly supported. Our metrics of phylogenetic uncertainty are imperfect because we do not have global complete knowledge of the phylogenetic likelihood landscape. Just as our optimization algorithms must be heuristic, our metrics of uncertainty must also be heuristic. Phylogenetic uncertainty typically arises when the ML phylogeny exists in a region of tree space with one or more strongly-supported nearby alternate trees. This region of space can be imagined as a broad hill, with the ML tree at the summit. In contrast, strongly-supported (i.e. certain) ML trees exists in a region of tree space that can be imagined as a sharp peak with no significant support

for nearby trees. Phylogenetic error occurs when the ML tree exists on the wrong peak. An errored tree may be strongly supported (on a sharp peak) or ambiguously supported (on a broad hill) – but either way, it's the wrong tree. My results in chapter II showed that Multimax does a better job finding the correct hill. My results in chapter III showed that trees on the same hill yield the same ancestral sequence.

Implications for the Future of Systems Biology

The analysis pipeline discussed in this dissertation was presented in the context of studying single gene families in isolation. However, nearly all biological phenotypes of scientific interest are produced by multiple genes working in concert. A research frontier of biology is to study the evolution of complex phenotypes as a consequence of multi-gene systems responding to particular environmental cues. Future progress in this direction will be made by extending the methods of phylogenetic ancestral reconstruction from single-gene studies to multi-gene studies. My results in chapter IV demonstrate this multi-gene paradigm for a simple system with three genes. I look forward to a not-so-distant future in which we can generate and test hypotheses about the evolution of complex ancestral systems, including ancestral regulatory networks, ancestral signal transduction networks, and the ancestral assembly of complex molecular machines.

Computational error and statistical uncertainty will play an important role in the future of systems-level evolutionary studies. When multiple genes are studied in tandem, error and uncertainty can become amplified. For example, in any reconstructed ancestral protein sequence, it is not uncommon for some proportion of sites to be ambiguously inferred (with low posterior probability support). A

rigorous experimentalist will explicitly test alternate molecular states at these ambiguous sites in order to observe their effect on reconstructed protein function. The total number of uncertain ancestral sites increases when reconstructing multiple ancestral proteins because there are simply more sequences being studied. This means that reconstructing a complex multi-gene ancestral system increases the degrees of uncertainty that must be systematically explored. In extreme cases, the amount of labor required to test all ambiguous states could become overwhelming. Therefore, as the field of systems biology adopts the methods of phylogenetic ancestral reconstruction, it becomes increasingly important to minimize ancestral uncertainty and maximize ancestral accuracy.

Software

There are currently few – if any – available software tools that chain together all the algorithms necessary for multiple sequence alignment, phylogenetic inference, and ancestral reconstruction into a unified toolkit. My software tools (URL = XX) provide a first attempt at automating this pipeline. My tools automatically make "smart" decisions during each pipeline stage, including time-consuming tasks like ML model selection. I encourage you to use my software suite in your future ancestral reconstruction projects.

APPENDIX A

MARKOV MODELS OF SEQUENCE EVOLUTION

The core idea of molecular Markov models is that characters substitute to other characters over time with some probability. The frequency of substitution events is assumed to have a Poisson distribution, and the probability of k events occurring in time t is:

$$P(k|t) = \frac{t^k e^{-t\mu}}{k!} \tag{A.1}$$

where μ is our assumed rate of evolution. The probability that any number of substitutions—from zero to infinity—occur in time t can be calculated by summing over all values k:

$$P(0 \leq k \leq \infty|t) = \sum_{k=0}^{\infty} \frac{\mu t^k e^{-t\mu}}{k!} = 1.0 \tag{A.2}$$

Expression A.2 is used to calculate the likelihood of a single phylogenetic branch for a single sequence site as follows.

Suppose we observe an evolutionary character—a single nucleotide or an amino acid—currently in some state x, where x is one of the letters in the nucleotide or amino acid alphabet. Also suppose we have a matrix R expressing the relative substitution rates between states. R is an n-by-n matrix, where n is the size of the alphabet. Finally, we have a vector π expressing the expected frequencies of each state. Putting all these elements together, x will mutate to state y over time t with probability calculated as follows:

$$P(x \rightarrow y|t) = \sum_{k=0}^{\infty} (\pi_x \pi_y R_{xy}^k) \frac{t^k e^t}{k!} \tag{A.3}$$

. . . where R_{xy} is the relative rate of x transitioning to y, and (R_{xy}^k) the extrapolated rate of $x \rightarrow y$ occurring over k steps. π_x and π_y are the frequencies of states x and y, otherwise known as the stationary frequencies. Expression A.3 is typically shown in a more compact form:

$$P(t) = \sum_{k=0}^{\infty} \frac{Q\mu t^k}{k!} = e^{Q\mu t} \tag{A.4}$$

. . . where the matrix Q equals $\Pi R - I$. Π is the diagonal matrix, where $\Pi[a, a]$ equals the equilibrium frequency π_a for state a in our alphabet. I is the identity matrix. The value μ is chosen such that the total rate of possible mutation is one: $1 = \mu \times \left(1 - \sum_a \pi_a R_{aa}\right)$. Whereas Expression A.3 calculates a single floating-point probability value, Expression A.4 calculates a matrix P of probability values for any state x mutating to any other state y. A description of the matrix algebra necessary to convert Expression A.3 into Expression A.4 is given in Bryant et al. (2005).

APPENDIX B

COMPUTING THE LIKELIHOOD OF A PHYLOGENY

The likelihood of an entire phylogeny is calculated by recursively applying Expression A.4 to all branches in the tree. The likelihood $L(t, \theta|D)$ of the tree t and model parameters θ, given sequence alignment D, is calculated as a product of likelihoods $\left(\prod_i L(t, \theta|D_i)\right)$ for each sequence site i (Expression B.1).

$$L(t, \theta|D) = \prod_i L(t, \theta|D_i) \tag{B.1}$$

The likelihood $L(t, \theta|D_i)$ of the phylogeny at site i is the sum of partial likelihoods $\sum_x L_x^v$ of the root node (call it node v) having state x at site i (Expression B.2).

$$L(t, \theta|D_i) = \sum_x L_x^v \tag{B.2}$$

Each partial likelihood L_x^v is calculated recursively, by descending from v along its branches t_1 and t_2 to nodes u_1 and u_2. Along each branch, we calculate the sum of probabilities of x mutating to some state y. (Expression B.3).

$$L_x^v = \left(\sum_y P(x \to y|t_1) L_y^{u_1}\right)\left(\sum_y P(x \to y|t_2) L_y^{u_2}\right) \tag{B.3}$$

$L_y^{u_1}$ and $L_y^{u_2}$ are the partial likelihoods of observing state y at nodes u_1 and u_2. These partial likelihoods are calculated by deeper recursion to the branches descending from nodes u_1 and u_2. Eventually, the recurrence arrives at a leaf node u_T. The partial likelihood $L_y^{u_T}$ of state y at node u_T equals 1.0 if u_T is state y in the sequence data; otherwise $L_y^{u_T}$ equals 0. Figure 31. illustrates the data structures involved in this recursion.

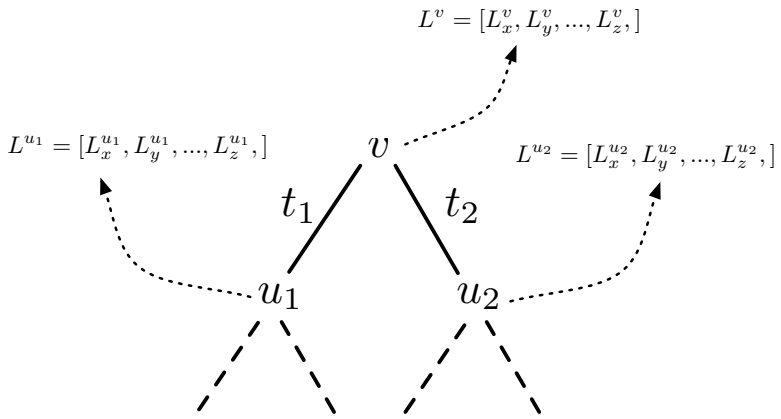


FIGURE 31. The recursive data structure for the likelihood algorithm. We pick an arbitrary root node v, with descendant branches t_1 and t_2 leading to nodes u_1 and u_2. We recursively calculate a vector of partial likelihoods for each node on the tree. For example, the vector L^{u_1} contains the partial likelihood $L^{u_1}_x$ of node u_1 existing as state x, the partial likelihood $L^{u_1}_y$ of u_1 existing as state y, etc.

Markov Models Make Simplifying Assumptions About Evolution

Markov models of molecular sequence evolution make three major several simplifying assumptions about the underlying evolutionary process. First, sites within an alignment are assumed to evolve independently from each. Second, the state-to-state substitution process is simplified to be time reversible. Finally, the substitution process is assumed to be ergodic, such that the expected frequency of each state is assumed to be static over evolutionary time. All three of these assumptions have been shown to be incorrect for specific empirical counterexamples, but for most protein families these assumptions seem to yield robust and accurate evolutionary inferences.

APPENDIX C

ALIGNMENT ERROR & ANCESTRAL RECONSTRUCTION

In this section I show there is a complex relationship between error in multiple sequence alignments and the downstream error in ancestral reconstruction. I observed that a collection of five alignment algorithms found significantly different alignments for simulated sequences evolved in a variety of controlled conditions. Further, these five alignments yielded downstream ancestral sequences that varied in length and accuracy. My results show that the choice of alignment algorithm has significant consequences for the accuracy of downstream evolutionary inference. Further work is required to dissect the mechanisms of alignment algorithms in order to understand why their accuracy varies in different evolutionary conditions.

Phylogenetic inference and all downstream analysis relies on the accuracy of the multiple sequence alignment (MSA). The goal of MSA is to identify the characters that are homologous – with shared evolutionary history – within a collection of sequences. MSA can be difficult because molecular sequences tend to acquire lineage-specific mutations over evolutionary time; the precise relationships between sequences may be unclear. There exist dozens of algorithms and software packages for MSA (Batzoglou (2005); Notredame (2007)). MSA algorithms primarily differ in two ways: their cost functions, and their strategies for hierarchically ordering the sequences.

Cost functions determine the relative penalties for inserting gap characters. These functions are based on a dynamic string matching algorithms, using the Needleman-Wunsch algorithm for global alignment (Needleman and Wunsch (1970)) or the Smith-Waterman algorithm for local alignment (Smith and

Waterman (1981)). Cost functions can be made arbitrarily complex to reflect various biological similarities of molecular characters and the underlying insertion/deletion process. It is computationally intractable to apply the cost function to multiple sequences simultaneously, except in trivial-sized problems (Gotoh (1990); Wang and Jiang (1994)). Rather, alignment software must employ an iterative approach, in which the most similar sequences are first aligned and then less similar sequences are progressively added until all the sequences have been incorporated. The results of MSA are highly dependent on the order in which sequences are ordered (Berger and Munson (1991); Landan and Graur (2009)).

The relative accuracy of different MSA algorithms has not been comprehensively studied (Schwartz et al. (2005); Thompson et al. (2005)). Previous work showed that different MSA algorithms result in significantly different alignments, and these alignments ultimately yield different ML phylogenies (Wong et al. (2008)). However, the effects of different MSA algorithms on the downstream accuracy of ancestral reconstruction has not been investigated.

Here, I compared the performance of five different algorithms under a range of evolutionary conditions. I observed these algorithms created alignments that were significantly and consistently biased to be overaligned or underaligned, depending on the algorithm. I reconstructed ML phylogenies and ancestral sequences for these alignments. I observed that the accuracy of ancestral sequences varied among the five alignment algorithms, but the accuracy of the alignment itself was a poor predictor of the ancestral accuracy. My results show that the choice of alignment algorithm has non-trivial ramifications for the accuracy of downstream evolutionary inference.

Methods and Materials

Simulated Sequence Alignment

In order to determine the extent to which a variety of alignment algorithms affect the accuracy of ancestral sequence reconstruction, I simulated random amino acid sequences evolving on a four-taxon tree:

(((ta:0.2,tb:0.2):0.1,tc:0.25):0.1,td:0.3);

Simulations were initialized with random ancestral sequences 400 amino acids long. The ancestors were then evolved along the branches of the tree using the JTT model of amino acid substitution, combined with one of six different distributions of insertions/deletions (indels). I used (i) a small-mean and (ii) large-mean negative binomial distribution, (iii) a small-mean and (iv) large-mean Zipfian (power) distribution, and a (v) small-mean and (vi) large-mean uniform distribution. I varied the probability of indel events from 0.0 to 0.4, in increments of 0.04. I repeated the simulation twenty times for each combination of indel model and indel probability.

Alignment of Simulated Sequences

I aligned the simulated sequences using the default settings in the software Amap (cite Do et al. (2005); Schwartz et al. (2005)), Clustal (Thompson et al. (1994)) , Mafft (Katoh et al. (2002)), Muscle (Edgar (2004)), and Prank (Loytynoja and Goldman (2008)).

Phylogenetic Inference

I inferred ML phylogenies for each alignment, using my own in-house modifications to PhyML version 3.0. Trees were optimized using Multimax (described in chapter II), using the default settings.

Ancestral Sequence Reconstruction

I reconstructed ancestral sequences for the most-recent-common ancestor of taxa *ta* and *tb* using maximum likelihood as implemented in PAML version 4.2 and an in-house GUI – named Lazarus – that controls PAML (Yang (2007); Hanson-Smith et al. (2010)).

Ancestral Error

For every alignment, I measured the accuracy of the length of the ML ancestral sequence by counting the total number of sites in the ancestor, sans indel characters. I compared this value to the total number of sites in the true ancestral sequence, which had been recorded during its simulation.

Results

Variability of Alignment Length

The five alignment algorithms found significantly different alignments for sequences simulated under all six indel models (Fig. 32.). The algorithms Prank and Clustal systematically under-aligned sequences, whereas Amap, Muscle, and Mafft systematically over-aligned sequences. Across all conditions, Mafft created

alignments with the most accurate lengths, whereas Prank created alignments with the least-accurate lengths.

Effect on Ancestral Sequence Error

The accuracy of ancestral sequence lengths significantly varied among the five alignment algorithms (Fig. 33.). Across all methods, the absolute amount of ancestral error increased when the evolutionary conditions had large indel rates, and also when the true indel length distribution had a large mean value. Amap, Clustal, Muscle, and Prank inferred alignments whose ML ancestral sequences were generally too long. Mafft, in contrast, inferred alingments whose ML ancestral sequences were nearly always too short. Overall, the most accurate ancestral sequences came from alignments inferred by Prank and Mafft.

Discussion

My results demonstrate that the choice of alignment algorithm has significant consequences for the accuracy of the alignment, and for the downstream accuracy of the length of inferred ancestral sequences. Further, my results suggest that alignment error is a poor predictor of ancestral error. The least-accurate MSA algorithm – Prank – ultimately produced some of the most accurate alignments. Prank's superiority for ancestral reconstruction may be understood by comparing its underlying mechanisms to other algorithms. Whereas older MSA algorithms, such as Clustal, attempt to minimize the total number of indel characters in an alignment, the Prank algorithm attempts to minimize the total number of indel events – where a single event may include multiple contiguous indel characters.

Prank's alignments seem to preserve a phylogenetic signal that is highly amenable to placing ancestral indel characters.

My results suggest that phylogenetic practitioners should embrace alignment uncertainty, and repeat their evolutionary analysis using several different MSA algorithms in parallel.

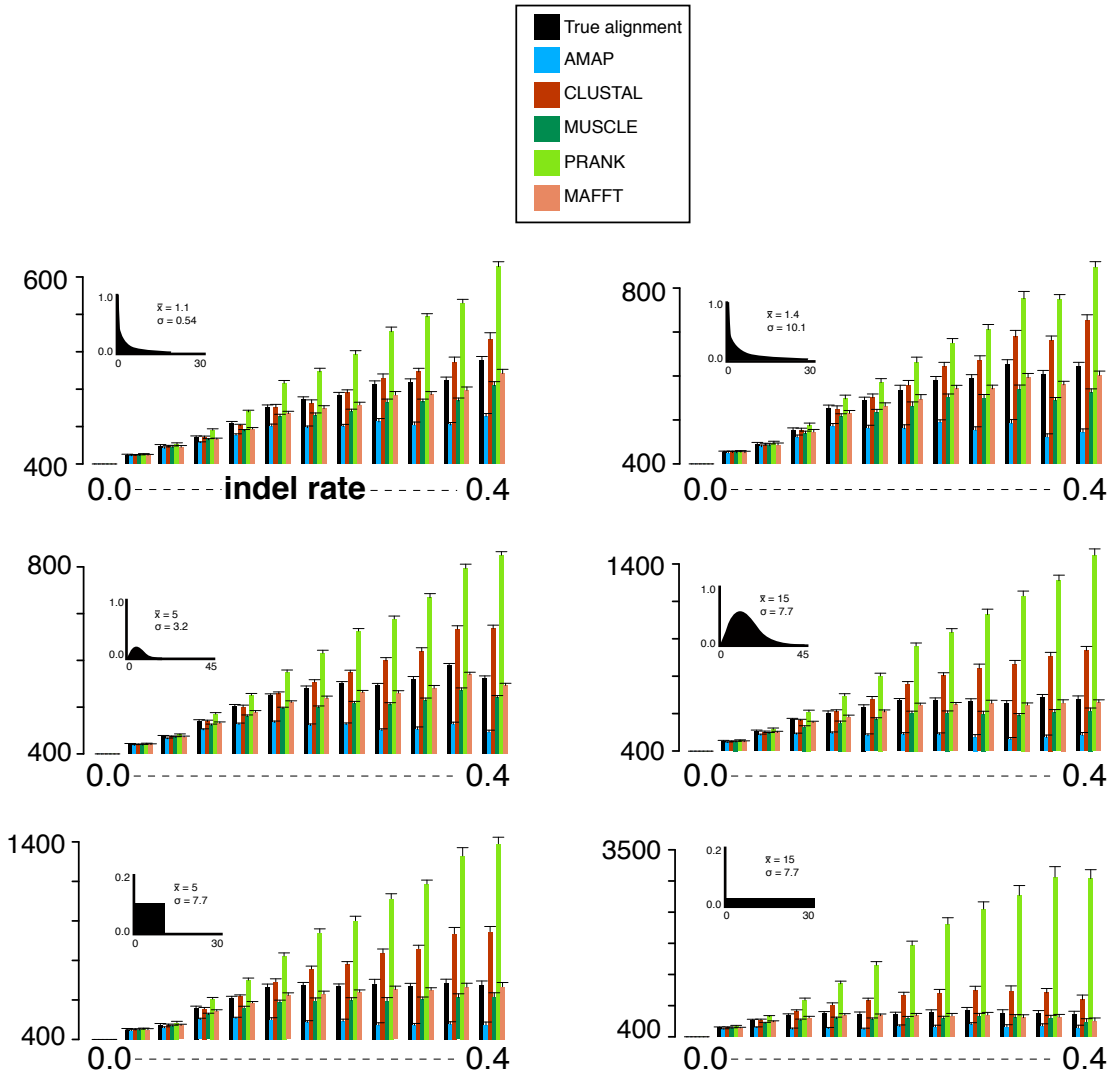


FIGURE 32. Alignment length versus insertion-deletion rate. Sequences were simulated on the tree described in section C, using six different models of insertion-deletion events (shown here in small insets). Vertical bars express the average length of the alignment for the given model, indel rate, and alignment algorithm. Error bars are standard error of the mean.

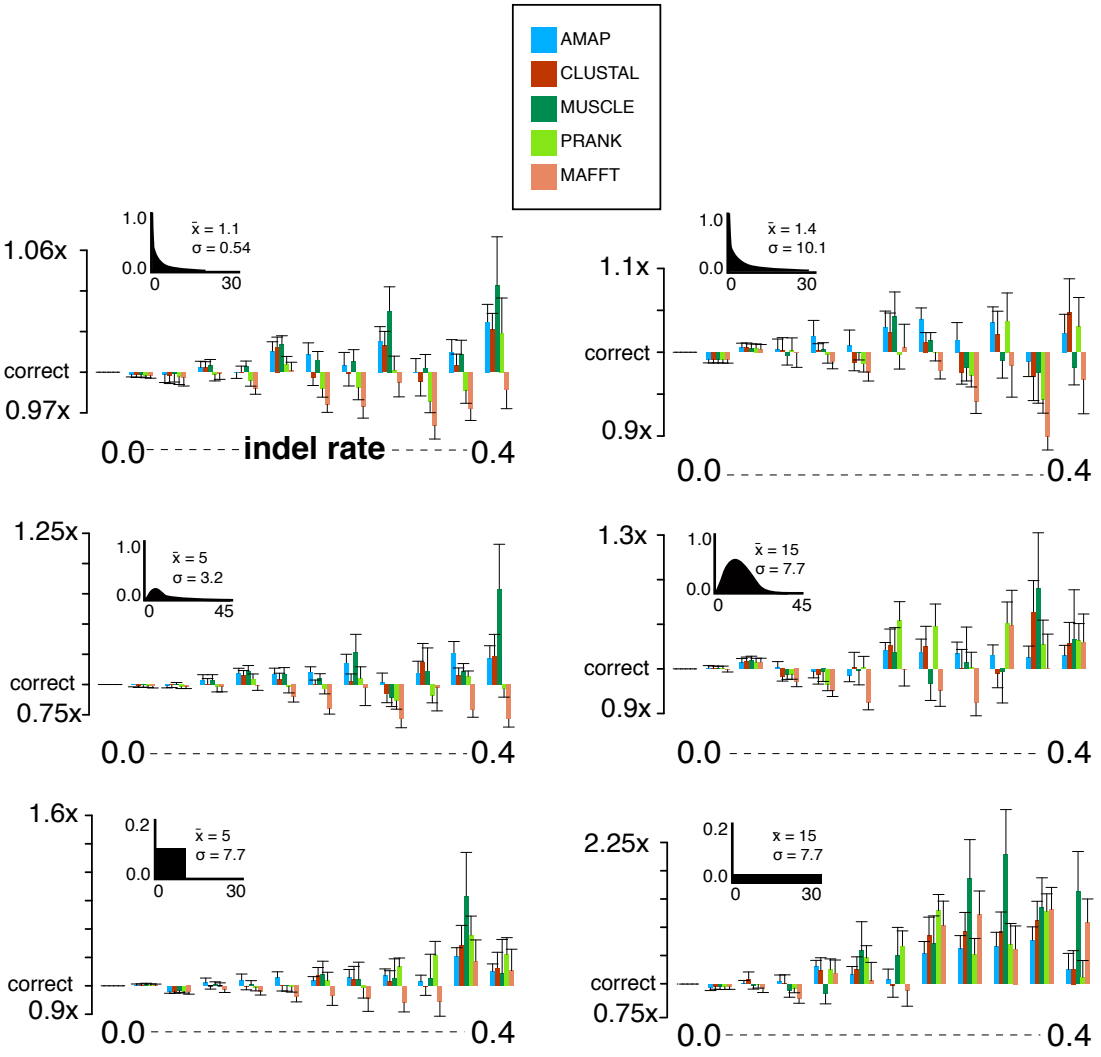


FIGURE 33. Ancestral length error versus insertion-deletion rate. Sequences were simulated on the tree described in section C, using six different models of insertion-deletion events (shown here in small insets). Vertical bars express the scaling factor by which ancestral sequences are too long or too short. Error bars are standard error of the mean.

APPENDIX D

V-ATPASE SUBUNIT PROTEIN SEQUENCES

The following list contains the GenBank accession IDs for protein sequences used in chapter IV. The IDs are labeled with the first letter of their genus, the full name of their species, and an integer number. 3, 11, and 16 indicate homology to yeast subunits c, c', and c", respectively.

M_musculus_16 NP_291095
O_mordax_3 ACO09611
D_rerio_3 NP_001098606
M_grisea_16 XP_369356
M_grisea_11 XP_366989
A_niger_3 XP_001399935
C_glabrata_3 XP_447321
A_terreus_11 XP_001214955
A_terreus_16 XP_001211600
C_parvum_16 XP_627363
P_vivax_16 XP_001616329
T_castaneum_3 XP_967959
X_tropicalis_3 NP_988893
G_zeae_3 XP_390178
L_elongisporus_3 XP_001526092
A_fumigatus_3 XP_001263225
M_grisea_3 XP_365764
S_pombe_3 NP_594799
C_albicans_3 XP_721376
B_bovis_16 XP_001612047
G_zeae_11 XP_388749
C_muris_3 XP_002141961
S_purpuratus_3 XP_797801
C_immitis_16 XP_001246494
M_mulatta_16 XP_001097275
Y_lipolytica_3 XP_505831
B_fuckeliana_16 XP_001552198
C_immitis_11 XP_001242880
A_mellifera_16 XP_392599
L_elongisporus_11 XP_001523616
T_annulata_16 XP_953463
N_vectensis_3 XP_001637733
S_Salar_16 NP_001134021
L_elongisporus_16 XP_001525467
C_neoformans_16 XP_773114

C_neoformans_11 XP_778255
P_knowlesi_16 XP_002261350
A_terreus_3 XP_001213329
A_pisum_3 NP_001155531
S_Salar_3 NP_001154112
P_marneffei_3 XP_002152865
P_stipitis_3 XP_001387092
A_gossypii_3 NP_984787
C_hominis_3 XP_667190
C_muris_16 XP_002142524
X_laevis_16 NP_001087741
M_mulatta_3 XP_001088617
M_brevicollis_16 XP_001742805
K_lactis_3 XP_454966
T_castaneum_16 XP_975026
C_intestinalis_16 XP_002131348
E_caballus_3 XP_001915231
G_zeae_16 XP_385476
C_globosum_3 XP_001229170
A_clavatus_3 XP_001271234
N_crassa_11 XP_965807
N_crassa_16 XP_964449
X_tropicalis_16 NP_001017064
P_anserina_11 XP_001907168
P_anserina_16 XP_001910317
H_sapiens_16 AAP36886
T_annulata_3 XP_952989
A_gossypii_11 NP_985409
A_gossypii_16 NP_983473
A_pisum_16 NP_001155679
N_crassa_3 XP_961418
V_polyspora_3 XP_001642185
T_adhaerens_3 XP_002112261
S_cerevisiae_3 NP_010887
P_marneffei_16 XP_002145395
D_hansenii_3 XP_460869
S_pombe_11 NP_593600
T_guttata_3 ACH45347
S_pombe_16 NP_594516
B_fuckeliana_3 XP_001553113
T_adhaerens_16 XP_002114348
S_cerevisiae_16 NP_011891
R_norvegicus_3 NP_033859
S_cerevisiae_11 NP_015090
B_taurus_3 NP_001017954

C_familiaris_3 XP_537002
P_anserina_3 XP_001911041
P_marneffei_11 XP_002147471
P_knowlesi_3 XP_002259621
T_guttata_16 NP_001232246
B_fuckeliana_11 CCD51873
C_cinerea_okayama_3 XP_001835649
P_falciparum_16 XP_001350256
C_intestinalis_3 XP_002132074
A_niger_11 XP_001391591
A_niger_16 XP_001397102
C_parvum_3 XP_627909
S_purpuratus_16 XP_790651
B_floridae_3 XP_002598155
Y_lipolytica_16 XP_505205
A_mellifera_3 NP_001011570
A_clavatus_16 XP_001275839
N_vectensis_16 XP_001638230
A_clavatus_11 XP_001274195
S_sclerotiorum_3 XP_001588693
P_falciparum_3 XP_001351750
D_discoideum_16 XP_644318
C_cinerea_okayama_11 XP_001835902
D_hansenii_16 XP_460013
D_hansenii_11 XP_458901
C_cinerea_okayama_16 XP_001830694
P_vivax_3 XP_001613765
C_immitis_3 XP_001239974
A_fumigatus_16 XP_755891
C_neoformans_3 XP_772642
H_sapiens_3 AAP36127
A_fumigatus_11 XP_753781
Y_lipolytica_11 XP_504637
M_brevicollis_3 XP_001743042
D_discoideum_3 XP_644319
P_stipitis_16 XP_001386908
C_albicans_16 XP_722165
C_albicans_11 XP_721376
P_stipitis_11 XP_001382501
X_laevis_3 NP_001082675
C_glabrata_11 XP_445959
E_caballus_16 XP_001916016
C_glabrata_16 XP_447739
V_polyspora_11 XP_001645235
V_polyspora_16 XP_001646358

C_hominis_16 XP_665533
C_globosum_11 XP_001222467
C_globosum_16 XP_001223715
O_mordax_16 ACO10130
D_rerio_16 NP_955855
B_bovis_3 XP_001609797
B_floridae_16 XP_002610356
M_musculus_3 NP_033859
C_familiaris_16 XP_539645
K_lactis_11 XP_452911
K_lactis_16 XP_454470
B_taurus_16 NP_001033127
S_sclerotiorum_16 XP_001590765
S_sclerotiorum_11 XP_001595091
R_norvegicus_16 AAH09169

REFERENCES CITED

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

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods, and other moulting animals. *Nature*, 387:489–493.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281.

Archibald, J. M., Logsdon, Jr, J. M., and Doolittle, W. F. (2000). Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in cct genes. *Molecular Biology and Evolution*, 17(10):1456–66.

Avise, J. C. (1998). The history and purview of phylogeography: a personal reflection. *Molecular Ecology*, 7:371–379.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, 18:489–522.

Baker, C. R., Tuch, B. B., and Johnson, A. D. (2011). Extensive dna-binding specificity divergence of a conserved transcription regulator. *Proceedings of National Academy of Science*, 108(18):7493–8.

Batzoglou, S. (2005). The many faces of sequence alignment. *Briefings in Bioinformatics*, 6(1):6–22.

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A., and Wheeler, D. L. (1999). GenBank. *Nucleic Acids Research*, 27(1):12–17.

Berger, M. P. and Munson, P. J. (1991). A novel randomized iterative strategy for aligning multiple protein sequences. *Computer Applications in the Biosciences*, 7(4):479–84.

Billera, L. J., Holmes, S., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733.

Blanchette, M., Green, E. D., Miller, W., and Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Research*, 14:2412–2423.

Brent, R. P. (1972). *Algorithms for Minimisation without derivatives (automatic computation)*. Prentice Hall.

Bridgham, J. T., Brown, J. E., Rodriguez-Mari, A., Catchen, J. M., and Thornton, J. W. (2008). Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genetics*, 4(9).

Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 307:97–101.

Bryant, D., Galtier, N., and Poursat, M.-A. (2005). *Mathematics of Evolution and Phylogeny*, chapter 2, pages 33–62. Oxford University Press.

Burks, C., Cinkosky, M. J., M.Fischer, W., Gilna, P., E.-D.Hayden, J., M.Keen, G., Kelly, M., Kristofferson, D., and Lawrence, J. (1992). Genbank. *Nucleic Acids Research*, 20:2065–2069.

Carroll, S., Grenier, J., and Weatherbee, S. (2005). *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Wiley.

Cavalli-Sforza, L. and Edwards, A. (1967). Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, 19(3):233–257.

Chang, B. S. W., Jonsson, K., Kazmi, M. A., Donoghue, M. J., and Sakmar, T. P. (2002). Recreating a functional ancestral archosaur visual pigment. *Molecular Biology and Evolution*, 19(9):1483–1489.

Chor, B., Hendy, M. D., Holland, B. R., and Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Molecular Biology and Evolution*, 17(10):1529–41.

Chor, B. and Tuller, T. (2005). Maximum likelihood on evolutionary trees: hardness and approximation. *Bioinformatics*, 1:97–106.

Clements, A., Bursac, D., Gatsos, X., Perry, A. J., Civciristov, S., Celik, N., Likic, V. A., Poggio, S., Jacobs-Wagner, C., Strugnell, R. A., and Lithgow, T. (2009). The reducible complexity of a mitochondrial molecular machine. *Proceedings of National Academy of Science*, 106(37):15791–5.

Darwin, C. R. (1859). *The Origin of Species*. John Murray.

Dean, A. M. and Thornton, J. W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Reviews Genetics*, 8(9):675–88.

Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005). Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340.

Dolezal, P., Likic, V., Tachezy, J., and Lithgow, T. (2006). Evolution of the molecular machines for protein import into mitochondria. *Science*, 313(5785):314–8.

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

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinaur Associates, Inc.

Fletcher, W. and Yang, Z. (2009). Indelible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–88.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45.

Forgac, M. (2007). Vacuolar atpases: rotary proton pumps in physiology and pathophysiology. *Nature Reviews Molecular Cell Biology*, 8(11):917–29.

Frattini, A., Orchard, P. J., Sobacchi, C., Giliani, S., Abinun, M., Mattsson, J. P., Keeling, D. J., Andersson, A. K., Wallbrandt, P., Zecca, L., Notarangelo, L. D., Vezzoni, P., and Villa, A. (2000). Defects in tcirg1 subunit of the vacuolar proton pump are responsible for a subset of human autosomal recessive osteopetrosis. *Nature Genetics*, 25(3):343–6.

Fukami, K. and Tateno, Y. (1989). On the maximum likelihood method for estimating molecular trees: uniqueness of the likelihood point. *J Mol Evol*, 28(5):460–4.

Gabaldón, T., Rainey, D., and Huynen, M. A. (2005). Tracing the evolution of a large protein complex in the eukaryotes, nadh:ubiquinone oxidoreductase (complex i). *Journal of Molecular Biology*, 348(4):857–70.

Gascuel, O. (1997). BIONJ: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–95.

Gaucher, E. A., Govindarajan, S., and Ganesh, O. K. (2007). Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature*, 451:704–707.

Gaucher, E. A., Thomson, J. M., Burgan, M. F., and Benner, S. A. (2003). Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, 425(18):285 – 288.

Goldstein, A. L. and McCusker, J. H. (1999). Three new dominant drug resistance cassettes for gene disruption in saccharomyces cerevisiae. *Yeast*, 15(14):1541–53.

Gotoh, O. (1990). Consistency of optimal sequence alignments. *Bulletin of Mathematical Biology*, 52(4):509–525.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, 59(3):307–321.

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

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*.

Harms, M. J. and Thornton, J. W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion Structural Biology*, 20(3):360–6.

Heath, T. A., Zwickl, D. J., Kim, J., and Hillis, D. M. (2008). Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Systematic Biology*, 57(1):160–6.

Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. (2011). Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A*, 108(19):7896–901.

Hillis, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology*, 47(1):3–8.

Hirata, T., Iwamoto-Kihara, A., Sun-Wada, G.-H., Okajima, T., Wada, Y., and Futai, M. (2003). Subunit rotation of vacuolar-type proton pumping atpase: relative rotation of the g and c subunits. *Journal of Biological Chemistry*, 278(26):23714–9.

Huelsenbeck, J. P. and Bollback, J. P. (2001). Empirical and heirarchical bayesian estimation of ancestral states. *Systematic Biology*, 50(3):351–366.

Huxley, J. (1942). *Evolution: The Modern Synthesis*. MIT Press.

Imamura, H., Takeda, M., Funamoto, S., Shimabukuro, K., Yoshida, M., and Yokoyama, K. (2005). Rotation scheme of v1-motor is different from that of f1-motor. *Proceedings of National Academy of Science*, 102(50):17929–33.

Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295):1161–6.

Jermann, T. M., Opitz, J. G., Stackhouse, J., and Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, 374:57–59.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1991). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282.

Kane, P. M. (2006). The where, when, and how of organelle acidification by the yeast vacuolar h+-atpase. *Microbiol Mol Biol Rev*, 70(1):177–91.

Katoh, K., Misawa, K., ichi Kuma, K., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.

Kelmanson, I. V. and Matz, M. V. (2003). Molecular basis and evolutionary origins of color diversity in great star coral montastraea cavernosa (scleractinia: Faviida). *Molecular Biology and Evolution*, 20(7):1125–33.

Kolaczkowski, B. and Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431(7011):980–4.

Kolaczkowski, B. and Thornton, J. W. (2007). Effects of branch length uncertainty on bayesian posterior probabilities for phylogenetic hypotheses. *Molecular Biology and Evolution*, 24(9):2108–2118.

Kolaczkowski, B. and Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, 25(6):1054–1066.

Kolaczkowski, B. and Thornton, J. W. (2009). Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, 4(12):e7891.

Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal Molecular Evolution*, 42:313–320.

Landan, G. and Graur, D. (2009). Characterization of pairwise and multiple sequence alignment errors. *Gene*, 441(1-2):141–7.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–109.

Lewontin, R. C. (1972). The apportionment of human diversity. *Evolutionary Biology*, 6:381–398.

Liberles, D., editor (2007). *Ancestral Sequence Reconstruction*. Oxford University Press.

Liu, R. and Ochman, H. (2007). Stepwise formation of the bacterial flagellar system. *Proceedings of National Academy of Science*, 104(17):7116–21.

Loytynoja, A. and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Science*, 102(30):10557–10562.

Loytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5884):1632–1635.

Lucena, B. and Haussler, D. (2005). Counterexample to a claim about the reconstruction of ancestral character states. *Systematic Biology*, 54(4):693–695.

Lynch, M. (2007a). The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, 8(10):803–13.

Lynch, M. (2007b). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of National Academy of Science*, 104 Suppl 1:8597–604.

Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y., and Koonin, E. V. (2007). Inventing the dynamo machine: the evolution of the f-type and v-type atpases. *Nature Reviews Microbiology*, 5(11):892–9.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search of similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer Series in Operations Research. Springer.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, 3(8):1405–1408.

Ohno, S. (1970). *Evolution by gene duplication*. Springer Berlin / Heidelberg.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., and Thornton, J. W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317(5844):1544–8.

Ott, M., Zola, J., Stamatakis, A., and Aluru, S. (2007). Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In *Proceedings of the 2007 ACM/IEEE conference on Supercomputing*, SC '07, New York, NY, USA. ACM.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53(5):673–684.

Pallen, M. J. and Matzke, N. J. (2006). From the origin of species to the origin of bacterial flagella. *Nature Reviews Microbiology*, 4(10):784–90.

Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular restoration studies of extinct forms of life. *Acta Chemica Scandinavia*, A(17):S9–S16.

Pereira-Leal, J. B., Levy, E. D., Kamp, C., and Teichmann, S. A. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol*, 8(4):R51.

Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z.-M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T. J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J. M., Gaucher, E. A., and Fernandez, J. M. (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol*, 18(5):592–6.

Pérez-Sayáns, M., Somoza-Martín, J. M., Barros-Angueira, F., Rey, J. M. G., and García-García, A. (2009). V-atpase inhibitors and implication in cancer treatment. *Cancer Treat Reviews*, 35(8):707–13.

Pollock, D. D., Zwickl, D. J., McGuirec, J. A., and Hillis, D. M. (2002). Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology*, 51(4):664–671.

Posada, D. (2001). The effect of branch length variation on the selection of models of molecular evolution. *J Mol Evol*, 52(5):434–44.

Powell, B., Graham, L. A., and Stevens, T. H. (2000). Molecular characterization of the yeast vacuolar h+-atpase proton pore. *Journal of Biological Chemistry*, 275(31):23654–60.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). Numerical recipes in c: The art of scientific computing. second edition.

Pupko, T., Shamir, I. P. R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, 17(6):890–896.

Raff, R. (1996). *The Shape of Life: Genes, Development and the Evolution of Animal Form*. University of Chicago Press.

Rambaut, A. and Grassly, N. C. (1997). Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 03(1):92–94.

Rogers, J. S. and Swofford, D. L. (1999). Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. *Molecular Biology and Evolution*, 16(8):1079–85.

Rokas, A., Krüger, D., and Carroll, S. B. (2005). Animal evolution and the molecular signature of radiations compressed in time. *Science*, 310(5756):1933–8.

Ryan, M., Graham, L. A., and Stevens, T. H. (2008). Voa1p functions in v-atpase assembly in the yeast endoplasmic reticulum. *Molecular Biology and Evolution*, 19(12):5131–42.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Sambrook, J. and Russel, D. W. (2001). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press.

Schultz, T. R. and Churchill, G. A. (1999). The role of subjectivity in reconstructing ancestral character states: A bayesian approach to unknown rates, states, and transformation asymmetries. *Systematic Biology*, 48(3):651–664.

Schwartz, A. S., Myers, E. W., and Pachter, L. (2005). Alignment metric accuracy. *arXiv*. on arXiv.org at http://arxiv.org/abs/q-bio.QM/0510052.

Shi, Y. and Yokoyama, S. (2004). Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proceedings of the National Academy of Science*, 100(14):8308–8313.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–7.

Smith, W. L. and Wheeler, W. C. (2004). Polyphyly of the mail-cheeked fishes (teleostei: Scorpaeniformes): evidence from mitochondrial and nuclear sequence data. *Molecular Phylogenetics and Evolution*, 32(2):627–46.

Sokal, R. R. and Sneath, P. H. (1963). *Principles of numerical taxonomy*. W.H. Freeman, San Francisco.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Steel, M. (1994). Recovering a tree from the leaf colourations it generates under a markov model. *Applied Mathematics Letters*, 7:19–23.

Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-dna interactions. *Nature Reviews Genetics*, 11(11):751–60.

Swofford, D. and Olsen, G. (1990). Phylogeny reconstruction. In Hillis, D. M. and Moritz, C., editors, *Molecular Systematics*, chapter 11, pages 411–501. Sinauer, Sunderland, MA.

Swofford, D. L. (2003). Phylogenetic analysis using parsimony (and other methods) 4.0 b10. Sineer Associates, Inc.

Taylor, J. W. and Berbee, M. L. (2006). Dating divergences in the fungal tree of life: review and new analyses. *Mycologia*, 98(6):838–49.

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

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–36.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., Kee, D. W. D., Li, T., Aris, J. P., and Benner, S. A. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics*, 37(6):630–635.

Thornton, J. W. (2004). Resurrecting ancient genes: Experimental analysis of extinct molecules. *Nature*, 5:366–375.

Thornton, J. W., Need, E., and Crews, D. (2003). Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. *Science*, 301(5640):1714–1717.

Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Ménard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.-M., Shapiro, J., Sheikh, B., Suter, B., Wong, S. L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H., and Boone, C. (2004). Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–13.

Ugalde, J. A., Change, B. S. W., and Matz, M. V. (2004). Evolution of coral pigments recreated. *Science*, 305:1433 – 1433.

Umemoto, N., Ohya, Y., and Anraku, Y. (1991). Vma11, a novel gene that encodes a putative proteolipid, is indispensable for expression of yeast vacuolar membrane h(+)-atpase activity. *Journal of Biological Chemistry*, 266(36):24526–32.

Umemoto, N., Yoshihisa, T., Hirata, R., and Anraku, Y. (1990). Roles of the vma3 gene product, subunit c of the vacuolar membrane h(+)-atpase on vacuolar acidification and protein transport. a study with vma3-disrupted mutants of saccharomyces cerevisiae. *Journal of Biological Chemistry*, 265(30):18447–53.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–48.

Wang, Y., Cipriano, D. J., and Forgac, M. (2007). Arrangement of subunits in the proteolipid ring of the v-atpase. *Journal of Biological Chemistry*, 282(47):34058–65.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319:416–417.

Xu, L., Shen, X., Bryan, A., Banga, S., Swanson, M. S., and Luo, Z.-Q. (2010). Inhibition of host vacuolar h+-atpase activity by a legionella pneumophila effector. *PLoS Pathogens*, 6(3):e1000822.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. *Journal of Molecular Evolution*, 51(5):423–32.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650.

Yokoyama, S., Tada, T., Zhang, H., and Britt, L. (2008). Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of National Academy of Science*, 105(36):13480–5.

Zheng, L., Baumann, U., and Reymond, J.-L. (2004). An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Res*, 32(14):e115.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, University of Texas, Austin.