CONVERGENCE AND EVOLVING AMINO ACID PROPENSITIES

by

STEPHEN T. POLLARD

B.A., Princeton University, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

PhD, Structural Biology and Biochemistry

Structural Biology and Biochemistry Program

2019

This thesis for the PhD, Structural Biology and Biochemistry degree by

Stephen T. Pollard

has been approved for the

Structural Biology and Biochemistry Program

by

Brad Bendiak, Chair

Mair Churchill

Robert Hodges

Nicholas Rodrigue

Date: December 13, 2019

Pollard, Stephen T. (PhD, Structural Biology and Biochemistry)

Convergence and Evolving Amino Acid Propensities

Thesis directed by Professor David D. Pollock

ABSTRACT

Understanding the relationship between structure and function of proteins is essential for studying protein dysfunction and disease and for engineering therapeutic proteins with novel functions. The details of how proteins perform their function and the effects of mutations on that function are two important facets of this relationship. Both protein structure and function have been shaped by evolution and so evolution can inform predictions about protein function and the effects of mutation.

One important aspect of protein evolution is the set preferences for each amino acid at every site in the protein, referred to as propensities. These preferences are affected by overall protein stability and thermodynamics, active sites, cofactor binding sites, and other factors that affect fitness. The propensities are often assumed to be constant over time and across the protein structure for varying reasons, including model simplicity and computational cost, but there is ample evidence that the propensities do vary in time and space. Molecular evolutionary studies of these phenomena include coevolution, epistasis, covarions, and heterotachy.

The main goal of the this thesis is to improve the computational methods of studying evolution and propose a new method for identifying how amino acid propensities change over sites and time.

The form and content of this abstract are approved. I recommend its publication.

Approved: Brad Bendiak

This thesis is dedicated to my wife, Rebecca, who supported me throughout graduate school. I would like to thank my parents for the help they have provided throughout my life. And finally to my dog, Calvin, who sat by me for many days of writing.

ACKNOWLEDGEMENTS

I acknowledge the support of the National Institutes of Health (NIH; GM083127 and GM097251) to David Pollock.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

I.1 Evolution and the definition of biology

Biology is the study of life, from the Ancient Greek words $\beta\iota o$ (bio), meaning "life", and $\lambda o\gamma\iota\alpha$ (logia) meaning "branch of study". This definition seems straight-forward, however surprisingly the definition of life has been debated for millennia. From Greek scholars like Aristotle to Renaissance philosophers such as Descartes, up to modern institutions like NASA, humans have been grasping for a definition of life that agrees with our reason and intuition [1, 2, 3, 4]. The great physicist Erwin Shroedinger even weighed in on the topic in his famous book titled "What is life?", defining life as a system that reduces its own internal entropy while increasing the entropy of its environment [5].

Some have gone so far as to say that the definition of life depends on its ability to evolve. NASA's definition of life requires evolution, defining life to be "a self-sustained chemical system capable of undergoing Darwinian evolution" [3, 4]. Theodosius Dobzhansky, one of the founders of the modern synthesis of evolution, expanded the centrality of evolution to biology, claiming famously that "nothing in biology makes sense except in the light of evolution" [6].

I.2 Early evolution

Even before Darwin, humans have wondered how all the diversity of different organisms came to be. Evolutionary biology is the field surrounding the wonders of the origins of life and how life has changed over time. The first major questions in the field were focused on taxonomy and trying to establish well defined families of species. Assuming that all life came from the last universal common ancestor (LUCA), scientists believed early on that the histories of every living thing must be a tree structure: the Tree of Life [7, 8, 9]. In order to reconstruct this tree, however, evolutionary biologists needed to know the amount of relatedness between different extant (meaning currently existing; as opposed

to extinct) species and approximately how long ago those different species had a common ancestor.

The best evidence to answer the latter question comes from paleontology discoveries of ancient bones. Some of these extinct species looked similar enough to extant species of today to predict that they are ancestors of current species. For example, discovering early mammal fossils helped shed light on when mammals first evolved and what they were like (e.g. [10, 11, 12, 13]). We can get an idea of approximately how quickly evolution happens from studies like this, but different species evolve at different rates, and so determining the time back to the LUCA is in fact a difficult undertaking [14, 15]. Early studies and hypotheses about the amount of relatedness among species were based on morphological differences, meaning body shape or non-sequence genetic information. Distances between species could be calculated using different biometrics such as the length of a femur. Once distances between species were estimated, trees could be constructed. Many different methods for building trees have been developed and will be discussed later.

Sequencing of biological polymers from different species introduced a new way of calculating distances between species and the new field of molecular phylogenetics. Beginning with protein sequencing data in the 1970s, biological sequences have become more important to evolutionary analysis, including RNA and DNA [16]. Eventually the majority of the information used to determine the relatedness among species shifted from phenotype to genotype. Huge amounts of genetic sequences were amassed from a wide variety of organisms. The more sequence data we have, the better we can study how evolution really works. We have enough information to build models to explain how that genetic information has been created. Our current models of evolution can explain some of our observations about biological sequences, however they cannot explain certain discoveries about convergent evolution. We still lack the computational techniques to test increasingly complex models of evolution on the acquired sequence data.

I.3 Modeling sequence evolution

An essential part of the pursuit of science is learning how the world works. Unfortunately the real world is infinitely complicated, and so we must make do with simplifications, which are called models [17]. In general, scientific and statistical models connect observable data with unobservable variables (sometimes called parameters). In phylogenetics, the observables often are sequences either from proteins, DNA, or any other kind of biological sequence, though sometimes other phenotypes are used. The unobservable variables are the evolution process, the evolutionary tree that connects the sequences together, and the substitution history (that is which substitutions happened on which branches at which sites).

The earliest methods to produce a tree from sequences minimized the number of substitutions required across the whole tree to explain the sequence data. This method is called "minimum evolution" or "maximum parsimony" [18, 19]. Part of the logic behind these methods is that substitutions are relatively rare and the simplest tree is probably the best one. One negative of this process is that it does not utilize a proper model of evolution that it is testing. One of the earliest models of sequence evolution, the Jukes-Cantor model, approximated the evolutionary process as a Poisson process with a single rate parameter λ: every substitution from one residue to another occurred at the same rate [20]. This rate is an instantaneous rate of substitution which results in an exponentially decreasing probability that the site has not substituted.

$$P(t) = \lambda e^{-\lambda t} \tag{I.1}$$

Kimura noticed that not all nucleotide substitutions seem to occur at the same rate. Nucleotides can be divided into two classes: purines and pyrimidines. Pyrimidines such as cytosine (C), thymine (T), and uracil (U) contain a single ring structure, while purines such as adenine (A) and guanine (G) have a double ring structure. Substitutions within

each class (called 'transitions', e.g. C to T) occur more frequently than substitutions across the classes (called 'transversions', e.g. C to A). Kimura introduced separate transition and transversion rates in order to model this observation [21]. Later differing stationary frequencies (π) for each residue and differing substitution rates among the residues were introduced into a general time reversible model (GTR; [22]). These stationary frequencies for amino acids could be interpreted as the amino acid propensities or relative fitnesses. The categories model (CAT) is an infinite mixture model that groups sites into classes and adapts the number of site classes to the complexity of the data [23]. Its goal is to capture across site variation in the amino acid frequencies, and the CAT-GTR model also captures rate matrix variation across sites [23].

Amino acid propensities have been well studied and included in many protein evolution models since their introduction in 1975 [24, 25]. Propensities (Π) are defined in terms of the fitness of each amino acid at a site by:

$$\Pi_i^x = \frac{\omega_i^x}{\sum_y \omega_i^y} \tag{I.2}$$

where x is the amino acid at site i, and ω_i^x is the fitness of amino acid x at site i [26]. The site-specific propensities have been shown to depend on secondary structures, a site's proximity to the surface, and interaction energies from nearby amino acids, tertiary structure, protein function, and protein-protein interactions [27, 28, 29, 30, 31, 32, 33, 34]. Overall protein stability and thermodynamics can also have strong effects on the propensities at a site [35, 36]. Measuring the amino acid propensities at a single site is limited by the amount of information at the site, and so propensity estimates are often averaged over many sites in an alignment using site-heterogeneous models. The number of groups and assignment of sites to groups can be predefined or estimated during the analysis [27, 37, 38, 23]. Gathering sites into distinct groups has been useful in finding propensity differences between different regions in a protein structure, but by averaging across sites, this method loses specific information about individual sites. Previous work

has taken into account secondary structures, proximity to surface, and interaction energies, and tertiary structure [27, 28, 29, 30, 31], however these novel types of models that include structural and biophysical information break through the barrier of averaging over a group of sites. Many previous attempts had strict groups of sites, rather than allowing sites to be influenced by multiple groups [27, 28, 23, 39, 40]. Sometimes this average model of evolution is seen as the model under which the sites have evolved, however recent evidence suggests otherwise. In this thesis, I show that the average model does not approximate the site-specific model well in important ways.

After grouping different sites together, the averaged propensities have often been assumed to be constant over time. Evidence that the evolution process changes over time have come from studies of coevolution, epistasis, covarions, and heterotachy, as well as convergence [30, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50]. Time- or branch-heterogeneous models allow the way in which evolution proceeds, or the evolutionary process, to change over time, but in doing so, they often introduce a large number of parameters. Methods to limit the number of parameters in a time-heterogeneous evolution process have been utilized including only allowing the G+C content to change over the tree, only allowing change at branch break points, and breaking the tree into groups of contiguous branches and allowing a single evolution process per group [51, 52, 53, 54].

This research proposes a method which will allow the study of how quickly propensities change over time and help guide efforts to model changing propensities toward an improved view of the evolution process. By providing better models, this work will affect many areas of study including ancestral reconstruction research, which aims to resurrect ancient proteins and predict the ancestral function. Incorporating changing constraints, in the ways that this research proposes, will improve the ability to resurrect ancient proteins, which will also lead to higher confidence in the ancestral function predictions and tests.

The site- and time-heterogeneous models can be viewed as averages of the instantaneous evolution process as it varies across sites and time, and a few models have been proposed

[43, 55, 56, 44, 45, 57]. Methods for studying the instantaneous site-specific propensities are limited, though recent advances using convergence are promising [48]. A site- and time-heterogeneous model would allow the instantaneous evolution process at a site to change even if the amino acid at that site does not change [35, 57]. Therefore, the same amino acid at two different sites or the same site at two different times could have different amino acid propensities. In Chapter V, I propose a novel and general method which will allow the amino acid propensities to change across sites and across time.

I.4 Evolutionary model and tree testing

The methods for testing and comparing how well models and trees fit the sequence data have changed and improved along with the models. Early on the models and techniques were tied tightly together, as with the "maximum parsimony" or "minimum evolution" methods [58]. These used implicit models of evolution, rather than explicitly structuring the model to be tested. Many ways have been developed to test the validity of the evolutionary models. There are two main philosophies of statistics and model comparison in the phylogenetics field: frequentist and Bayesian. Both agree that in order to test the model or compare it against other models, the likelihood of a model given the data must be calculated. In phylogenetics, often the model is the model of the evolution process and the data is the multiple sequence alignment. The phylogenetic tree that connects the sequences is sometimes considered known data and sometimes it is a variable part of the analysis. Models can be broken down into the parameters of the model and model structure, which is how those parameters come together to produce probabilities of events.

In the frequentist perspective, usually the goal is to find the maximum likelihood model structure, model parameters, and tree for a given multiple sequence alignment. The first application of this method to phylogenetics was in 1963, where the authors attempted to fit the data to a model of minimum evolution, which is a form of maximum likelihood statistics [58, 18]. This method is popular for models for which there is a closed-form solution to the likelihood calculation, because the maximum likelihood parameter values

can simply be calculated. When there is not a closed-form solution, then various different optimization techniques can be used to search the parameter space and find the maximum likelihood values.

In the maximum likelihood framework, models are compared by assuming a null model and an alternative hypothesis or model. The likelihoods are compared and the null model is rejected in favor of the alternative if the alternative model's likelihood is significantly higher than the null model's likelihood. How much higher the likelihood must be is determined by how many more degrees of freedom (or parameters) the alternative model has than the null. With enough data, the expected distribution of the likelihood ratio between alternative and null hypotheses follows a χ^2 distribution. The null is rejected if the p-value of the observed likelihood ratio is above a chosen significance level, with 0.05 as a common choice, where a p-value is the probability of observing a likelihood ratio at least as high as one actually observed under the null.

Bayesian methods are sometimes chosen over maximum likelihood approaches for a number of reasons including not having a closed form solution for the likelihood of every model and potential biases [59, 60]. Bayesian statistics uses Bayes' theorem in order to update our belief about the world from a set of prior understandings of the world to a set of posteriors. This view naturally works well with the scientific method and slow, stepwise refinement of models of phenomena. Usually the proposed new model is slightly more complicated than the old model in a nested manner, that is the structure of the more complicated model can match the structure of the less complicated model with certain parameter values. A simple example of this nesting would be testing whether the sites in a multiple sequence alignment can be split up into two sets of sites with different substitution rates, or if all the sites share a single substitution rate. If the two rates in the two-rate model are the same, then the structure is identical to the single rate model. Because of this, the more complicate model will always have an equal or higher maximum likelihood that the simpler model.

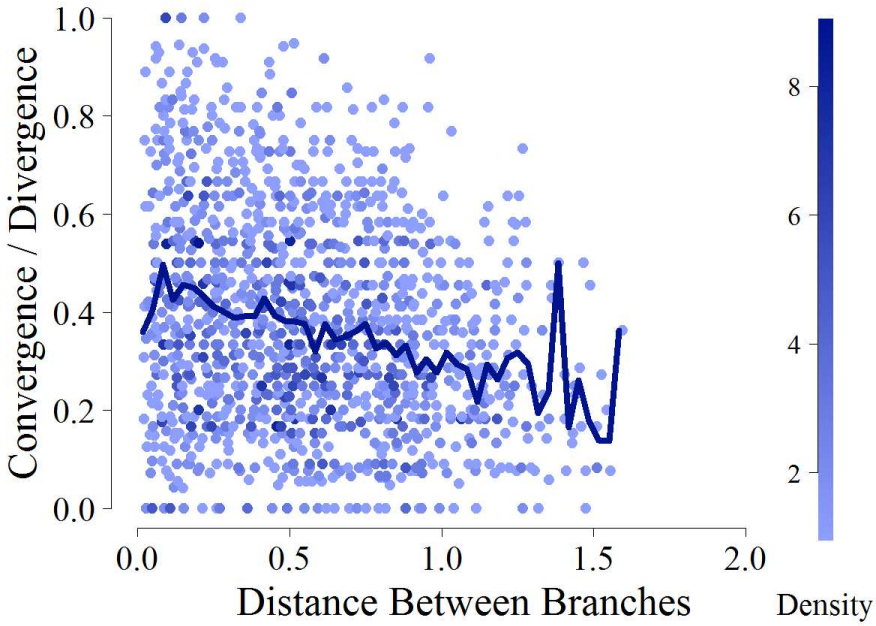


Figure I.1: Convergence calculated from the mitochondrial data set with same ancestors.

The goal of Bayesian statistics is not to determine the maximum likelihoods of the models considered, rather it is to estimate the posterior distributions of the parameters in the models. One can then construct a credible interval for each parameter in question as an estimate of that parameter. Given these parameter distributions, there are a number of methods to compare models including using the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or Bayes factors [61, 62, 63, 64].

I.5 Prior evidence for amino acid propensities varying across sites and time

In Goldstein et al., we recently showed that at any given point along a protein's evolution, the average site specific constraint is much higher than previously predicted [48]. This was shown by estimating the average amount of convergence between two sister branches and then calculating the expected constraint (see Figure I.1). We estimated that the average site was constrained to around 4 amino acids, based on the observed convergence levels.

We analyzed commonly used evolution models and estimated the amount of convergence

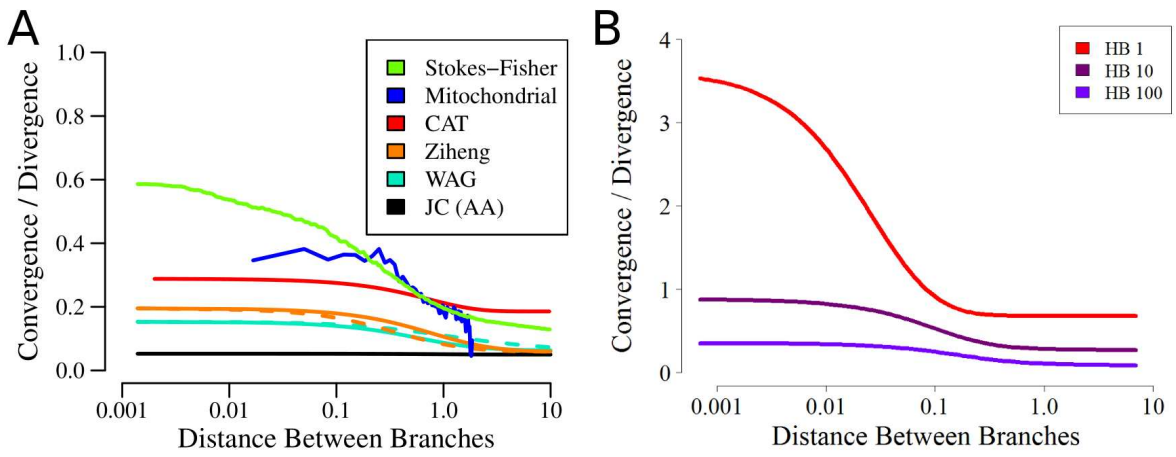


Figure I.2: A) Convergence predicted by common models. B) The expected convergence from the Halpern and Bruno model fit to trees of different lengths: 1 (red), 10 (purple), 100 (blue) evolutionary units.

in these models to determine the constraint. We used the simple Jukes-Cantor model (JC; equal rates of amino acid exchange), a codon model (Z; Zihengian), and Z with Gamma distributed rate parameters, and the CAT-60 model (CAT), which includes variation in constraint among sites [39, 20, 65]. The CAT-60 model was the most constrained and allowed around 6 amino acids per site, which is a significant decrease in the amount of constraint per site. Figure I.2A shows that the model based approach for estimating the amount of expected convergence performs poorly. The site-specific models perform better than the site-homogeneous models, but still not well. The convergence from the Stokes-Fisher model, which allows propensities to change over time, matches the observed convergence well [35]. All of the models analyzed, except for the Stokes-Fisher model, integrate over sites and time. By integrating over sites and time, these models reduce the average constraint per site. As shown by the expected convergence, this averaged constraint clearly does not match the instantaneous constraint shown in the mitochondrial data. These models also do not directly take into account structure and it is known that structure is important in constraints.

How well one can measure the site-specific propensities can also greatly affect one's ability to predict quantities of interest about the proteins, such as the amount of expected

convergence. In preliminary data, the Halpern and Bruno model, which allows every site to have a unique set of constant propensities, was fit to sequences evolved under the Stokes-Fisher model, which allows propensities to change [35]. A Halpern and Bruno model fit using a short amount of evolutionary time (1 unit corresponding to an average of 1 substitution per site) produced estimates of convergence that are far too high (see Figure I.2B). If the fit encompasses a much longer time, 100 units of evolutionary time, the convergence is underestimated. One can under- or overestimate the convergence using HB depending on how long your tree is. This could indicate that the model is estimating the propensities poorly.

In Goldstein et al., we showed that the amount of convergence between two branches on a mitochondrial phylogenetic tree decreases as the branches grow further apart [48]. We also calculate the convergence and divergence when the same site on both branches has the same amino acid, the convergence still decreases with time (see Figure I.1). This result can only be explained by the propensities changing.

The location of a site in the protein structure can greatly influence the site-specific amino acid preferences [27, 28, 29]. One result from this is that amino acids are clearly not distributed uniformly throughout a protein. As an example, let's consider Cytochrome C Oxidase. Some amino acids are equally distributed throughout the protein complex, such as Leucine, shown in Figure I.3A in red. Other amino acids are clearly concentrated at the hydrophilic caps of the transmembrane protein, such as Arginine shown in Figure I.3B in red.

These preferences could be heavily influenced by the hydrophobicity or polarity of the amino acids. Figure I.4 shows the Cytochrome C Oxidase and Cytochrome B structures with the polar and the hydrophobic amino acids in purple and yellow, respectively. Clearly the sites near the middle of the transmembrane alpha helices prefer hydrophobic amino acids, while the ends of the alpha helices prefer polar amino acids. Often amino acid preferences are modeled as discrete regions with discrete sets of preferences. For example,

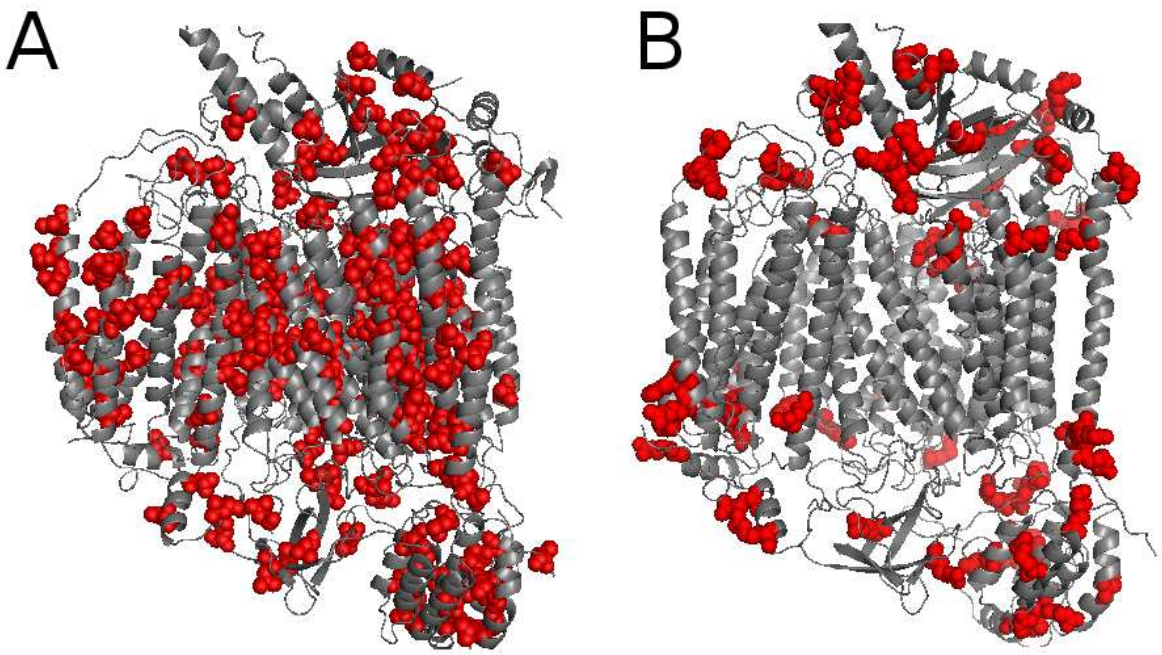


Figure I.3: Leucine (A) and Arginine (B) highlighted on the Cytochrome C Oxidase structure.

the structures in Figure I.4 could be divided into the region exposed to the solvent and the region buried in the membrane.

The new method proposed in Chapter V can estimate how the amino acid preferences change across tertiary structure, solvent exposure, buriedness, etc. by allowing each site to evolve independently of other sites and not averaging the evolution process across sites. The model proposed is distinct from previous site-focused methods such as the CAT model, in that it does not average over sites at all.

The main goal of the this thesis is to improve the computational methods of studying evolution and propose a new method for identifying how amino acid propensities change over sites and time. I motivate the creation of new models using inferred molecular convergence. I show how amino acid propensities can shift dramatically and that large shifts correlate with adjacent substitutions. I propose a new way of testing models by integrating over likely trees using a Bayesian method named "Markov Katana", which is novel and builds off previous techniques. In this research, I bring together the ideas of

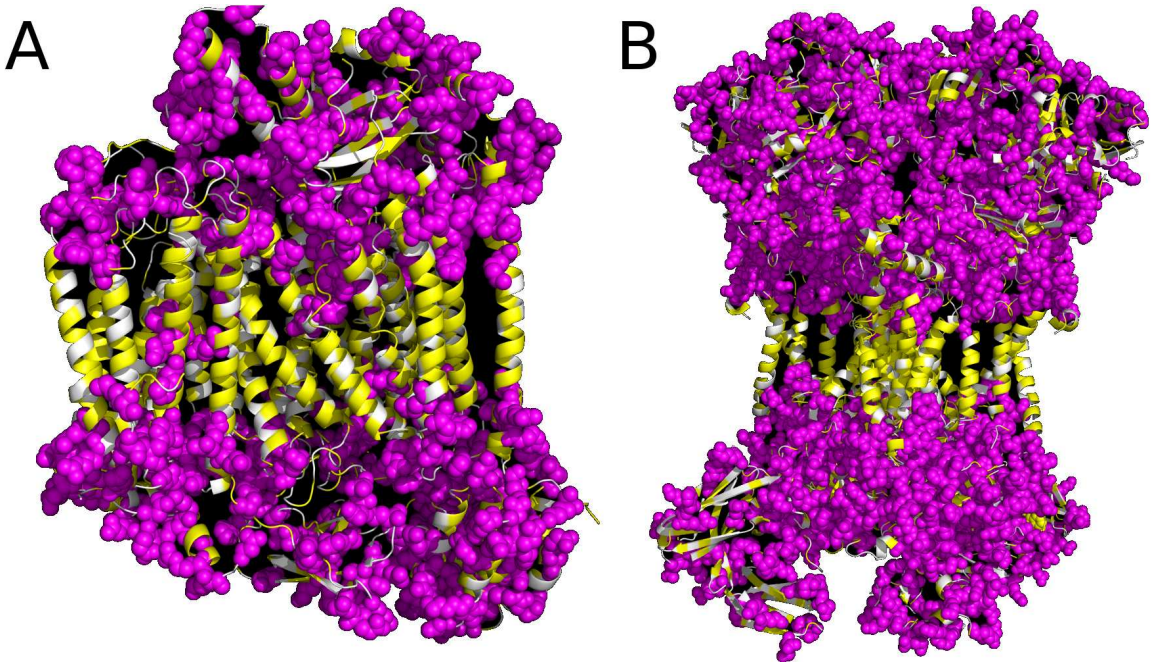


Figure I.4: Cytochrome C Oxidase (A) and Cytochrome B (B) structures with polar (purple) and hydrophobic (yellow) amino acids highlighted.

phylogenetics, molecular convergence, and changing amino acid propensities in order to achieve an integrated understanding of protein evolution.

I.6 Chapter overviews

In Chapter II we discuss the differences between mechanistic models and phenomenological models. Models of sequence evolution are used ubiquitously in biology from phylogenetic reconstruction to the analysis of adaptation, coevolution, and convergence. The structure of the model used affects these analyses, and it is therefore preferable to use good models. The field of molecular evolution is currently undergoing an important transformation due to large increases in the ability to collect and analyze massive amounts of data. Here we briefly review the history of molecular evolution and then discuss how evidence of epistasis and convergent molecular evolution helps overturn traditional models of protein evolution. We conclude by discussing desired features in a simple mechanistic model of protein evolution that is more compatible with patterns observed in real and simulated protein evolution. The distinction between mechanistic and phenomenological

models is used to advocate for better models in Chapter IV and to support the model used in Chapter V.

Chapter III explores convergence and demonstrates its ability to shed light on the average constraint at a site and to disrupt phylogenetic inference if not handled properly. Convergence is a central concept for phylogenetic inference because it can occur when two lineages respond in the same way to similar adaptive pressure. Convergence may thus serve as a signal of adaptive response, and when it occurs it provides information on the replicability of particular adaptive responses. Comparative genomics approaches have begun to allow the analysis of convergence at all levels from the phenotype to physiological systems to proteins and other functional macromolecules, thus allowing researchers to dissect the molecular mechanisms that give rise to phenotypic convergence. In turn, this allows study of the replicability of convergence at all levels, from repeated modifications of the same genes to repeated modifications of the same protein positions to the same amino acids. It is particularly important to understand how species divergence leads to altered constraints affecting convergence at the molecular and various systems levels. These altered constraints may affect the probabilities of adaptive as well as non-adaptive convergence. A major focus of research on molecular convergence concerns model-building and empirical techniques to distinguish ubiquitous but variable levels of non-adaptive convergence from more rare but interesting adaptive convergence events. "Parallelism" is often used to discretely categorize different types of convergent events, but its use is controversial and often contradictory among different authors. Because of this, we suggest that its use be discontinued in favor of focusing on the various evolutionary issues it is intended to embody.

In Chapter IV, I use the observed amount of non-adaptive convergence in a mitochondrial protein alignment to learn about the average constraint at sites. I also use the decrease in convergence over time to infer important characteristics about the evolution process. Convergence is a central concept in evolutionary studies because it provides

strong evidence for adaptation. It also provides information about the nature of the fitness landscape and the repeatability of certain evolutionary processes, and can mislead phylogenetic inference. To understand the role of adaptive convergence, we need to understand the patterns of nonadaptive convergence. Here, we consider the relationship between nonadaptive convergence and divergence in mitochondrial and model proteins. Surprisingly, nonadaptive convergence is much more common than expected in closely related organisms, falling off as organisms diverge. The extent of the convergent drop-off in mitochondrial proteins is well predicted by epistatic or coevolutionary effects in our "evolutionary Stokes shift" models and poorly predicted by conventional evolutionary models. Convergence probabilities decrease dramatically if the ancestral amino acids of branches being compared have diverged, but also drop slowly over evolutionary time even if the ancestral amino acids have not substituted. Convergence probabilities drop-off rapidly for quickly evolving sites, but much more slowly for slowly evolving sites. Furthermore, once sites have diverged their convergence probabilities are extremely low and indistinguishable from convergence levels at randomized sites. These results indicate that we cannot assume that excessive convergence early on is necessarily adaptive. This new understanding should help us to better discriminate adaptive from nonadaptive convergence and develop more relevant evolutionary models with improved validity for phylogenetic inference.

In Chapter V, I propose a novel and general method to determine how amino acid propensities shift over time and across sites. First I estimate the amino acid propensities at all sites for each ancestral branch in the phylogenetic tree, then I calculate the differences between the amino acid propensities at the ancestral and descendant sides of all branches. I hypothesize that substitutions cause larger shifts at adjacent sites, and search for substitutions at adjacent sites which may have caused the shift in amino acid propensities. I compare the distributions of shifts of amino acid propensities at sites with or without adjacent aamino acid substitutions to determine that substitutions cause larger propensity

shifts at adjacent sites.

In Chapter VI, I propose a novel method for exploring tree space in a phylogenetics context. This is usually achieved using progressive algorithms that propose and test small alterations in the current tree topology and branch lengths. Current programs search tree topology space using branch-swapping algorithms, but proposals do not discriminate well between swaps likely to succeed or fail. When applied to datasets with many taxa, the huge number of possible topologies slows these programs dramatically. To overcome this, we developed a statistical approach for proposal generation in Bayesian analysis and evaluated its applicability for the problem of searching phylogenetic tree space. The general idea of the approach, which we call "Markov Katana", is to make proposals based on a heuristic algorithm using bootstrapped subsets of the data. Such proposals induce an unintended sampling distribution that must be determined and removed to generate posterior estimates, but the cost of this extra step can in principle be small compared to the added value of more efficient parameter exploration in Markov chain Monte Carlo analyses.

Our prototype application uses the simple neighbor-joining distance heuristic on data subsets to propose new reasonably likely phylogenetic trees (including topologies and branch lengths). The evolutionary model used to generate distances in our prototype was far simpler than the more complex model used to evaluate the likelihood of phylogenies based on the full dataset. We demonstrate that this method can be used to efficiently estimate a Bayesian posterior.

This prototype implementation indicates that the Markov Katana approach could be easily incorporated into existing phylogenetic search programs and may prove a useful alternative in conjunction with existing methods. The general features of this statistical approach may also prove useful in disciplines other than phylogenetics.

I.7 Contributions

Chapter II is a published review of mechanistic protein evolution models. I contributed substantially in the writing and produced one of the two figures.

Chapter III is a published encyclopedia article I wrote with my professor for the Encyclopedia of Evolutionary Biology. I contributed early drafts of the article and worked on revisions with my professor.

Chapter IV is a published paper about convergence and models of protein evolution. I made many of the figures for the paper and contributed to writing the paper.

Chapter V is a potential paper proposing a new method of identifying changing amino acid propensities. I developed the model, analyzed the data, and wrote the paper.

Chapter VI is a paper in revision based on reviewer comments proposing a new method of sampling trees in a Bayesian context called Markov Katana. Early program development was done by Seena Shah with some contributions by Kenji Fukushima, however I improved the method substantially, produced the results, and wrote the paper.

CHAPTER II

MECHANISTIC MODELS OF PROTEIN EVOLUTION*

II.1 Abstract

Models of sequence evolution are used ubiquitously in biology from phylogenetic reconstruction to the analysis of adaptation, coevolution, and convergence. The structure of the model used affects these analyses, and it is therefore preferable to use good models. The field of molecular evolution is currently undergoing an important transformation due to large increases in the ability to collect and analyze massive amounts of data. Here we briefly review the history of molecular evolution and then discuss how evidence of epistasis and convergent molecular evolution helps overturn traditional models of protein evolution. We conclude by discussing desired features in a simple mechanistic model of protein evolution that is more compatible with patterns observed in real and simulated protein evolution.

II.2 Introduction

The field of molecular evolution is concerned with how molecules evolve, and the forces that determine their evolutionary path. For functional molecules such as proteins, RNA and regulatory DNA, the main factors include mutation, how neutral variants spread in a population, and the differential fitness of organisms containing variants. Because most individuals in most populations lived in the inaccessible past, the field is generally focused on using data from currently living (extant) populations to infer past processes, including gene duplications, population divergence (speciation), and phylogenetic relationships. Since we acquired the ability to sequence proteins and genetic material, the power of utilizing this rich evolutionary record has been demonstrated repeatedly in terms of understanding phylogenetic relationships, predicting the timing and order of species

*Portions of this chapter were previously published in *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*, 2017, and are included with the permission of the copyright holder. Authors include David D. Pollock, Stephen T. Pollard, Jonathan A. Shortt, and Richard A. Goldstein.

divergence events such as those among human/chimp/gorilla ancestors, as well as looking specifically at evolutionary processes that occur at this molecular level such as when we predict the functional importance of individual positions or regions of molecules [66].

These advances were achieved despite our lack of understanding of the mechanistic aspects of functional molecular evolution. We are currently in the midst of an important transformation of our mechanistic description of how functional molecules evolve, primarily through a better understanding of the importance of epistasis and coevolution and how to incorporate them into evolutionary models. This transformation will strongly impact our ability to resolve conflicts in species and gene phylogenies, predict function and adaptation of function, predict the timing of molecular and systems-level evolutionary events, and predict the functional effect of mutations. To understand the mechanistic view, we first briefly review here the history of how molecular evolution has usually been described, followed by discussion of the molecular evidence that directly contradicts many existing evolutionary models. This will be followed by an overview of recent theoretical advances, which demonstrate the opportunities for simplification and better modeling, despite the potential for epistatic interactions to introduce overwhelming complexity.

II.2.1 A Brief History of Molecular Evolution

Since the work of Mendel and Morgan, it has been clear that mutations give rise to variants that can affect higher-level phenotypes such as the wrinkled surface of peas or the white color of a fly's eyes. Variants were reasonably considered as subject to natural selection that would alter the expected evolutionary trajectory of these variants, eliminating variants that give rise to deleterious phenotypes (negative or purifying selection) and increasing the frequency of initially rare variants that give rise to beneficial phenotypes (positive selection). The subsequent development of mathematical descriptions of these processes, especially through the pioneering work of Wright, Fisher, and Haldane, gave rise to the field of population genetics. Although Wright in particular advocated a stochastic approach to population genetics, which considered the role of a finite or even small

population sizes, the simpler and more tractable deterministic approach assuming very large/infinite population sizes tended to dominate for decades during the middle of the 20th Century (reviewed in [67]). Although the deterministic approach allows standing variation in populations due to e.g., mutation/selection balance and overdominance, and transient variation during selective sweeps, this was essentially an adaptationist description of molecular evolution. The implicit assumption is that most traits are highly adapted if not perfectly optimized, and that positively selected changes, when they do rarely occur, are due to changes in an often vaguely described 'adaptive landscape', whose hyperdimensionality did not prevent it from being represented in a two-dimensional plot.

As the scientific community learned the nature of transcription, translation, the genetic code, and how to sequence proteins and DNA, it became clear that some nucleotide variants were unlikely to impact phenotypic traits as much as others. This was then largely incorporated into population genetics theory, especially through the work of Kimura and his 'neutral theory' of molecular evolution [68]. Once enough sequence data accumulated, it became clear that many variants, including synonymous and non-coding mutations, were also probably essentially neutral, and should be included in neutral theory. Because of this, a better appreciation for the importance of stochastic processes began to dominate towards the end of the 20th Century. Although neutral theory was clearly anti adaptationist in the sense that it was no longer viable to believe that *all* mutations were subject to meaningful levels of natural selection, in retrospect it was still highly adaptationist in the sense that the mutations that *mattered* for selection were all (or nearly all) considered to be deleterious. Most proteins were implicitly considered so optimized that the effect of a mutation, if it had an effect, must be deleterious. Despite the dominance of purely neutral theory during this time, theoreticians such as Gillespie and Ohta developed approaches that incorporated more variable degrees of selection, and laboratory biologists such as Powers and Watt evaluated potentially idiosyncratic systems in which protein variants appeared to be sustained (not fixed) due to varying selection

along gradients and overdominant selection [69, 70].

The generality and purity of neutral theory was primarily broken by a combination of events. More examples of variants that were maintained by selection were discovered, and variation at the Major Histocompatability Complex (MHC) played a big role in convincing many neutral evolution proponents that positive selection mattered (e.g. [71]). MHC was important both because it contains a great deal of long-term standing variation (trans-species polymorphism) that cannot be explained by neutral theory, and because it is a good example of ongoing selection due to constantly varying host-parasite interactions. Other examples of molecular cat-mouse chases involving protein-protein interactions arose in viral proteins, venom-prey, and male-female or mother-offspring conflicts (e.g. [72, 73]). All of these produce proteins with evolutionary histories of amino acid substitution rates greater than neutral expectations, what is called diversifying selection. Finally, with the pioneering work of Yang and others, it became possible to detect brief bursts of amino acid substitution greater than neutral expectation along ancestral branches in phylogenetic trees (e.g. [74]). These are generally interpreted as 'adaptive bursts', driven by changing selective requirements (sometimes identified and sometimes not). There are valid concerns about the statistical certainty with some of these approaches in some cases, but the overwhelming impression is that adaptive bursts are moderately common and identifiable within the vast diversity of life. For example, in our work with snake mitochondrial genomes we identified perhaps the largest known temporary adaptive burst in multiple proteins at the base of snake diversification [75]. This example provides evidence of adaptation, but also a large enough sample of substitutions enriched for adaptive change that we can characterize the differences in evolutionary patterns in adaptive and nearly neutral substitutions [70]. With the sequencing of the first two snake genomes, it has also held up as a general systems-level metabolic adaptive phenomenon.

Although Ohta's 'nearly neutral' theory combined with evidence for occasional bursts of adaptive change should give us a healthy respect for the fluctuating nature of molecular

evolution, we would argue that nothing discussed so far deviates too much from a modified adaptationist paradigm. Yes, not all variation affects functional adaptation, and yes, sometimes the meaning of 'adapted' changes, but overall these arguments are compatible with mostly constant adaptive pressure at each amino acid position. Missing are explanations for observed epistasis and coevolutionary interactions among amino acid residues, related observations of heterotachy (changes in evolutionary rates over time), and why substitution rates among amino acids differ among positions in proteins. To begin finding explanations, we can move to three-dimensional and experimental considerations, and determine that particular amino acid substitutions have particular effects on stability (changes in the free energy of folding, as measured by $\Delta\Delta G$) or function (e.g., ligand binding or measurable enzymatic parameters). However, experimental results are expected to be low resolution compared to the sensitivity of evolution and the effects measured in a laboratory may be different than what is important to selection. Experimental results are therefore considered to be generally informative but not definitive, and need to be interpreted with care. Furthermore, there are strong practical limits to the amount of data that can be collected. Computational predictions have questionable utility [76, 77, 78], with their limited accuracy of $\Delta\Delta G$ prediction that further decreases with multiple substitutions, and binding strength predictions are even more limited. In any case, case-specific measurements do not amount to a general theory of how evolution proceeds [77]. Knowledge progresses on all fronts, but we focus here on our multi-pronged approach, which involves empirical statistical modeling of sequence evolution in the context of phylogenetics, simulation of protein evolution as a thermodynamic system to better understand non-intuitive aspects of how functional molecules evolve, and the continued development and application of theory analogous to statistical mechanics to understand the mechanics of functional molecule evolution.

II.3 Modeling Principles and Empirical Statistical Models of Molecular Sequence Evolution

II.3.1 Empirical Statistical Modeling and Phylogenetics

Because we are advocating for more mechanistic models of protein evolution, it is useful at this point to discuss how empirical statistical models of molecular evolution are compared, what we mean by 'mechanistic' models, and how mechanistic models differ from phenomenological models. The statistical models that we discuss are those used to analyze sequence data, and the fundamental calculation in these models is to determine the probability that the data would have been produced if the model had been operating with particular parameter settings. In a frequentist approach, one compares models by finding the parameter combination that is most likely to have produced the data, while a Bayesian approach compares models by integrating the posterior probability over reasonably likely parameter settings, and simultaneously incorporating prior probabilities of models and parameter settings. Good reviews of this topic can be found elsewhere [79, 80].

Empirical statistical models can differ substantially in their theoretical foundations. Here we emphasize the difference between phenomenology and mechanism-based models. We define pure phenomenology as simple, theory-free measurement, such as might be done to count the number of individuals in a population of organisms, or to measure the height of a person. In the context of molecular evolution, an example of a mostly phenomenological approach might be to measure the frequencies of amino acids at each position in a sequence alignment, or in each protein, or in an entire set of proteins. Other alignment-based measures such as the fraction of sites that are unvarying or correlations between amino acids observed at different sites might also be considered primarily phenomenological. The statistical questions for a purely phenomenological measurement are mostly limited to reproducibility, accuracy, and perhaps how the quantity changes over time. These measurements are a good start, but are of limited utility unless

we are able to interpret their meaning, significance, and range of applicability. We usually would prefer to understand what the site-specific amino acid frequencies can tell us about the protein, and its relationship with other proteins, or use the divergence between sequences to estimate evolutionary distances.

An alternative to pure phenomenology is to represent the salient aspects of the process that resulted in the current sequences using a model that embodies some theoretical mechanism. Choosing which aspects to include and how they are represented generally involves a mixture of empirical (phenomenological) observations and (mechanistic) representations of the underlying biology. For example, modern substitution matrices are usually estimated using a phylogenetic tree, which can be considered a mechanistic model for how the species diverged during the course of evolution, part of the process by which the sequences were produced. Another example is the analysis of the DNA sequences that code for proteins using the Genetic Code, which constitutes a mechanistic model of how DNA is converted to protein. The most popular rate models for representing protein evolution include the empirical observations that some amino acids are more common than others, some substitutions are more common than others, and some sites change more frequently than others. These are represented phenomenologically with a set of amino acid equilibrium frequencies, a symmetric 'exchangeabilities' matrix, and a (generally Gamma distributed) distribution of rates. Models used for identifying positive selection include the mechanistic consideration that DNA substitutions in protein-coding regions can be synonymous or non-synonymous, but often ignore the observation that some amino acid changes are more likely than others. In both these cases, the probability of substitution from one amino acid or from one nucleotide to another is simply inferred from the number of sequence differences, and thus is a phenomenological component of the model.

There can be borderline cases as well, where the empirical results can be justified from the underlying biology; for instance, the difference between transition rates (between

purines A and G or pyrimidines C and T) and transversion rates (between purines and pyrimidines) can be rationalized by considering the chemical structure of DNA. There are also numerous instances where these phenomenological representations are used to gain mechanistic insights, such as in inferences of positive selection or in the analysis of substitution matrices to determine physicochemical protein properties [81, 82]. There are, conversely, always observations and biological knowledge that is ignored by these models. Many of these simplifications were required due to our lack of knowledge of molecular evolution and the limits of computational resources and sequence data availability at the time in which the models were constructed. In other instances, the phenomenological representations are in conflict with basic molecular biophysics, or are internally inconsistent. For instance, the site-specific rates of amino acid substitutions reflect the degree of selection acting on that site, resulting in a restriction in the amino acids that are appropriate for that site (generally not modelled), which causes the reduced substitution rate (which is modelled). It is, in general, impossible to reconcile the empirical amino acid equilibrium frequencies in these models with the observed overall substitution rate [83].

Historically, the basis of empirical statistical models used in molecular evolutionary analysis has nearly always been that there are a certain number of states (e.g., nucleotides, amino acids, or codons) with constant substitution rates of exchange among them. These substitution rates might be different for different classes of sites or different genes or genomic regions, and occasionally have been allowed to change at discrete points on the phylogenetic tree. Substitution probabilities, $P(t)$, along branches of length t in the phylogenetic tree were then usually (and often still are) calculated first by spectral decomposition to obtain the eigenvalues (Λ) and eigenvectors (S) of the instantaneous rate matrix (Q), and then by calculating $P(t) = Se^{\Lambda t}S^{-1}$. The implicit mathematically necessary but rarely discussed assumption in these approaches is that Q holds over long periods of time. Thus, even if the average Q is well estimated, it may not be an accurate

reflection of the process at any single point in time. This is a problem for a number of reasons, chief among them being that the parameters of the instantaneous rate matrix identified will depend on the particular phylogenetic tree considered, and how long the branches on that tree are, and that this averaged rate matrix may not be accurate for any site at any time during the evolutionary process, obscuring the actual nature of the evolutionary change. This problem was to some extent recognized early on when it was found that PAM matrices determined using many closely related proteins produced very different results than BLOSUM matrices determined with more distantly related proteins [84, 85, 86].

In the last decade or so, more and more Bayesian approaches have incorporated augmented data methods that allow one to avoid time-consuming and computationally expensive spectral decomposition and repeated matrix-vector-matrix multiplication to obtain the substitution probabilities along branches, $P(t)$. Focusing on our own method encoded in the program *PLEX*, we partially sample substitutions to the nearest short branch region to augment the data; in combination with uniformization [65, 87] of substitution rates this can be much faster than complete augmentation of fully specified substitution histories. This program was designed to allow greater flexibility in allowing substitution probabilities to differ among positions in a molecule and over time, but this can create an explosion of complexity, or at least an explosion in the number of adjustable model parameters, and the question is how to develop appropriate models that reflect the underlying biology. Models that include rate heterogeneity (e.g., [88, 89, 90, 23, 91], for example, are limited by the amount of sequence necessary to estimate parameters. They also treat each site in a protein as independent from all others without considering the protein molecule as a whole. Thus, the consequences of selection are modeled but the mechanism of selective action is still treated as an unknown. For reasons that will become clear below, we do not think that continuing to divide sites into more and more small substitution categories is a fruitful or mechanistically-justified approach.

II.3.2 Epistasis and coevolution

The concept of epistasis, that the effect of variants in combination is not always an additive sum of their individual effects, is well known from the early days of genetics and biochemistry. From a biochemical perspective, function and three-dimensional structure arise from interactions among amino acid residues, and if one residue in a protein changes, it is natural to presume that it may alter the effect of a change at another position. Experimentally, epistasis is easily detected by finding mutants in a protein that are deleterious to function, and then selecting for 'compensatory' mutants that allow the protein to recover [92]. Early mutagenesis studies in lysozyme also clearly established the prevalence of compensatory relationships among amino acids in protein cores [93]. As with the adaptationist/neutralist arguments of the last century, however, questions arise about how often epistatic changes occur during evolution, and how important the role of positive selection is in preserving these changes. The observation of rampant epistasis among amino acids in proteins [94] promotes a more nuanced view on protein evolution and the substitution process, a view in which the probability of substitution at each site is dependent on which amino acids occupy nearby and other interacting sites. A few evolutionary concepts have been extremely important to understanding the role of epistasis in functional molecular evolution: molecular coevolution, deep evolutionary inference, and convergence. The concepts of coevolution and epistasis overlap [95], but here we will view coevolution as the long-term evolutionary consequences of epistatic effects, and define it (as in [42]) as what occurs when a substitution at one position alters the propensity to accept substitutions at other positions. It is worth noting, however, that with epistatic changes (such as the biochemist's compensatory changes), it is usually considered that at least one change has a phenotypic or selective effect, whereas coevolution can proceed even if every substitution involved is entirely neutral. Past coevolution between individual residue positions is difficult to prove, especially for small or moderately diverse sequence datasets, but the cumulative evidence across many residues that coevolution is pervasive,

and evidence that there is a strong relationship between coevolution and structural proximity is overwhelming (e.g., [42]). It has also been noted that residues that are pathogenic in humans are surprisingly often the most frequent residues in related species [74]. Furthermore, even those trying to downplay the role of epistasis and fluctuating amino acid propensities in protein evolution have tended to produce data that confirm it, reducing the argument to questions of the size of the fluctuations under different conditions [96, 36]. Finally, recent papers have demonstrated the ability to filter coevolutionary information in very large and diverse bacterial phylogenies to find sufficiently adjacent amino acid residues that they can be used to predict protein structure [97, 98, 99].

The problems for simple evolutionary mechanistic theory that arise from deep (ancient) protein evolutionary inferences were recently reviewed [83], and we refer readers to that paper for details. However, the basic problem is that mutation rates are sufficiently large that neutral substitutions should have saturated individual positions, such that multiple substitutions will have thoroughly obscured the utility of neutral substitutions for inferring deep phylogenetic relationships that we are often interested in (such as mammalian divergence or deeper). Functional molecules such as proteins don't appear to saturate so quickly though (partly reflected in the differences between PAM and BLOSUM matrices, described above), and molecular phylogenetic analyses have long relied on such molecules to resolve ancient phylogenetic questions. Although we agree that this observation does not prove epistasis [100] as claimed by Kondrashov [101], it seems likely that the level of coevolution among amino acids that has already been demonstrated is sufficient to cause this effect.

II.3.3 The importance of convergence

Convergence "occurs when two biological traits in two separate lineages independently evolve to similar end points" [95]. It has long played an important role in evolutionary theory at the organismal level because convergent evolution of similar complex morphologies is seen as a strong sign of adaptation to similar selective forces in the environment [102,

103]. Convergence at the molecular level has been seen as a relatively rare phenomenon, but the huge increase in genomic data and dense taxonomic sampling in recent years has led to an upswing of papers detecting molecular convergence. These efforts have seen a number of false starts, however, beginning with mitochondrial genomes [104] and continuing with convergence in echolocating mammals [105]. An obstacle is that current evolutionary models do a poor job at predicting levels of convergence [106], but further problems arise when indirect methods of detecting convergence are used, and detection of convergence can be conflated with phylogenetic errors [106, 107, 108].

Molecular convergence is also beginning to play a big role in understanding epistasis and the mechanics of nearly neutral molecular evolution, and that is because of its relationship to propensity and constraint [48]. To see this, consider evolution at three sites, all with resident amino acid alanine (A), as shown in Figure II.1. Considering only the six amino acids shown (resident alanine and five possible substitutions), at site 1 the amino acids have equal propensity, and nearly equal (1 out of 5) probabilities of convergence (modified only by differences in mutation rates, particularly transition and transversion rates, as shown). At site 2, however, the propensities for S and T (and the resident, A) are much higher than for P, V, and G, meaning that substitutions at the site are almost completely constrained to S and T. If these propensities don't change over time, then the probability of convergence along two different evolutionary lineages given substitutions along both lineages at this site (both with ancestral state A) is nearly 50% (modified by relative mutation rates) because there are only two practical choices of substitutions. Site 3 has the same relative distribution of propensities among the amino acid alternatives to the resident amino acid as site 2 does, but the resident amino acid is far more fit. Thus, site 2 and site 3 will have the same probability of convergence if there are two substitutions at that site at different lineages in a phylogenetic tree, but site 3 is much less likely to substitute, and thus to converge, at all.

Understanding convergence as a biological consequence of constraint allows us to

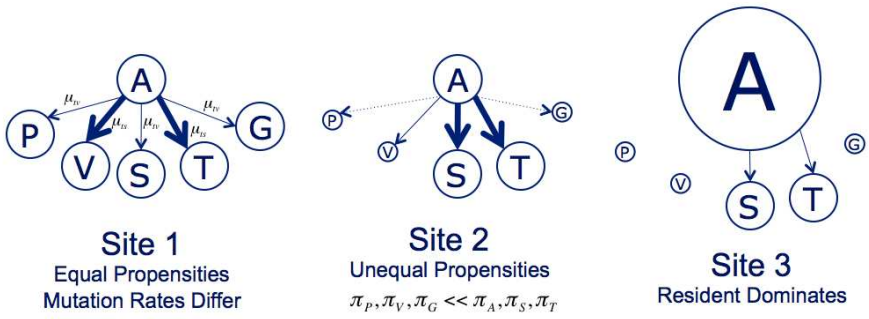


Figure II.1: Convergence depends on constraint. In the examples, at Site 1 the amino acids shown (A, alanine; P, proline; V, valine; S, serine; T, tyrosine; and G, glycine) have equal propensities (illustrated as size of circles), so evolution is unconstrained and substitution (indicated by thickness of arrows) is determined by the mutation rate for each type of mutation (transition, μ_{ts}; or transversion μ_{tv}) required to change the codon from alanine. At Site 2 the greater propensities of serine and tyrosine mean most substitutions will be to one of these two amino acids, and the remaining substitution rates are reduced to thin or dashed lines. At site 3, there are few substitutions due to the overwhelmingly large propensity of the resident amino acid, alanine; substitutions to serine and tyrosine are reduced, and other substitutions are so rare as to be essentially absent.

better understand why current evolutionary models do such a poor job at predicting convergence levels. Firstly, when protein positions with different levels of constraint are combined, the combined average model is often less constrained than any of the individual models. Consider one position that can substitute from alanine to serine or tyrosine, and another that can substitute from the same starting point, alanine, to proline and valine. If substitution probabilities between these sites are combined, one might expect that they both could substitute to any of the four amino acids, reducing the expected probability of convergence by half (25% expected probability rather than the actual 50%). A similar logic applies if the process of substitution changes at a single position at two very distant time points. Applying the previous example, the actual probability of convergence is near 50%. The position can substitute from alanine to serine or tyrosine over a *short time separation*, but if the position has switched to only accepting proline and valine at some *distant point* on the phylogenetic tree, then the probability of convergence would have fallen to zero. Thus one can understand that epistasis and coevolution, which by definition

alter substitution probabilities, have the necessary effect of reducing the probability of convergence over time [48]. If the evolutionary process differs among positions and over time, which appears to be the case [48], static evolutionary models would appear to have almost no hope of predicting levels of convergence, although just as a stopped clock may correctly predict the time of day, they may occasionally do so by chance. It is also worth noting that because neutral convergence is equivalent to homoplasy, and inferring levels of homoplasy is one of the main points of evolutionary models in phylogenetics [106], this result has implications for the reliability of phylogenetic inference using functional molecules, although the extent of the problem is currently uncertain.

As an aside, it should be noted that because evolutionary models do such a poor job of predicting levels of convergence, they probably cannot be used for this purpose with any degree of reliability. The problem is particularly insidious because commonly used standard models of amino acid substitution (such as JTT [109] or WAG [110]; see [83] review) are so broad and unconstrained that they actually don't change predicted convergence levels much over time, even with different ancestral amino acids and the inclusion of the genetic code [48]. Thus, a user would be highly confident of their results even when they shouldn't be. In contrast, moderately constrained but time-invariant models such as CAT models [39] or Halpern-Bruno models [88] interact strongly with the genetic code, and predict that global convergence levels will decrease over time even though the process at each site is not changing. For this reason, great care is needed to distinguish the effect of lowered convergence levels due to previous divergence and the structure of the genetic code; the trivial convergence caused by prior substitution is probably the strongest signal in any protein dataset [48], and does not demonstrate epistasis and fluctuating constraint. We therefore do not think that convergence predictions based on incorrect models and branch lengths, as in [111, 112], are reliable. Instead, we recommend that branch pairwise convergence levels should be compared to branch pairwise double divergence levels, and both only for cases of a common ancestral amino acid [106, 48,

113, 112]. Such convergence/divergence measures are also not subject to error due to fluctuation of average branch lengths among genes or gene regions. Because convergent molecular evolution may occur in response to both adaptive and non-adaptive causes, it is critical that we obtain a better understanding and use good means to predict non-adaptive convergence, the better to detect adaptive convergence when it does appear. Understanding the difference between adaptive convergence and non-adaptive convergence requires a better understanding of the evolutionary forces that govern the substitution process and the variability in site-specific constraints over time.

II.3.4 The evolutionary Stokes shift and the role of thermodynamic models

Sensitive readers may at this point be slightly concerned because if, as seems to be the case, amino acid propensities and therefore substitution rates and convergence levels fluctuate due to pervasive epistasis, then it would seem that there is little predictability to molecular evolution, especially if these fluctuations are random. However, the fluctuations are not actually random in the sense that they are undirected or unconstrained. A clear sign of this is the evolutionary Stokes shift [36, 35]. The basic idea of the evolutionary Stokes shift is that when an amino acid is substituted at a site, proteins tend to equilibrate to the newly resident amino acid through epistatic substitutions at other sites [35]. At the same time, other non-resident amino acids, including the previously resident amino acid, are not necessarily stabilized, and may wander away from or into stability states comparable to the resident amino acid [35], thus affecting the probability of substitution. The two components of the evolutionary Stokes shift can be termed 'contingency' (the necessary wandering of an amino acid in stability space to a similar stability level as the resident amino acid) and 'entrenchment' (the tendency of epistatic changes to stabilize the newly resident amino acid), and it has been shown that mutations that fix are contingent on previous substitutions [114].

The evolutionary Stokes shift was originally discovered as a consequence of modeling the evolution of functional proteins as thermodynamic, folded entities [35]. Surprisingly,

even a quite simple energy function, in conjunction with the need to be stable in a particular fold and not spend much time in other folds, can produce patterns of contingency and entrenchment [35]. Indeed, modifications of the model and inclusion of functional effects directly (for example, through ligand binding) do not seem to strongly affect the general result [114](Goldstein unpublished data), and the expected decrease in reversion rates after substitution has been shown to generalize to arbitrary fitness landscapes [100]. The direction of predicted stabilities between pairwise differences in diverged real proteins that had been crystallized showed remarkable agreement with our thermodynamic model proteins [35], and even skeptics have tended to produce measured stability data for substitutions in divergent proteins that are in rough agreement with theoretical predictions [96, 115, 36, 35], although the number of protein measurements is necessarily small.

Because modeling the evolution of functional proteins as thermodynamic, folded entities appears to reproduce many important features of protein evolution that are not explained by static models [48, 35], it is worthwhile to consider further what these models are doing and how we are using them. In Figure II.2, it can be seen that while the frequencies in the WAG model are distributed relatively evenly and are constant, the thermodynamic models (sometimes called Stokes-Fisher models) produce highly variable frequencies at a single position, over time changing the relative magnitudes and often the order or amino acid propensities. Major differences also occur among positions [35]. These differences occur despite the fact that the underlying amino acid interaction model that drives stabilities, a 20x20 interaction matrix, is no more complicated than the 20x20 WAG substitution matrix. The difference lies in the mechanism, which includes the requirement that the propensities and substitution rates are caused by the effect of substitutions or potential substitutions on the stability of the entire protein sequence.

The use of a simplified thermodynamic models to simulate evolution can reproduce some of the most perplexing features of protein evolution (epistasis, coevolution, the evolutionary Stokes shift, and changing nearly neutral convergence over time), and may

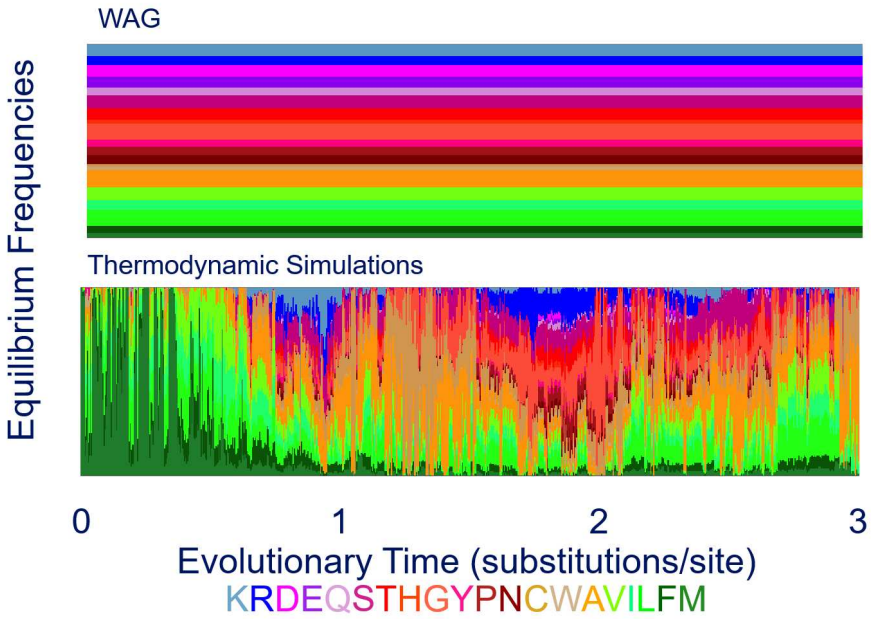


Figure II.2: Changes in equilibrium amino acid frequencies (propensities) over time. Results for the WAG model (top) and in thermodynamic simulations (bottom). WAG frequencies stay constant over time, while in the thermodynamic model constraints and equilibrium frequencies vary greatly.

indicate that the thermodynamic three-dimensional folded nature of functional molecules has an impact on their evolution. It also helps to illustrate the utility of developing a more detailed mechanistic approach. Simple mechanisms or processes can produce complicated-seeming results that are nearly impossible to sort out from a purely empirical perspective, but simplicity can be revealed and predictive power greatly improved by focusing on understanding the mechanism that produced these results. Such a scenario seems to be the case with molecular evolution; without a mechanism for how substitution rates are generated, we are faced with trying to find rate matrices for each site, and then for shorter and shorter periods of time, until the point where there is no more data to collect and resolution is still lacking. For example, Figure II.3 shows an example where the substitution rate between threonine and alanine appears to change across the vertebrate mitochondrial tree. The threonine to alanine rate (and the rate of reversion) seems much higher in birds than it is in mammals, and very different amino acid propensities would be predicted if the range of taxa were birds, versus birds plus crocodiles, turtles and

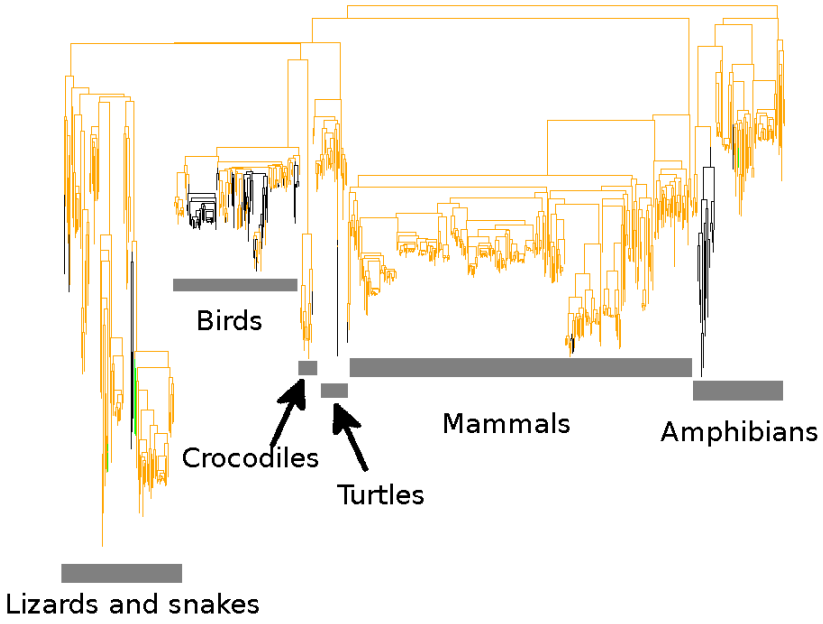


Figure II.3: Substitutions along a site in cytochrome c oxidase from tetrapod mitochondria. Branches are colored by orange=threonine, black=alanine, green=asparagine. Alanine is produced from threonine by a first codon position transition mutation, while asparagine is produced by a second codon position transversion mutation.

mammals, or among all the (tetrapod) vertebrates.

With a mechanism, however, it is possible that data collection can be focused on understanding the simpler question of how amino acids interact, which may be informative across all sites. To be clear, we are not saying that our model demonstrates that there is a single distribution of interactions at all sites; our model runs on a single distribution and reproduces many salient features of real protein evolution, indicating that careful work will need to be done to see if different context-dependent interaction models are truly needed. We view the thermodynamic models as more of a null hypothesis indicating the complexity that can be produced through thermodynamic evolution alone; to demonstrate a strong effect of context-dependence, a truly site-specific effect on the mechanism, one now needs more than just to show that their average rates, observed over a finite time, are significantly different.

Another key feature of our thermodynamic model is that at its base is a Hamiltonian-Potts model, i.e., an energy model. However, because it is a selected energy model, the amino acid propensities and interactions are not directly inferable from the energy function, and vice versa, as would be the case, for example, in inferring molecular conformation distributions. In the next section, we focus on explaining recent work towards understanding the theoretical dynamics of this situation, and how such theory can be used to inform on relative substitution rates and the strength of the evolutionary Stokes shift [116].

II.4 Towards a Statistical Mechanics Theory of Molecular Sequence Evolution

II.4.1 Introduction to the statistical mechanics of evolution

Up to this point, we have discussed mostly the evidence from statistical empirical models and from thermodynamic simulation models that jointly point to the idea that there is something about thermodynamics that may explain important features of molecular evolution that are incompatible with current models. In this section we consider the utility of a statistical mechanics framework for this explanation, following our recent work developing this theory [116]. The topic is difficult conceptually because we need to simultaneously include terms from classical statistical mechanics, thermodynamics, and transition state theory to discuss the folding of molecules and their ability to act as catalysts, but also develop terminology to discuss the application of statistical mechanics and transition state theory to the evolution of the sequences that code for these same proteins. For example, for this reason we discuss the stability of sequence X, as $\Phi(X)$, which is defined to be in the same direction as fitness (increases in stability correspond to increases in fitness), and is simply the negative of $\Delta G_{folding}(X)$, the free energy change of sequence X upon folding to a structure that carries out a function. This allows a smooth transition in discussion as to how the results may extend to other fitness functions, including ligand binding, catalysis, and signal propagation.

To separate out the structural component that underlies nearly all protein function, we consider that the probability that the protein is folded at thermodynamic equilibrium is equivalent to fitness [117, 35, 60]. This is partly based on our experience that when fitness is incorporated into thermodynamic models, proteins will crystalize into a single folded structure that tends to be marginally stable, as do real proteins [118]. Because we want to understand how substitution rates at a site come about, we can focus theoretical attention on a single representative site, k, and consider how substitutions at this site alter overall protein stability, and on this basis whether they will be accepted during the course of evolution. This is an entirely reasonable proposition because we have defined fitness to be determined by protein stability. Indeed, we unsurprisingly find in our simulations that substitutions between two amino acids can be extremely well predicted by the distributions of relative stability contributions made by each amino acid, and using Kimura's formula to predict substitution probability from effective population size (N_e) and relative fitness.

It is useful to pause here a moment and consider what this may indicate about real proteins. The distributions of contributions to stability for amino acids in real proteins are unlikely to be exactly what we get in our simulations because it may depend on the target structure and the true interaction energies between these amino acids, and may be modified by other functional constraints. We see wide variation in the distributions depending on which amino acids are involved and the average rate of substitution at a site, however, and it seems reasonable that these are factors in real proteins as well. Although we do not know the magnitude of these fluctuations or the rate at which individual sites in real proteins move in stability space, the fluctuation of stability contributions observed in simulations matches the observation of fluctuations in stability seen in real proteins, and explains the observed decrease in convergence probabilities with time of divergence [48].

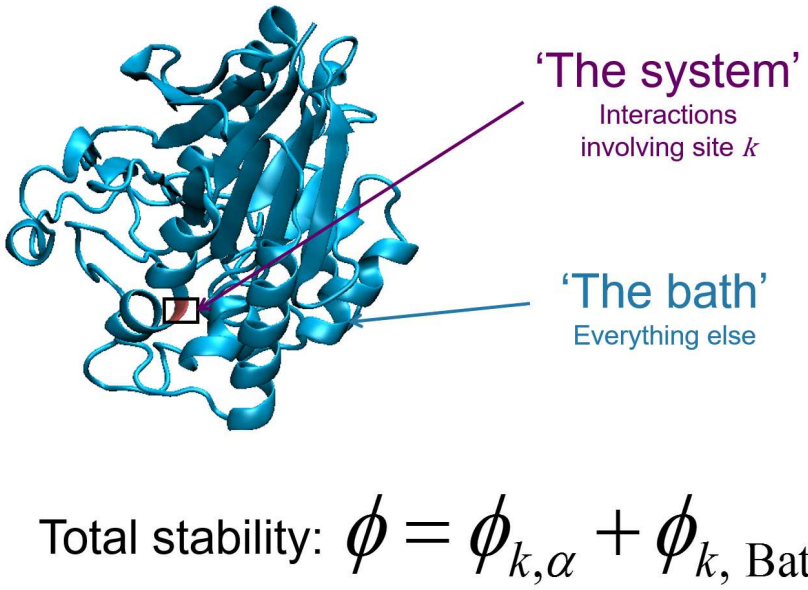


Figure II.4: Total stability. The total stability is divided into the contribution from the amino acid at a site k and stability contributions due to interactions among amino acids not including site k.

II.4.2 Can we obtain a mechanistic entropic explanation for the magnitude of the evolutionary Stokes shift and how it explains substitution rates?

The substitution rate between two amino acids depends on the amount of variation in the stability of the resident amino acid and how much covariation there is between the stability contributions of possible replacements. Ongoing work [116] centers around the idea that we can convert sequence space into a statistical mechanics framework by considering, in addition to the stability contribution of an individual site, the stability contribution of the remaining interactions not involving the site, the latter corresponding to the 'Bath' in analogy to classical statistical mechanics. Both of these sum up to the total stability: $\phi = \phi_{k.\alpha} + \phi_{k.Bath}$ (Figure II.4).

The mechanistic process can then be visualized first by considering the forces of sequence entropy (the number of sequences, Ω) and selection (depending on N_e and other factors), which conspire to tightly constrain total stability (ϕ, Figure II.5). There are no selective constraints on the relative proportion of $\phi_{k.\alpha}$ and $\phi_{k.Bath}$ that sum up to ϕ,

however, and so for that proportion entropy alone dominates. Because there are so many more interactions involving the bath contribution to stability, it tends to move towards lower stability values that have larger number of sequences (Figure II.5). It is only able to do this, however, if the individual site contribution to stability increases to compensate and keep the total approximately constant.

II.5 Conclusion

We have described here the role of mechanisms and phenomenological descriptions as components of statistical empirical models, and described recent developments in mechanistic descriptions of the evolution of functional molecules, such as proteins. The role of fast thermodynamic evolutionary simulations is pivotal in discerning how proteins, as thermodynamic entities, should evolve, and what sorts of effects thermodynamics have on evolutionary outcomes. These thermodynamic models provide a potential explanation for patterns of epistasis, coevolution, average substitution rate differences over long periods of time, molecular convergence changes over time, and the evolutionary Stokes shift, which are fundamental problems for current statistical empirical models. We believe that a statistical mechanic-like treatment of protein sequence evolution points to a mechanistic explanation for many, if not all, of these phenomena, with the added benefit that it may greatly reduce the number of phenomenological parameters needed for future statistical empirical models of evolution.

II.6 Acknowledgments

We acknowledge the support of the Medical Research Council (UK) (MC_U117573805) to RAG and the National Institutes of Health

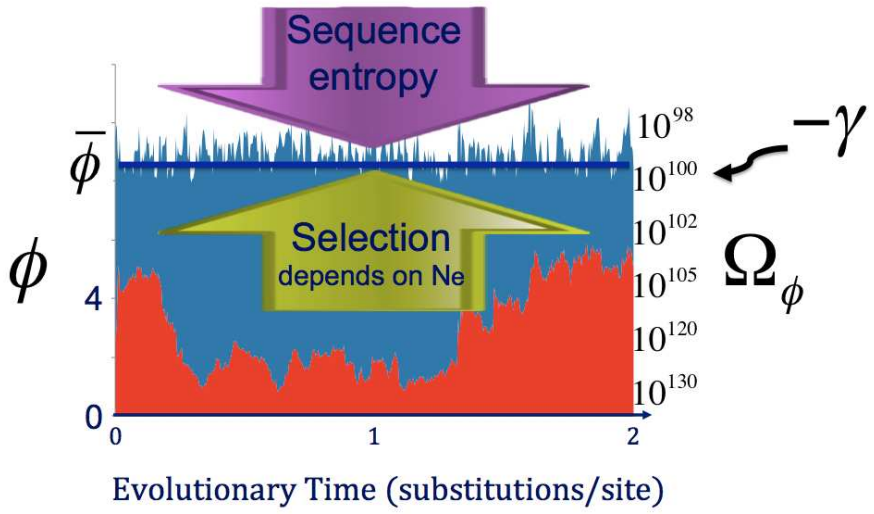


Figure II.5: Constraints on stability. The portion of the total stability occupied by the site-specific interactions and the remaining bath interaction. The constraints on stability due to entropy and selection are indicated. Bath and site-specific interactions are shown in blue and red, respectively. Plausible example numbers of sequences at each stability value (to the left, in kcal/mol) are shown to the right.

CHAPTER III

PARALLEL AND CONVERGENT MOLECULAR EVOLUTION*

III.1 Glossary

Adaptation It refers to long-term evolutionary modification and maintenance of functional traits or molecules in an organism in response to natural selection. Adaptation does not necessarily imply optimization because a trait may have constraints in how much it can be modified, and also because it is sufficient to improve a trait until it is "good enough".

Ancestral reconstruction It is the process by which ancestral traits or sequences that no longer exist in living organisms are inferred. In recent decades, ancestral protein molecules have been reconstructed and expressed in the laboratory to test their function and obtain more direct inference of how functions have evolved over time. Because of the multiplicative accumulation of uncertainty in ancestral reconstruction among all sites in an alignment, because of coevolution or epistatic interactions among sites, and because of laboratory experimental uncertainty, such inferences cannot be assumed to be exact reconstructions of sequence and function, but are often presumed to be strong indicators of trends in functional evolution.

Branch pairs They are two branches (sometimes called edges) on a phylogenetic tree that are compared as paired lineage segments for which the amount of convergent evolution can be measured.

Convergent evolution It occurs when two biological traits in two separate lineages independently evolve to similar end points. At the molecular level, a key goal is to distinguish between convergent events that have arisen because of adaptation, and possibly neutral convergent events that occurred by random drift and were thus not driven by positive natural selection.

*Portions of this chapter were previously published in *Encyclopedia of Evolutionary Biology*, 2016, and are included with the permission of the copyright holder. Authors include David D. Pollock and Stephen T. Pollard.

Divergence It refers to the process of traits or sequences becoming more different between lineages over evolutionary time by the accumulation of differences in each lineage that are not convergently replicated in the other.

Evolutionary distance It refers to the expected number of substitutions that occurred between species or along branches of a phylogenetic tree. Evolutionary distance in neutral evolution is the product of the mutation rate and time, and thus its expectation is proportional to time if the mutation rate is constant, although the outcome of actual number of substitutions along a lineage is subject to stochastic variation.

Homology It refers to traits in an organism that are descended from a common ancestor. In molecular evolution, in gene families that have incurred duplications, homologous genes are divided into **orthologs**, which are only related by speciation events, and **paralogs**, which are related by gene duplication events.

Homoplasy It is a cladistic term that refers to traits that cannot be resolved as having occurred just once on a phylogenetic tree, and thus must have arisen by convergence. Such traits cannot be resolved parsimoniously with other traits on a phylogenetic tree, and thus confound phylogenetic inference using cladistics methods.

Long branch attraction It is the phenomenon by which the accumulation of non-adaptive convergent events, or homoplasies, which is predictably greater between longer branch pairs than between shorter branch pairs, will lead to false phylogenetic signal that tends to "attract" or falsely join long branches if the true phylogenetic signal is not sufficiently strong.

Multiple sequence alignment It is the process by which positions in different sequences are ordered with respect to each other such that all corresponding sites in the alignment represent inferred homologous positions.

Neutral evolution It is evolution that occurs by random drift, meaning that the probability of fixation of new alleles is proportional to their starting frequency. At sites that evolve neutrally without selective constraint (and without certain interfering mutation

repair processes), the substitution rate is expected to be equal to the mutation rate.

Parallel evolution It is sometimes used as a synonym for convergent evolution and homoplasy, or may refer to convergent evolution that occurs in experimental replicates, or in closely related species under similar ecological pressures to adapt. At the molecular level, it may sometimes be used to indicate convergent evolution with the same or similar molecular mechanism, or involving the same genes, and at the amino acid level it may refer to convergent evolution from the same ancestral amino acid. In this article we advocate that the term should be deprecated.

Phylogenetic inference It is the process by which evolutionary relationships, or branching order and timing, among species or sequences, is inferred.

Position In a functional molecule such as a protein may refer to the amino acid location along the sequence and in three-dimensional space. Positions in difference sequences may be related to each other as inferred homologous sites in a sequence alignment.

Sites In an alignment of amino acids or nucleotides refer to aligned positions that are inferred to be homologous, and thus related by substitutions alone rather than insertion or deletion events.

Sequences It refers to the ordered components of biological molecules such as DNA (made up of four different deoxyribonucleic acids), RNA (made up of four different ribonucleic acids), or proteins (made up of twenty different amino acids).

III.2 Abstract

Convergence is a central concept for phylogenetic inference because it can occur when two lineages respond in the same way to similar adaptive pressure. Convergence may thus serve as a signal of adaptive response, and when it occurs it provides information on the replicability of particular adaptive responses. Comparative genomics approaches have begun to allow the analysis of convergence at all levels from the phenotype to physiological systems to proteins and other functional macromolecules, thus allowing researchers to dissect the molecular mechanisms that give rise to phenotypic convergence. In turn, this

allows study of the replicability of convergence at all levels, from repeated modifications of the same genes to repeated modifications of the same protein positions to the same amino acids. It is particularly important to understand how species divergence leads to altered constraints affecting convergence at the molecular and various systems levels. These altered constraints may affect the probabilities of adaptive as well as non-adaptive convergence. A major focus of research on molecular convergence concerns model-building and empirical techniques to distinguish ubiquitous but variable levels of non-adaptive convergence from more rare but interesting adaptive convergence events. "Parallelism" is often used to discretely categorize different types of convergent events, but its use is controversial and often contradictory among different authors. Because of this, we suggest that its use be discontinued in favor of focusing on the various evolutionary issues it is intended to embody.

III.3 Introduction

The idea of convergence is central to the study of evolution because it addresses the key concepts of adaptation and replicability. Roughly, convergence happens when two biological traits in two separate lineages independently evolve to similar end points. The concept can cover a broad range of evolutionary events, from convergence of function and morphology to gene duplication, expression levels, and sequence. At the organismal scale, convergence is usually assumed to involve adaptation in response to similar selective pressures from the environment, but at the molecular scale this is not always so. There is a great deal of excitement to the current study of molecular evolutionary convergence because in the age of genomics, large amounts of information are becoming available that can be used to elucidate the molecular mechanisms of phenotypic convergence. Researchers can now ask detailed questions about whether the convergence of aquatic mammals to life in the sea [119], echolocation in bats and dolphins [105, 120], or song in different groups of birds [121] involve similar genes expressed in similar locations with similar amino acid changes.

III.4 Integrating molecular convergence from molecules to phenotypes

To understand and dissect the mechanisms of convergence, it is necessary to consider how phenotypic convergence of organisms is achieved across multiple organizational levels, from molecules to phenotypes. Adaptive pressures act on organisms as a whole, and the response will be integrated from amino acid changes that alter functional aspects of proteins, regulatory and transcriptional changes that alter when and where molecules are expressed, and interactions among system components. We discuss here just a few of the many examples available in the literature, particularly pointing out a couple of noteworthy recent genome-wide convergence analyses.

Molecular convergence of amino acids in proteins was first observed in lysozyme as an adaptation to expression in the acidic environment of the stomach as a digestive enzyme in cows (ruminants) and colobine monkeys [122]. This was later augmented by observation of amino acid convergence in RNAses in the same environments [123, 124]. A recent example is that of convergence in the molecule prestin, involved in adaptation to echolocation in dolphins and bats [125, 126]. The example of adaptive convergence at positions involved in modulating proton transport in cytochrome C oxidase between snakes and agamid lizards [106] is the largest and densest (per amino acid position) example of convergence that we are currently aware of, and is large enough to allow characterization of the types of amino acids and positions involved in convergence (Figure III.1). The adaptive burst in the ancestor of all snakes is also notable for having generated follow-on convergent events in the subsequent phylogeny of snakes, some of which are also convergent with amino acid changes in the ancestor of another tubular legless squamate, *Rhineura* [75]. These events provide examples of convergence to the same amino acid at the same position, physicochemical convergence to amino acids with the same functional group (hydroxyl, or -OH) at the same site, and convergence to add positively charged amino acids in a tightly clustered region at the base of a proton channel (some at the same amino acid position and some at different positions).

Convergence on the genome level has been observed in laboratory experiments involving phi bacteriophage, E. coli in high temperatures, and yeast under nitrogen starvation [127, 128, 129]. A general observation under the conditions of these studies (involving selection on a large number of possible genes) is that although adaptive changes were observed repeatedly in the same genes (that is, convergent usage of genes), it is rare to observe convergence of the same amino acid at the same position in the same gene because there are so many alternative adaptive responses. Foote and colleagues [119] considered convergent evolution in the much larger mammalian genomes that adapted to a marine environment. Echoing the viral and microbial results, they showed that while there was widespread convergence between three different marine mammals at the amino acid substitution level, they were unable to separate what might have been adaptive from the noise of what is expected with such a large number of statistical comparisons. Further analysis of genes showing signs of positive selection in multiple marine mammal lineages detected convergent changes plausibly related to adaptation to a marine environment, but the point remains that the proportion of convergent changes is small, and many adaptive responses are not convergent in replicates under both experimental and natural conditions.

Molecular adaptation is also thought to occur rapidly through regulatory change. Zhang and colleagues assessed expression levels of genes associated with song learning in birds and found that increased expression levels of certain genes were often convergent in the song-learning nuclei [121]. In a somewhat more complex example, convergent evolution of the transcription factor SP1 in mammals and birds (to different amino acids, but at the same position) appears to have convergently altered its structure and binding specificity. This then caused convergent evolution at hundreds of SP1-regulated binding sites in both mammals and birds [130]. This system further demonstrated follow-on convergence in regulatory networks, as there were later convergent amino acid substitutions in homologous SP transcription factors in various mammal and bird lineages that co-regulate some of

the same regulatory modules by binding to SP1 binding sites. Notably, these convergent changes in the co-regulatory paralogs were at the same sites as the convergent changes in SP1, and to the same amino acid as in chicken SP1.

III.5 A conceptual understanding of convergence and parallelism

Despite its importance, and despite its deceptively simple definition, the exact meaning and characterization of convergence is not always clear. This echoes the lack of clarity surrounding the term "adaptation" itself, which can be difficult to distinguish from selection and can be notoriously difficult to prove. The uncertainty over what is and is not convergence is further confounded by the common use of the term "parallel evolution", which is sometimes used as a synonym of convergence or in different ways that can appear mutually exclusive to one another. The intended meaning of the authors is not always clear. At the organismal level, parallel evolution embodies the idea that closely related organisms might respond to similar selective pressures in similar ways. As discussed in a comprehensive review by Arendt and Reznick [131], the idea of parallel evolution thus sometimes applies exclusively to phylogenetic considerations (how closely related the two species are), exclusively to mechanistic considerations (whether the pathways used to effect the phenotypic change are the same), or some mixture of the two. These authors argue strongly and convincingly for the general use of "convergence" and deprecation of the use of "parallel evolution", a usage also followed in the comprehensive review by Christin, Weinreich, and Besnard [132] and at the MapOfLife website (mapoflife.org). At the molecular level, the use of parallel evolution has sometimes been focused down even to the trivial level of the individual amino acid, adding another usage to distinguish the cases where the common ancestor is the same (parallel) or different (convergent) [133]. Using the same data to argue nearly the opposite of Arendt and Reznick, Rosenblum and colleagues [134] advocate for using the term "convergence" only for the phenotypic level, and "parallelism" only for the molecular level.

The clearest message from these discussions in the literature is that the difference

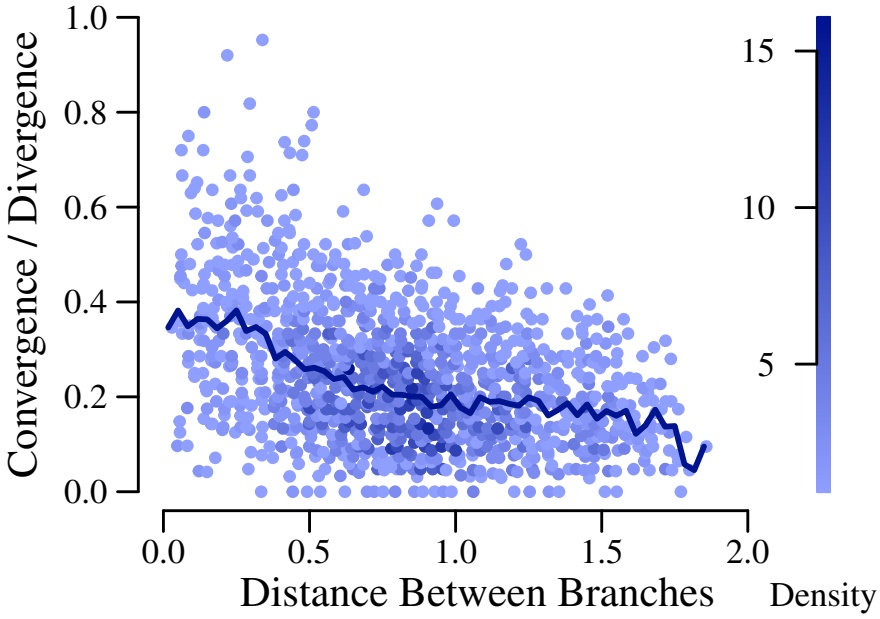


Figure III.1: The effect of convergence on phylogenetic tree reconstruction. This figure is modified from Castoe et al., 2009. This figure shows that the mitochondrial data links the acrodont lizards (blue box) in a sister relationship with the snakes (orange box), although most nuclear and morphological evidence suggests that they should be grouped with iguanid lizards (green box), as shown by the red arrow. Detailed analysis showed that the mitochondrial phylogeny was driven by convergent amino acid substitutions at otherwise usually conserved sites. When the 500 codons that most support the mitochondrial tree are removed, the remaining mitochondrial data suggests the same tree as the nuclear data. A similar result was seen if the top 5% most convergent amino acid sites are removed.

between parallel and convergent evolution is unfortunately muddled, highly controversial, and unlikely to be settled soon. From a practical standpoint, this means that a reader interested in the subject will need to search for both terms, and will need to read carefully to try to understand the usage intended by each individual author, to the extent that it is clear.

In this article, we will follow the usage of Arendt and Reznick, using "convergence" as the general term and advocating that the use of the controversial term "parallelism" should be deprecated. This is strongly motivated in part by our own studies of convergent evolution in mitochondrial genomes [106] and on coevolution and epistasis [35], and in part

because it makes writing on the topic simpler and more clear. In considering convergent amino acid changes in proteins, for example, we are most concerned with which convergent amino acids caused a convergent structural or functional change, and not whether the convergent events on each lineage arose from a common ancestor. Because of coevolution and epistasis, the effect of an amino acid replacement may be dependent on the entire protein, and the effect cannot be assumed to be constant over time. Furthermore, given the variation in the evolutionary process across positions in a protein, it is clear that sorting sites by whether or not they have a common ancestor has the perhaps unintended consequence of biasing the groups depending upon rates of evolution at each position. The rate at which sites evolve affects their tendency to contribute to adaptive convergence because of the well-known relationship between function and conservation: the most functionally important positions are usually most conserved and also most likely to be utilized to effect adaptive change [75].

The position we take on usage should not be construed to imply that the distinctions that authors are trying to make when they use the term "parallel evolution" are unimportant. On the contrary, they involve extremely important questions about how adaptation works, what matters in terms of the changes that occur during an adaptive event, and the degree to which adaptive pathways are replicable. It is clear that the answers to these questions are tied up in the degree of constraint in the adaptive potential of the biological system, and how those constraints change over time. What functional modifications are theoretically accessible to a system, or to an individual protein, at any point in evolutionary time? How does the realm of accessible modifications change over time? Do slightly different adaptive pressures applied to an organism result in similar or very different molecular evolutionary responses? What amount of functional change in a protein or a regulatory system as it evolves over time will lead to different evolutionary responses to similar adaptive pressures? It is our belief that rather than to say an adaptive event "is parallel" as opposed to convergent, it is better to be clear on

precisely what questions are being asked and answered about the degree of taxonomic similarity between the lineages in question and the degree of functional similarity between any of the molecular components.

III.6 Discriminating adaptive and non-adaptive molecular convergence

Perhaps the greatest problem in the study of molecular convergence is how to discriminate adaptive molecular convergence from non-adaptive molecular convergence. There have been a number of claims of large-scale adaptive molecular convergence over the last decade [105, 104] that have later been thoroughly refuted [106, 107, 111]. The primary reason for these mistaken claims is that commonly used evolutionary models do not adequately predict expected levels of convergence under neutrality [106], although in some cases use of indirect tests of convergence and failure to apply proper comparative controls may also have played a role [107, 111]. It has been thought that more realistic models that vary across sites may be adequate to predict convergence levels [135], but it has been recently shown that the situation is complicated and may require models that vary over time as well as across sites [48]. Unfortunately, such complex models are not well specified with existing data sets, and we can expect intense future analysis and development of approaches to determine the best way to predict molecular convergence using models of amino acid substitution.

In the absence of good predictive models, an alternative approach is to use empirical observation. Such an approach is based on the idea that we expect adaptive molecular convergence between lineages (or branches on a phylogenetic tree) to be rare or moderately rare, whereas we expect that non-adaptive convergence is everywhere. Thus, in principle we can infer the amounts of convergence levels among branch pairs and then determine which branch pairs have excessive convergence relative to the main distribution. However, one additional factor to be considered is that we expect the amount of convergence to somehow be scaled by the total amount of evolution on each of the two branches being compared. To see this, consider that to be convergent, a pair of substitutions must occur

at the same site on both branches, and the probability of substitution at each site and branch will be proportional to the total amount of substitution along that branch.

An example of such an empirical approach was developed and implemented by Castoe and colleagues in their analysis of adaptive convergence between ancestral snakes and agamid lizards [106]. This approach avoids the direct dependence of convergence on branch lengths by making use of a discovered strong general correlation between convergence (abbreviated C) and paired divergence (abbreviated D), defined as substitutions in two separate lineages that independently evolve to different end points (as opposed to the same end point in convergent events). In this analysis, we found that predicting convergence from branch lengths was much worse than prediction from paired divergence events, which may be related to inaccuracies in the evolutionary model used to predict the mitochondrial phylogenetic tree as well as variation in rates among sites and over time, which may have made the average branch lengths inapplicable to the convergence predictions for the sites involved.

The idea that the ratio C/D of paired convergence and divergence events should be constant and would thus arise naturally from the data makes some degree of intuitive sense. This is because given that paired substitutions at the same site on two different branches have occurred, all else being equal the probability that such events are convergent or divergent should be equal among branch pairs. However, it should be considered that all else is not necessarily equal, in that the ancestral amino acids may have changed over evolutionary time, the composition of the sites that make up the paired substitutions may have changed, and the evolutionary process at some or all sites may have changed. We recently found that non-adaptive amino acid convergence rates do decrease over evolutionary time of separation between branch pairs, with dependence on the specific ancestors involved, the rate of substitution at the sites involved, and a fluctuating evolutionary process due to epistasis or coevolution [48]. Thus, these are important interacting effects, and it is a major challenge to incorporate them adequately in future

predictions of adaptive convergence.

III.7 Molecular Convergence, Ancestral Reconstruction and Phylogenetic Inference

Phenotypic and molecular convergence can both in principle be inferred through the observation that distantly related groups of species contain similar traits, while many other species, more closely related to each of the convergent species than they are to each other, do not. However, as we begin to understand how molecular convergence events have led to phenotypic convergence, we generally need to pinpoint when the convergent events occurred with as much accuracy as possible. In other words, we need phylogenetic trees with diverse representation of relevant species to break up critical branches on the trees as much as possible, and then we need to infer on which branches the putative convergent events have occurred. It is useful to do this to understand the timing and possible environmental correlates, but it is essential to do this to limit the large numbers of neutral or unrelated substitutions that are bound to occur on long unbroken branches. To make these inferences, one must also perform ancestral reconstruction to estimate the state of each position at nodes (branching points) of the phylogenetic tree, a process that has error and which may be biased at the sequence as well as the functional level [136, 60].

Although phylogenies are necessary to dissect the mechanisms of molecular convergence, both adaptive and non-adaptive convergence may interfere with phylogenetic inference. The example of Castoe and colleagues [106] makes it clear how adaptive convergence can positively mislead phylogenetic reconstruction by systematically and falsely linking distant branches. This can be understood by seeing that if there are a lot of convergent events, the false signal at the sites involved may overwhelm the true phylogenetic signal at the more reliable neutral and well-behaved (or at least not systematically biased) sites. It may then be more parsimonious to resolve the changes on a tree that would make the convergent sites appear to have substituted only once. This problem is made worse by

the tendency of adaptive convergence to occur at functionally important sites that are otherwise conserved, and thus often thought to constitute a more reliable signal than variable sites. Although we expect adaptive and convergent events of the magnitude and density of the snake example to be rare, smaller events are more difficult to detect, may be more common, and if they occur between closely related and difficult-to-resolve branch pairs they may easily overwhelm the true phylogenetic signal.

Non-adaptive and possibly neutral convergence should have less obviously deleterious effect on phylogenetic inference because it is widespread and not focused on particular lineages. However, if the substitution models used in phylogenetic inference do not adequately predict levels of convergence (often called homoplasy in this context because the effect is worst for cladistic methods such as parsimony that do not have an explicit model), then there will be bias towards falsely joining long branches. Thus, the observation that most current models inadequately predict non-adaptive convergence levels is of considerable concern. Furthermore, the recent discovery of exceedingly high convergence levels in closely related lineages, with a decrease in these levels over time, is also of obvious concern [48]. Because this is also expected to be unbiased towards any particular branch pair, it appears that the greatest concern is that it adds considerable noise to inference of the most difficult phylogenetic branching problems, and that certainty and confidence in some phylogenetic results may be poorly understood.

III.8 Conclusion

Although there are open questions in how to best detect molecular convergence and how to improve our models to understand it better, genome-wide studies as well as focused biochemical studies have begun to reveal general mechanisms of how molecules effect phenotypic convergence. These are exciting times for the study of molecular convergence. Broad generalizations are beginning to emerge, but big questions remain about when and why adaptation to environmental conditions repeats itself in different organisms with different levels of relatedness. Elucidating the details promises to keep evolutionary

biologists occupied for many decades to come.

Selected Further Reading

Arendt, J. & Reznick, D., 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends in ecology & evolution, 23(1), pp.26–32.

Castoe, T. a et al., 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proceedings of the National Academy of Sciences, 106(22), pp.8986–91.

Christin, P.-A., Weinreich, D.M. & Besnard, G., 2010. Causes and evolutionary significance of genetic convergence. *Trends in genetics : TIG*, 26(9), pp.400–5.

Foote, A.D. et al., 2015. Convergent evolution of the genomes of marine mammals. *Nature genetics*.

Goldstein, R.A. et al., 2015. Non-Adaptive Amino Acid Convergence Rates Decrease Over Time. Molecular biology and evolution.

Thomas, G.W.C. & Hahn, M.W., 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. Molecular biology and evolution, pp.1–49.

Weinreich, D.M. et al., 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science (New York, N.Y.), 312(2004), pp.111–114.

Zhang, G. et al., 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, 346(6215), pp.1311–1321.

List of Relevant Websites

MapOfLife (mapoflife.org)

CHAPTER IV

NONADAPTIVE AMINO ACID CONVERGENCE RATES DECREASE OVER TIME*

IV.1 Abstract

Convergence is a central concept in evolutionary studies because it provides strong evidence for adaptation. It also provides information about the nature of the fitness landscape and the repeatability of evolution, and can mislead phylogenetic inference. To understand the role of adaptive convergence, we need to understand the patterns of nonadaptive convergence. Here, we consider the relationship between nonadaptive convergence and divergence in mitochondrial and model proteins. Surprisingly, nonadaptive convergence is much more common than expected in closely related organisms, falling off as organisms diverge. The extent of the convergent drop-off in mitochondrial proteins is well predicted by epistatic or coevolutionary effects in our "evolutionary Stokes shift" models and poorly predicted by conventional evolutionary models. Convergence probabilities decrease dramatically if the ancestral amino acids of branches being compared have diverged, but also drop slowly over evolutionary time even if the ancestral amino acids have not substituted. Convergence probabilities drop-off rapidly for quickly evolving sites, but much more slowly for slowly evolving sites. Furthermore, once sites have diverged their convergence probabilities are extremely low and indistinguishable from convergence levels at randomized sites. These results indicate that we cannot assume that excessive convergence early on is necessarily adaptive. This new understanding should help us to better discriminate adaptive from nonadaptive convergence and develop more relevant evolutionary models with improved validity for phylogenetic inference.

*Portions of this chapter were previously published in Molecular Biology and Evolution, 2015, volume 32, issue 6, and are included with the permission of the copyright holder. Authors include Richard A. Goldstein, Stephen T. Pollard, Seena D. Shah, and David D. Pollock.

IV.2 Introduction

Although evolution mostly proceeds by accumulation of differences between groups, numerous examples of convergent evolution exist, where similar solutions are found to similar evolutionary problems. Well-known morphological examples include eyes and wings, but an increasing number of examples are known at the molecular level, including proteins involved in echolocation in bats and cetaceans [120, 137, 105], foregut fermentation proteins in monkeys and cows [122], transcription factors in mammals and birds [130], and mitochondrial proteins among different snakes [75], and mitochondrial proteins between snakes and agamid lizards [106].

Such convergence at the molecular level can both confound and inform evolutionary analyses. Convergent evolution can result in erroneous phylogenetic trees by showing strong support for incorrect topologies [106]. However, replicated evolution to the same trait or amino acid in different lineages provides convincing evidence of adaptation [75]. In addition, convergent evolution can provide important information about the adaptive landscape; the relationship among genotype, phenotype, and fitness; the constraints acting on evolutionary processes; and the role of chance and necessity in evolution.

Statistically meaningful analyses of adaptive convergence rely on estimates of the likelihood that such convergence could occur by chance in the absence of adaptation. Such analyses generally rely on standard models of evolution [104, 105], but it is now clear that these models are woefully inadequate, drastically underestimating the levels of nonadaptive convergence [106]. We need to improve our ability to predict the amount of expected nonadaptive convergence if we want to avoid errors in phylogenetic relationships, make accurate inferences of adaptive evolution, and investigate what convergence tells us about the fitness landscape and evolutionary process.

Two assumptions common to most evolutionary models are that evolutionary processes are homogeneous among sites in an alignment, and over time. It is becoming increasingly clear that both assumptions are unjustified. Different distributions of amino acids are

found in buried locations in the protein structure, exposed locations, tight turns, trans membrane helices, disordered regions, and locations of functional significance, indicating different selective constraints at these different types of locations. These differences are embodied in some mutation selection models [88, 138] and mixture models [37, 139, 23].

The evolutionary process at individual sites can also vary as a result of changes in structure, function, physiological role, or context of the corresponding location in the protein structure [29, 140, 49, 32, 35]. In addition, in the presence of epistatic or coevolutionary interactions between sites, the process at one site will change due to substitutions that occur at other coupled sites [35, 36]. There has been increasing evidence for the importance of epistatic interactions. For instance, Bloom et al. performed measurements on influenza proteins and observed that the effect of a substitution on the thermodynamic properties depended on the amino acids found in other positions [96, 36]. Pollock et al. demonstrated how amino acid propensities at a site will adjust over time after a substitution, such that the resident amino acid (and others with similar physicochemical properties) tends to be the most favorable amino acid at that site, an effect they termed an "evolutionary Stokes shift" [35]. As a result, the selective constraints at each site will shift to follow the changing occupant at that site.

The amount of amino acid variation in a protein can be decomposed into the variation allowed due to the site- and time-specific constraints, plus the effect of variation in those constraints among sites and over time. As a result, models that neglect variation in evolutionary constraints over sites and time tend to underestimate the magnitude of instantaneous selective constraints at individual sites, resulting in an underestimation of the expected amount of neutral convergence. In addition, temporal heterogeneity in selective constraints may induce time dependence to the neutral rate of convergence. We therefore set out to quantify the frequency of convergence in a data set of mitochondrial proteins and investigate changes in convergence patterns over time. We then compare these results with predictions from standard models, as well as simulated proteins evolving

under purifying selection for thermodynamic stability, similar to simulations used in Pollock et al. [35]. We then consider what the results indicate about the process of protein evolution.

IV.3 Results

IV.3.1 Convergence Decreases with Time in Vertebrate Mitochondrial Proteins

We examined convergence events occurring on distinct branches in a phylogenetic tree (supplementary fig. S1, Supplementary Material online) from a concatenated alignment of all 13 mitochondrial protein sequences from over 600 vertebrate mitochondrial genomes. A fixed amino acid substitution model mtMam [141] with site rate variation with five gamma distributed rates was used to infer the substitutions. When comparing the substitutions on two distinct branches, a pair of substitutions on each branch at the same site can be classified as either a convergence event (C) if the substitutions are to the same amino acid, or as a paired divergence event (D) if the substitutions resulted in different amino acids. Bayesian estimation was used to obtain the C and D totals for each branch pair considered. For short branches, C and D would both be roughly proportional to the product of the two branch lengths, suggesting that the branch length dependence could be minimized by considering the ratio of convergence and divergence events, C/D. This is supported by previous analyses showing that C and D are highly correlated and that D is a better predictor of C than branch lengths [106]. For display purposes, only substitutions with greater than 90% posterior probability were considered in calculating C/D for figures IV.1 – IV.4, although all significance and credible region estimates were obtained by integrating overall ancestral state uncertainty (see Methods and supplementary Methods, Supplementary Material online). Distances between branches were calculated as patristic distances along the phylogenetic tree, measured between the ancestral nodes on each branch. Note that we do not assess or make use of the state of the site in the more ancient common ancestor of both branches. Distances are given in units of expected number of

replacements per site.

The observed C/D ratios depend strongly on the distance between branches (fig. IV.1), a result that might seem surprising in the context of standard time- and site-homogeneous models of substitution, simply because if the model does not change one might think the C/D ratio would not change either (we elaborate further on these expectations below). The ratio is extremely high (0.4) for the shortest distances between branches, falling to below 0.2 for the most separated branches. The 99% credible regions for the expected C/D ratios over time are shown in supplementary fig. S2, Supplementary Material online, and they are nonoverlapping until later times when the ratios fall below 0.2 (supplementary table S1 Supplementary Material online). The variation in ratios among branch pairs is high, with ratios for short to medium branch distances (<0.5 replacements per site) ranging from zero to nearly one. Notably, this high variation mostly arises from biological variation in the expected ratio among branch pairs, not from poor estimation of ratios with few C or D counts (see model predictions below). There is also a strong dependence on whether or not the amino acid is different in the ancestral sequences of the two branches (fig. IV.2). When the ancestral amino acids are identical, the average ratio starts at about 0.45 and drops to 0.2, whereas when the ancestral amino acids are different the average ratio is approximately constant at 0.08.

These results strongly suggest that amino acid propensities and therefore substitution possibilities at each site are initially highly constrained. We can calculate an effective number of accessible residues by considering the size of the alphabet of states m that would result in a particular value of C/D if all substitutions were equally likely (i.e., a Jukes Cantor [JC] model; [20]). As shown in the supplementary Material, Supplementary Material online, if there is only a short evolutionary distance between the branches so that neither amino acid has changed, $C/D = \frac{1}{m-2}$. An initial C/D ratio of 0.45 therefore indicates an effective number of only 4.2 accessible residues per site. (These initial measurements are taken over the period of time represented by the length of the branches,

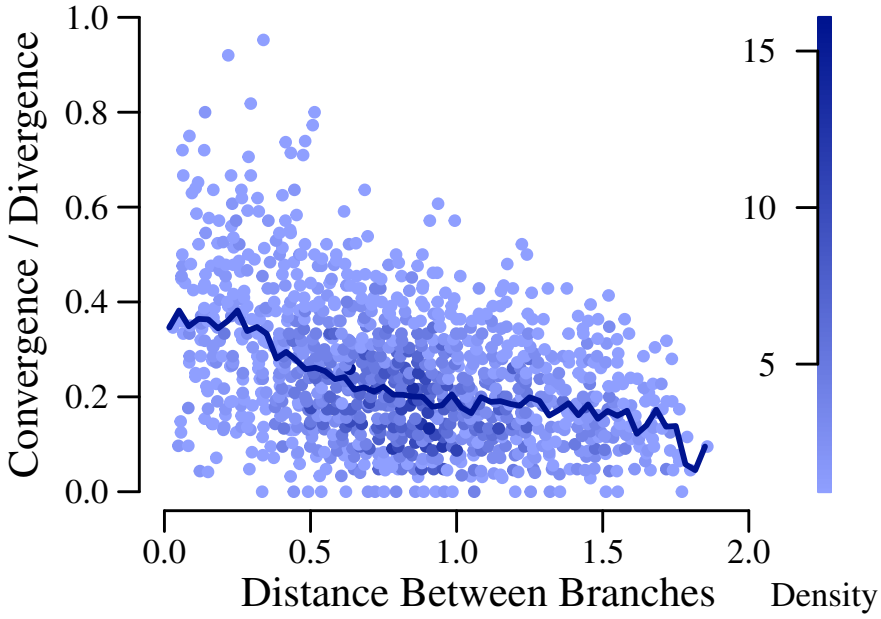


Figure IV.1: Change in convergence over time in mitochondrial proteins. The convergence over branch-paired divergence ratio (C/D) was estimated for all eligible pairs of branches in the mitochondrial phylogeny. In order to help visualization of the data, overlapping data points were merged into single points with the color determined by the density of dots merged, with blue intensity gradient as shown in the scale to the right. We used a threshold of $D \geq 20$ for inclusion in this graph. The distance between branches shown is the patristic distance between the ancestral nodes of each branch, measured in average number of amino acid replacements per site. The blue line shown is a running average with window size 0.03.

and given the falloff in the C/D ratio, the initial instantaneous ratio, prior to sequence divergence, may have been substantially higher). When the ancestral amino acids are known to differ, C/D is equal to $\frac{m-2}{m^2-3m+3}$, and therefore a ratio of 0.08 is equivalent to an effective number of 13.4 residues per site.

These results also strongly support the idea that amino acid constraints change over time, because the C/D ratios drop even when the ancestral amino acid is identical (fig. IV.2A). Although the changing convergence probability in the overall data set (fig. IV.1) can be understood by changing mixtures of sites with the same ancestral states (fig. IV.2A) and different ancestral states (fig. IV.2B), it does not appear possible to explain the drop in convergence seen in figure IV.2A based on changing site composition. If

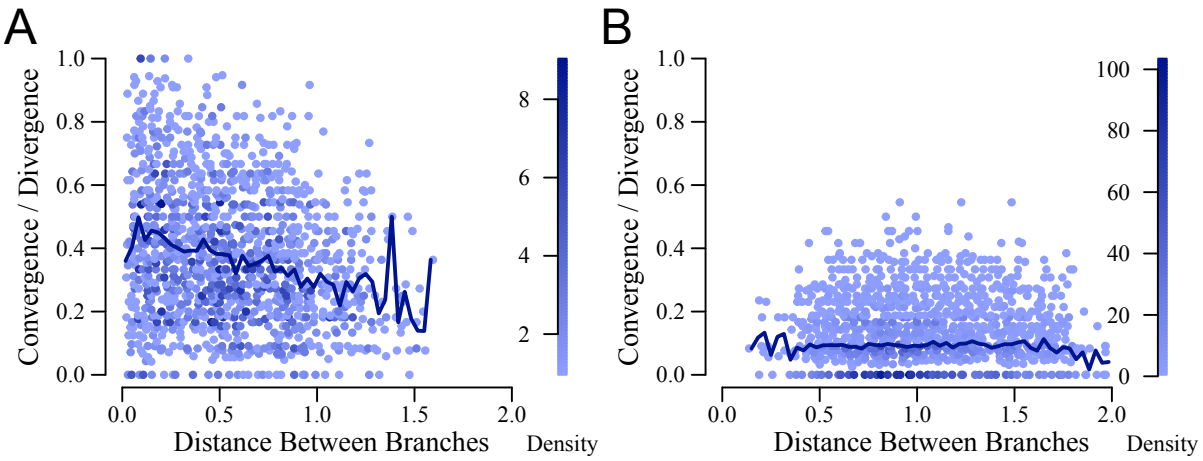


Figure IV.2: Mitochondrial protein convergence for identical and different ancestral amino acids. The convergence over paired divergence ratios were estimated, merged, and colored as described in Figure IV.1, except that events were separated into two categories depending on whether the ancestral amino acid at a site was the same (A) or different (B). We used a threshold of $D \geq 10$ for inclusion in these graphs.

anything, as discussed below, the bias in composition due to removal of evolved sites should remove low constraint (low convergence probability) sites, which would increasingly produce a bias for sites with higher convergence probabilities. The 99% credible region for the slope of a linear model fit to the data from figure IV.2B shows a clear decrease in C/D ratios with divergence (–0.109, –0.094; supplementary fig. S3, Supplementary Material online). We therefore conclude that the constraints are likely changing over time.

In contrast to the strong apparent initial constraint, once the amino acids at a site diverge, the number of amino acids acceptable at a site is quite high, drastically reducing the chance of convergence. (Recall that a C/D ratio of 0.08 corresponds to an effective number of 13.4 accessible residues per site.) To determine if sites retain information about convergence probabilities in the case of different ancestral amino acids, we resampled the substitutions among all sites, maintaining the same ancestral amino acids for each substitution. For example, if a branch has a substitution at site 5 from alanine to glycine, we collected all the substitutions from alanine on all branches and at all sites, then replaced the glycine with an amino acid randomly chosen from the descendant amino acids of the collected substitutions. The C/D was then recalculated for every branch pair

using these resampled substitutions. The results are shown in supplementary figure S4, Supplementary Material online. The C/D ratios for these resampled replacements are essentially the same as for the observed ratios (fig. IV.2B), indicating that, conditional on the different ancestral states, the sites provide no further detectable information about convergence probabilities.

To further understand these results, we partitioned the sites roughly evenly into three conservation classes. For recently diverged branch pairs, the average C/D ratio was highest for conserved sites, starting at about 0.6 and falling to below 0.2 (fig. IV.3A). In contrast, the average C/D ratio at variable sites was initially only slightly above 0.2 and fell quickly to near 0.1 (fig. IV.3B). As with the overall ratios, the fast- and slow- evolving sites may have started out with much higher C/D ratios, but the ratio dropped off too quickly to measure over finite branches. This would particularly affect the fast-evolving sites, and we cannot know for sure if the differences between figures IV.3A and B are due to an inherently higher convergence probability at more conserved sites or if they occur because highly variable sites reach equilibrium much faster. The results for identical and different ancestral states at each site for each conservation level (supplementary figs. S5 and S6, Supplementary Material online) are similar to the results for the complete data set (fig. IV.2), albeit noisier. It is worth noting, however, that the C/D ratio from identical amino acids at conserved sites (supplementary fig. S3A, Supplementary Material online) also falls off over time, indicating that the effect of fluctuating constraints over time on convergence probabilities is strong even for the most conserved sites.

IV.3.2 Relationship of Convergence with Time under Different Evolutionary Models

Given the convergence results for the mitochondrial data, we wanted to know the degree that these results are predicted by existing substitution models. We first simulated data along the mitochondrial tree under two different models and then we inferred the ancestral sequences and calculated the C/D ratios using the same method as with

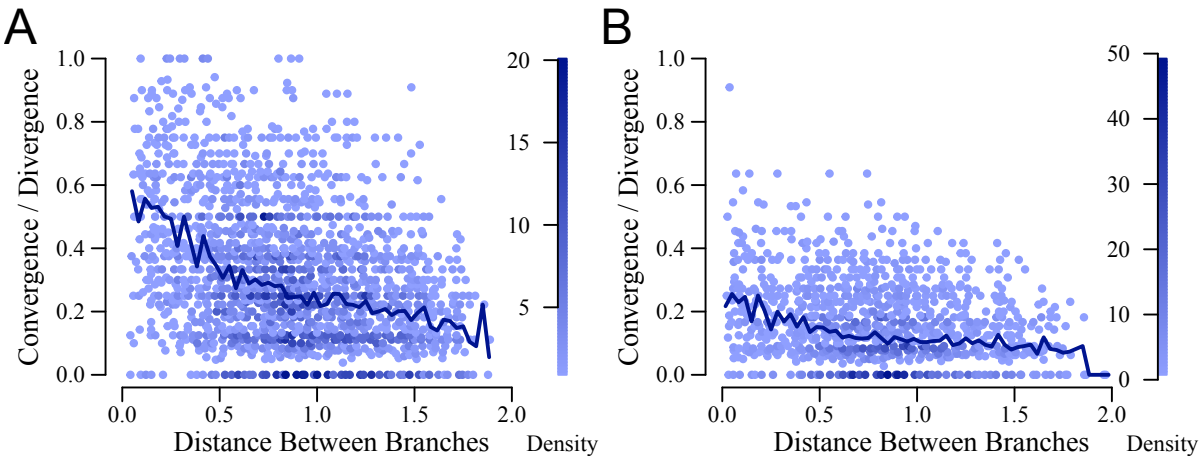


Figure IV.3: Mitochondrial protein convergence for conserved and variable sites. The data and visualization are the same as in Figure IV.1, except that ratios were estimated separately for conserved (A) and variable (B) sites. We used a threshold of $D \geq 7$ for A and $D \geq 10$ for B.

the mitochondrial data. The two models we used were the following: An amino acid substitution matrix (Whelan and Goldman model, WAG) [142], which neglects differences among sites and over time but accounts for differences in the rates of exchange among different amino acids; and the recent thermodynamic-based Stokes–Fisher (SF) model [35], which allows for coevolution (epistasis) among sites, and thus allows for different processes among sites and over time. We wish to avoid the possible impression that the SF models we use are designed to accurately reflect the true model of evolution. Instead, the SF model was constructed to generate semirealistic simulations that have many salient aspects of evolution similar to real proteins and can therefore guide us to better interpret observations on real protein evolution.

The "WAG" model results in a low and slightly decreasing C/D ratio that is not highly variable (fig. IV.4A). In contrast, the SF model results (fig. IV.4B) are a remarkably good match to the mitochondrial data results (fig. IV.4B insert). This indicates that the variance observed in the mitochondrial data is not just the result of estimation error, and the general shape of the curve is a fundamental expectation for evolution of complex functional molecules such as proteins, and is not specific to mitochondrial proteins.

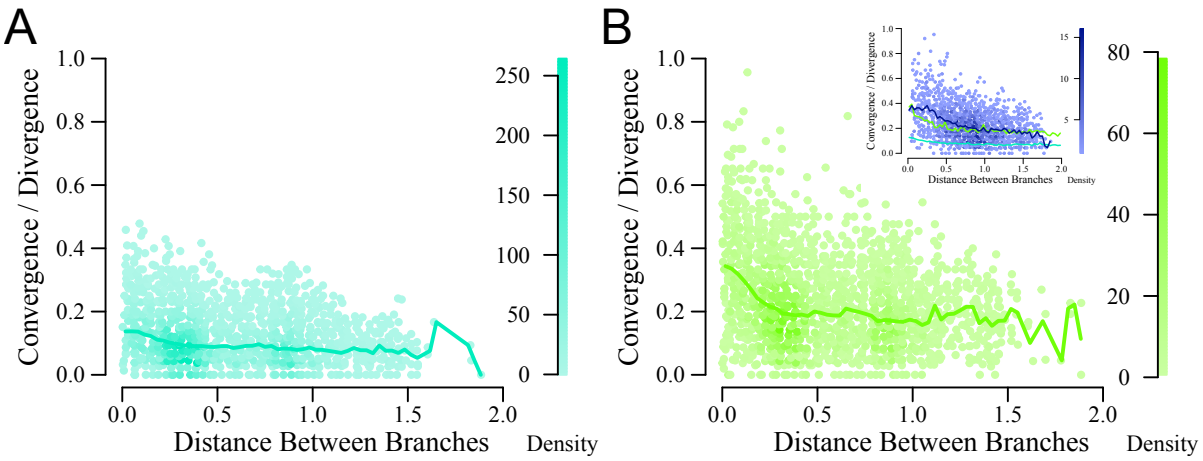


Figure IV.4: Convergence in simulated data. Protein evolution was simulated along the mitochondrial tree using the WAG substitution model (A) and Stokes-Fisher protein evolution model (B). C/D ratios were calculated using the same method as with the mitochondrial data (figs. IV.1 – IV.3) and were visualized the same as in figure IV.1. The inset in (B) shows the SF and WAG averages along with the mitochondrial data average (in blue, as before), for comparison. We used a threshold of $D \geq 20$ for A and $D \geq 20$ for inclusion in B

To further dissect the basis for the observed effect, we analyzed additional models of varying complexity, including the simple JC model (equal rates of amino acid exchange), a model with codon structure (Zihengian model, Z), WAG and Z with gamma-distributed rate parameters, and the CAT-60 model (CAT), which includes variation in constraint among sites [20, 22, 39]. These models, except the CAT-60 model and WAG, had their parameters fitted to data from SF simulations that used a star phylogeny, instead of the mitochondrial phylogeny. The form of these models allowed exact calculations of the expected mean C/D ratio, including the potentially higher initial instantaneous C/D ratio inaccessible to analyses on trees with finite branch lengths.

There is a decreasing C/D with time in all models of evolution (fig. IV.5), although it is barely perceptible for the JC model, and it is a relatively small effect for the WAG and Z models. For models with a constant site-specific process over time (all models except SF), the change in ratio is mostly attributable to the difference in the number of available convergent states depending on whether the ancestral state is the same or different. It is

interesting that the WAG and Z models both have small but slightly different responses to adding site specific rate variation, with WAG somewhat delaying its drop in convergence levels and the Z model accelerating the drop. We speculate that convergence levels in the WAG model are more dependent on slower exchanges, whereas in the Z model the drop in faster sites takes precedence.

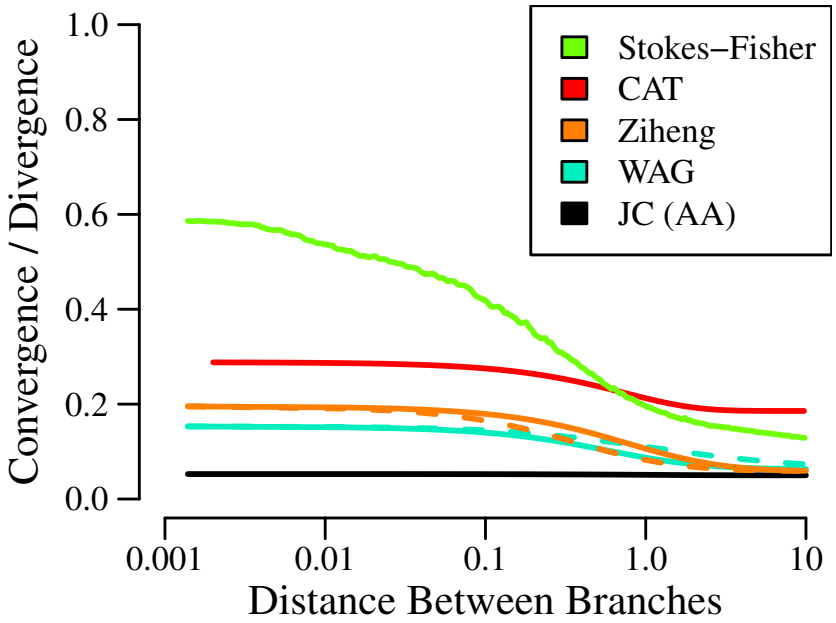


Figure IV.5: Convergence in simulated data under models of different complexity. The C/D ratios shown are from exact calculations on specific models. The different models shown (see legend for line colors) are JC, WAG, WAG plus gamma, Z, Z plus gamma, and SF. The results for WAG and Ziheng with gamma rate variation are shown with a dashed line. Note that unlike previous figures, the distances are on a log scale.

The results for that CAT model are especially notable when broken down into same and different ancestral states (fig. IV.6A). Although the C/D ratios for diverged ancestral states are somewhat higher than the equivalent results from the mitochondrial (fig. IV.2B) data, the truly notable observation is that the C/D ratio for sites with the same ancestral amino acid actually increases over time under the CAT model. In principle, if one observes C/D ratios for the same set of sites that change neither their ancestral amino acids nor their propensities over time (as in the CAT model), then their C/D ratio must remain

constant. However, the set of amino acids is changing in this case because the sites that evolve more rapidly are more likely to have differing ancestral amino acids (about 80% of sites by the most divergent timepoint; see fig. IV.6A inset). Unsurprisingly, the sites that change tend to have higher entropy than the sites that do not, and the sites with unchanged amino acids have less average entropy over time (fig. IV.6A inset). The increase in C/D ratios in the CAT model for sites with the same ancestral amino acid is thus explained by the lower average entropy (and thus greater constraint) over time at those sites that remain unchanged. This result is clearly exactly opposite the results from the mitochondrial data, in which sites with the same ancestral amino acids have clearly decreasing C/D ratios with time. Although it is possible to conceive the evolutionary models that are constant over time but still result in decreasing C/D ratios in this case (i.e. if low entropy sites all had extremely high mutation rates), such models would appear rather artificial and would have to overcome the naturally higher substitution rates of high entropy (low constraint) sites. It is much easier (and perhaps more natural) to explain these results with models that involve fluctuating constraints over time, of which the SF model is but one example.

Finally, we described above that sites with different levels of sequence conservation behaved differently in terms of their drop in C/D ratio. To understand this better, we separated out the instantaneous C/D ratio expectations for sites in the SF star phylogeny simulations corresponding to buried, partially buried, and exposed locations in the protein structure. From this we can see that indeed the buried sites start out with a higher C/D ratio of slightly over 0.8, and retain a higher ratio throughout the evolutionary simulation (fig. IV.6B). In contrast, the exposed sites start out with a lower ratio of about 0.5, and are always lower. This implies that the instantaneous site specific constraints under the SF model are highest at buried (more slowly evolving) sites and lowest at exposed sites. By analogy, this suggests that similar factors are at play in producing the real mitochondrial protein differences in C/D ratios between slow- and fast-evolving sites observed in figure

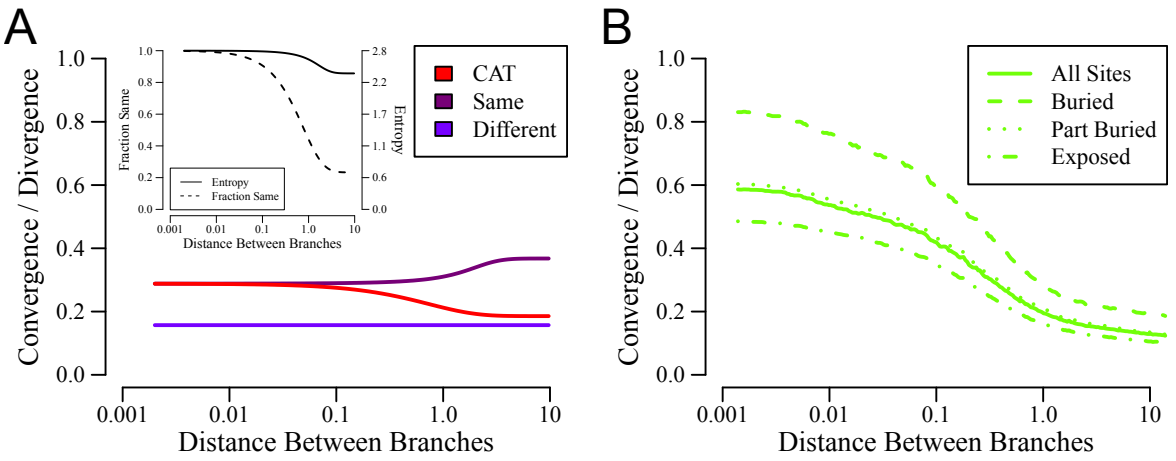


Figure IV.6: Convergence for the CAT model and the SF model with sites segregated by burial in structure. (A) The C/D ratios are shown for the overall CAT model (red), as well as C/D ratios depending on whether the ancestral states are the same (purple) or different (blue). The inset shows the fraction of sites that have the same ancestral state (dashed line) and the entropy averaged over all sites that have the same ancestral state (solid line). (B) The C/D ratios shown are the same as the SF runs in figure IV.5, but sites were determined to be buried, partially buried, or exposed based on fraction exposed surface area for the corresponding amino acid in the protein structure.

IV.6.

IV.4 Discussion

Current treatment of convergent events generally assumes that nonadaptive convergence at the molecular level is well predicted by simple time-averaged and site-averaged models. However, our analysis of real proteins and model-based simulations demonstrates that the rate of convergence changes over time, and can be extremely high for recently diverged proteins. The convergence data presented here provide additional evidence that our understanding of how proteins evolve needs to be fundamentally revised. The patterns of convergent evolution observed may cause difficulties for phylogenetic reconstructions, but can also provide important information about adaptation and adaptive bursts, as well as allowing us to investigate the underlying topology of the fitness landscape and the nature of the substitution process.

Convergence probability is closely related to the number of amino acids that are

acceptable at a given site at a given time. If a small hydrophobic amino acid is required, the probability that two acceptable substitutions in different lineages will result in the same small hydrophobic amino acid can be quite high. Constraints at another site requiring large flexible amino acids will result in a similarly high probability of convergence. If the substitution model is inferred by averaging over different sites, or the same site at different times, including instances where only small hydrophobic, or large flexible, or aromatic, or charged amino acids are required, the result is a model with few constraints that allows a wide variety of different amino acids. These simple models will overestimate the number of acceptable amino acid substitutions and underestimate the probability of convergence.

As indicated above, the high rate of convergence and the strong dependence of the convergence rate on evolutionary distance strongly suggest the importance of variation in the substitution rate across sites and over time. The idea of fluctuating amino acid substitution rates over time is an important feature of evolutionary Stokes-shift theory [35]. According to this theory, the fitness of an amino acid for any site, and therefore the propensities for the amino acid at that site, is dependent on how well suited it is to the environment formed by the amino acids at neighboring and interacting sites. As substitutions at neighboring sites alter the environment of a site, the amino acid propensities of that site will also be altered, resulting in fluctuating substitution rates at that site. Homologous but divergent proteins in other species will likely have fluctuated differently, meaning that the sets of acceptable amino acids at each position will diverge with evolutionary distance, causing a falloff in the convergence probability. In Stokes-shift theory, divergence in substitution models at a site is strongly coupled to substitutions at that site, so the convergence rate will also be significantly lower following a substitution, consistent with the data shown in figure IV.2.

The SF model makes three additional predictions. First, as the selection at different sites in the protein will be of different and fluctuating magnitude, there should be large

differences and fluctuations in the convergence probability, as shown in figure IV.4B. Second, we would expect more buried locations to be under more stringent constraints, resulting in a higher convergence probability than exposed locations, as shown in figure IV.6B. Third, as also shown in figure IV.6B, we expect the selective constraints at buried locations to diverge slowly because the residues around such locations are also buried and evolve slowly, resulting in a slower decline in the convergence probability with increasing evolutionary distance. All these predictions are matched by the observations of mitochondrial proteins (figs. IV.1 and IV.3).

Both heterogeneity of selection at different sites in the protein and fluctuations in selection over evolutionary time can cause models that neglect these effects to underestimate convergence rates. In particular, the CAT model [39], which includes spatial variation and excludes temporal variation, generates initially high C/D ratios that decline over evolutionary distance in a similar manner as the SF model (fig. IV.6A). Similar drops in C/D ratios can also be seen in other highly parameterized site-specific models of spatial variation (data not shown). However, the effect of spatial versus temporal variation can be distinguished by considering the evolutionary distance dependence of C/D ratios from the same ancestral states. As shown in figure IV.6A, this ratio increases with evolutionary distance when a model is used (CAT) that includes only spatial variation. Sites with fewer constraints are more likely to undergo changes, and therefore less likely to have the same ancestral states at longer divergence times. As a result, as shown in figure IV.6A inset, the sites with the same ancestral states become increasingly the highly constrained sites with lower sequence entropy. As more constrained sites have higher C/D ratios, this means that C/D for these sites will increase with evolutionary distance. In contrast, when there are temporal changes in selection, diverging sequences will increasingly be under different selective constraints. This can result in a decreasing C/D ratio with increasing evolutionary distance, as observed in figure IV.2A. A fluctuating temporal component is not surprising, as no plausible biophysical model would allow site-specific constraints to

remain fixed in the face of divergence in the rest of the protein, and there is other strong evidence for coevolution (or epistasis) among residue positions [42, 35, 36].

The effects of fluctuating and poorly estimated neutral convergence may have substantial effects on phylogenetic inference. Although truly neutral convergence is expected to be unbiased to any particular phylogenetic solution, it may well add considerable noise that would mask true phylogenetic signal. The distance dependence of the convergence probability may also interact in complex ways with the well known phylogenetic problem of long-branch attraction [143], and we expect that extensive analyses will be necessary to sort out such interactions. Furthermore, it is clear that our new understanding of fluctuating substitution processes suggests a multitude of new questions about how protein evolution operates and the role of convergence analysis in understanding protein evolution. Can we use convergence to better estimate instantaneous constraints? Can we understand the role of interactions between different amino acid substitutions at different distances in a protein structure, and how substitutions at those positions affect the probability of convergence? Can we use convergence estimates over different lengths of time to better understand the rates of fluctuation in constraints both with and without substitution at a target site? The inclusion of variation in the substitution process across sites and over time - details that standard models currently lack - should be included in future evolutionary models to obtain more accurate descriptions of protein evolution.

IV.5 Materials and Methods

IV.5.1 Convergence calculations on mitochondrial proteins

Thirteen genes encoded in the mitochondrial genome were downloaded from GenBank for 641 tetrapod species. Separate alignments of amino acid sequences for every gene were made using ClustalX [144]. Aligning a selection of the sequences using PRANK [145] yielded similar downward-sloping results, although there are differences in the height of the early curve (supplementary fig. S7, Supplementary Material online). The mutation pattern in genes across the mitochondrial genome has a complex pattern of

changing asymmetry [146, 147] that is not embodied in current phylogenetic reconstruction programs. We therefore made our phylogeny using only cytochrome oxidase 1 (CO1), which has the least asymmetric mutation rates among vertebrate mitochondrial genes [146, 147]. We partitioned the CO1 data by codon positions and determined the preferred model for the three data partition using the Akaike Information Criterion [148, 149] in MrModeltest v2.2 [150]. The Bayesian consensus tree was determined using the model for each partition (integrating over model parameters) and MrBayes 3.0b4 [151, 152].

The alignments for all genes were concatenated and taxa with a large number of gaps (4500 of 3,596 sites) were removed, leaving 629 taxa. PLEX [87] was used to infer the ancestral sequences and substitutions along the (maximum likelihood) CO1 tree. A fixed amino acid substitution model mtMam [141] was used along with site rate variation with five gamma distributed rates. PLEX analyses were run for 400,000 Markov chain Monte Carlo generations after 100,000 generations of burn-in. All branch pairs except sister branches and branch pairs where one branch was the ancestor of the other were considered. For the significance calculations (supplementary Materials, Supplementary Material online), we sampled the complete set of ancestral node states on the phylogenetic tree every 100 generations. Double substitutions for each branch pair were determined by finding all sites that changed between ancestor and descendant on both branches in the pair in that generation. Double substitutions that ended at the same descendant amino acid in both branches were counted as convergent events, while the remaining double substitutions that ended at different descendant amino acids were counted as divergent events. For simplicity of display, for figures IV.1 – IV.4 the average C/D ratios were calculated using only sites with >90% posterior probability of having a substitution, and branch pairs were included in the average only if D was greater than a specified cutoff (see figure legends). The number of inferred substitutions along the tree were counted to classify sites as conserved (60 or fewer substitutions) or fast-evolving sites (75 or more substitutions). We estimated how well we could infer the C/D ratios using this method

by simulating sequences evolving over the mitochondrial tree and then comparing the C/D ratios from the known ancestors with the inferred ancestors. The results are shown in supplementary figure S8, Supplementary Material online.

IV.5.2 Stokes-Fisher Model

The SF model used to simulate protein evolution in this study has been described previously [60, 117]. It is based on modeling the evolutionary process where the fitness of the protein is the probability that the protein would be folded in a particular "native" structure under equilibrium conditions.

The free energy $G(S, C_k)$ of a protein sequence S in a particular conformation C_k was calculated based on the sum of pair-wise energies between amino acids that are in contact in that conformation (i.e., have their C_β atoms closer than 7Å), using the contact potentials determined by [153] based on their analysis of protein structures. To calculate the free energy of folding $\Delta G_{Fold}(S)$, we calculated $G_{NS}(S)$, the free energy for the native state (the conformation of the 300-residue purple acid phosphatase, PDB 1QHW, [154]) as well as a large ensemble of alternative folds. We assumed that the distribution of the free energies of the large ensemble of thermodynamically relevant unfolded and alternative conformations can be represented by a Gaussian distribution with sequence-dependent average $\bar{G}(S)$ and variance $\sigma(S)^2$, which we estimated by calculating the average free energy and variance of the free energies of the sequence in the conformation of the first 300 residues of 55 different structurally diverse protein structures. Assuming that a large set (10^{160}) of possible unfolded structures with free energies are drawn from that distribution, we can then calculate $\Delta G_{Fold}(S)$ and therefore the probability $P_{Fold}(S)$ that the protein would be folded at equilibrium. As in previous work, we considered the fitness of a sequence $\omega(S)$ to equal the probability that it folded to the native state.

For the star phylogeny simulations, we initialized a protein sequence by choosing 300 codons at random (ignoring stop codons), using the standard genetic code to determine the encoded amino acids. We then computed the codon substitution model at each site

in the protein at each point in time. The mutation rate to all possible alternative codons Ω_{ij} was constructed using the K80 nucleotide model($\kappa = 2$)[21], disallowing multiple nucleotide changes. For each nonsynonymous mutation, ω' of the resulting sequence was computed based on the value of $\Delta G_{Fold}(S')$, the free energy of folding for this sequence, and the corresponding folding probability $P_{Fold}(S')$. This fitness was then compared with the fitness of the premutated sequence ω; the mutation rate was multiplied by the acceptance probability calculated using the Kimura formula for diploid organisms [155, 156, 157]:

$$Q_{ij} = \Omega_{ij} \frac{1 - e^{-2s}}{1 - e^{-4N_e s}} \tag{IV.1}$$

where $s = \frac{\omega' - \omega}{\omega}$, with N_e , the effective population size set equal to 10^6.

The simulation proceeded for a sufficient number of generations such that the stability of the protein reached equilibrium (i.e., the average fitness was approximately constant over time and across independent runs). Equilibrium was reached due to mutation–selection balance, the point where stabilizing mutations are relatively uncommon and have smaller relative fitness benefits, while destabilizing (but marginally acceptable) mutations are greater in number.

For the star phylogeny simulations, 100 replicate sequences were evolved to approximate equilibrium and then split into 10 lineages diverging from one another to produce sequences related by a star phylogeny. Each lineage was evolved for a distance of 10.0 synonymous nucleotide substitutions per nucleotide site from the common ancestor (on average, 6.95 amino acid replacements per amino acid position). We estimated the expected C/D ratios from these data.

To calculate the expected C/D ratios, we considered the instantaneous codon–codon substitution rate matrices given the constraints at each site in the two proteins and the current codons at this site. We then calculated the rate at which a double transition to the same amino acid would be observed in both lineages, compared with the rate at which

a double transition to different amino acids would be observed. C and D were summed over all sites, with the ratio of these quantities computed for that pair of proteins. We then averaged C/D overall pairs of proteins in each star phylogeny, and over all star phylogenies. The details are provided in the Supplementary Material, Supplementary Material online. The observed C/D ratios found in figure IV.4B were obtained from simulating SF over the mitochondrial tree and then inferring the ancestors and C/D ratios using the same method as for the mitochondrial data.

IV.5.3 Phenomenological Substitution Models

We also considered the expected C/D ratio for a variety of phenomenological substitution models, as more fully described in the supplementary Material, Supplementary Material online. We again considered a site in two homologous proteins i and j. We then calculated the probability that every pair of amino acids (or codons) would be observed in proteins i and j. We then used the substitution model to calculate the rate at which these amino acids would undergo a double substitution to the same or different amino acids (or codons coding for the same or different amino acids). C and D were calculated by summing over all possible amino acids (or codons) for sequences k, i, and j. When a gamma distributed rate distribution was used, we also summed C and D over four different rate categories. The ratio then yielded the C/D ratio.

IV.6 Supplementary Materials

Supplementary figures S1–S8 and table S1, and material are available at Molecular Biology and Evolution online (http:// www.mbe.oxfordjournals.org/).

IV.7 Acknowledgements

We thank Todd Castoe and Jason de Koning for contributions to the alignments and tree reconstruction, and Nicolas Lartillot for providing us with the parameters for the CAT 60 model. Richard A. Goldstein and David D. Pollock designed the research, analyzed the data, and wrote the paper. Richard A. Goldstein, Seena D. Shah, Stephen T.

Pollard, and David D. Pollock performed the research. This work was supported by the National Institutes of Health (grant number R01 GM083127) and the Medical Research Council (MRC) UK.

IV.8 Supplementary Materials

IV.8.1 Estimation of C/D ratio for Figure IV.1

To better demonstrate the downward curve of the C/D ratio versus distance graph in Figure IV.1 and our confidence in the curve, we determined the credible intervals for the C/D ratios for the Figure IV.1 data split into 20 windows using Markov Chain Monte Carlo (MCMC) simulations (one for each window). The 99% credible intervals for the C/D ratio for each window are presented in Table S1 below and are graphed in Figure IV.2. The high effective sample sizes and tight intervals show that we are quite confident in our estimations. There is a clear downward trend in the first five windows, as their intervals did not intersect at all. The C/D ratios continue to decrease steadily after the first five windows. The last few windows contained fewer data points and so the C/D ratio estimates were less constrained.

IV.8.2 The common ancestor convergence ratio decreases

We also wanted to test our confidence that the C/D ratios in Figure IV.2A were decreasing. We used a linear model and estimated the slope using an MCMC. After burnin, the slope never came close to zero as the 99% credible region was between -0.109 and -0.094. We sampled the posterior distribution well with a high effective sample size and so we are confident that the slope is indeed negative. We reject the null hypothesis that the slope is zero in favor of the alternative that the slope is negative with a p-value less than 0.01. The 99% credible linear fits are graphed in Figure IV.3.

IV.8.3 Estimate of effective number of accessible amino acids

We can calculate an effective number of accessible residues by considering the size of the alphabet of states m that would result in a particular value of C/D if all substitutions were equally likely (i.e., a Jukes Cantor model [20]. If there is only a short evolutionary

distance between the branches, and the current amino acids in the two branches are likely to be the same, each amino acid can change to $m-1$ other amino acids. The probability that a substitution occurring at each branch would result in the same amino acid would be $\frac{1}{m-1}$, resulting in $C/D = \frac{\frac{1}{m-1}}{1-\frac{1}{m-1}} = \frac{1}{m-2}$. When the amino acids are different, the probability that two substitutions would result in the same amino acid is slightly reduced, as neither substitution can result in either of the original amino acids. The probability that a substitution of one amino acid does not result in the other amino acid is $\frac{m-2}{m-1}$. The probability that the second amino acid substitution results in the same new amino acid is $\frac{1}{m-1}$, resulting in probability of two substitutions yielding a convergent change of $\frac{m-2}{(m-2)^2}$, giving an expected C/D of $\frac{m-2}{m^2-3m+3}$. For large divergence times, the two amino acids have probability $\frac{1}{m}$ of being the same, and probability $\frac{m-1}{m}$ of being different, so the probability of a double substitution resulting in a convergent change is $(\frac{1}{m})(\frac{1}{m-2}) + (\frac{m-1}{m})(\frac{m-2}{(m-1)^2}) = \frac{m^2-3m+3}{m(m-1)(m-2)}$. This corresponds to C/D=$\frac{m^2-3m+3}{m^3-4m^+5m-3}$.

IV.8.4 Exact averages: Phenomenological substitution models

The various phenomenological substitution models (JC, Jukes-Cantor; general time reversible (GTR); Z, Ziheng Yang's codon model, and CAT, the Lartillot group's CAT_{60} model) specify a substitution matrix Q (where Q_{ij} is the substitution rate from state i and state j) as well as π_i, the equilibrium frequencies of state i [22, 39]. These states are amino acids for the JC, GTR, and CAT models, and codons for the Z model. Consider two points on a phylogenetic tree separated by an evolutionary distance τ, as shown in Figure IV.1, where the two states of the sequence at these points are i and j respectively. The probability that these two states would be found at these sites, given a simple reversible Markov process described by Q and π, is given by $\pi_i[e^{Q\tau}]_{ij}$ (Yang 2014). Given these two states, the probability that substitutions occur from i to m and j to n ($m \neq i, n \neq j$) during short intervals of length δ is $Q_{im}Q_{jn}\delta^2$. This would be convergent if $m = n$ and divergent if $m \neq n$. The probability of a set of convergent substitutions is therefore

$$C(\tau) = \delta^2 \sum_{i,j} \pi_i [e^{Q\tau}]_{ij} \sum_{m \neq i,j} Q_{im} Q_{jm} \tag{IV.2}$$

while the corresponding probability of a set of divergent substitutions is

$$D(\tau) = \delta^2 \sum_{i,j} \pi_i [e^{Q\tau}]_{ij} \sum_{m \neq i} Q_{im} \sum_{n \neq j,m} Q_{jn} \tag{IV.3}$$

$$= \delta^2 \sum_{i,j} \pi_i [e^{Q\tau}]_{ij} (\sum_{m \neq i} Q_{im})(\sum_{n \neq j} Q_{jn}) - C(\tau) \tag{IV.4}$$

or by substitution process k, then we can write

$$C/D(\tau) = \frac{\sum\limits_{i,j} \pi_i [e^{Q\tau}]_{ij} \sum\limits_{m \neq i,j} Q_{im} Q_{jm}}{\sum\limits_{i,j} \pi_i [e^{Q\tau}]_{ij} ((\sum\limits_{m \neq i} Q_{im})(\sum\limits_{n \neq j} Q_{jn}) - \sum\limits_{m \neq i,j} Q_{im} Q_{jm})} \tag{IV.5}$$

As can be seen, $C/D(\tau)$ is independent of δ as long as δ is sufficiently short. For real data, δ is must be sufficiently large so that sufficient substitutions can occur. For the theoretical models we can take the limit as $\delta \to 0$, in which case it is the ratio of substitution rates rather than the ratio of substitution probabilities.

Some of the phenomenological models incorporate spatial variation, in that some sites evolve faster or slower (when a Gamma distribution of rates are used in the GTR or Z models) or when there are completely different mechanisms at different sites (as in the CAT model). This is incorporated into the calculations by considering that there is a set of substitution models where substitution process k is characterized by substitution rate Q^k and equilibrium frequencies π^k. If P_k is the proportion of sites characterized by substitution process k, then we can write

$$C/D(\tau) = \frac{\sum\limits_{k} P_k \sum\limits_{i,j} \pi_i^k [e^{Q^k\tau}]_{ij} \sum\limits_{m \neq i,j} Q_{im}^k Q_{jm}^k}{\sum\limits_{k} P_k \sum\limits_{i,j} \pi_i^k [e^{Q^k\tau}]_{ij} ((\sum\limits_{m \neq i} Q_{im}^k)(\sum\limits_{n \neq j} Q_{jn}^k) - \sum\limits_{m \neq i,j} Q_{im}^k Q_{jm}^k)} \tag{IV.6}$$

Note that, as we are counting the total number of convergent versus divergent changes at each branch over all sites, $C(\tau)$ and $D(\tau)$ are independently summed over substitution models.

IV.8.5 Averages: Stokes Fisher model

For calculating the expected value of $C/D(\tau)$ for the Stokes Fisher model, we performed one hundred sets of evolutionary simulations, where each set consisted of ten different lineages arranged as a star phylogeny, each lineage in the set starting from the same initial sequence. This provided us with $100\frac{10!}{8!2!}$ pairs of lineages. Consider lineages k and l, both having evolved for evolutionary time $\frac{\tau}{2}$ from the same initial sequence, i.e., they are separated by evolutionary distance τ, as in Figure IV.1. At these instances in evolutionary time, we know the identity of the amino acid at each site s in the protein ($A(k, s, \frac{\tau}{2})$ and $A(l, s, \frac{\tau}{2})$ in lineages k and l, respectively) and the Stokes Fisher model provides us with the instantaneous substitution rate matrices at this site in these lineages at that time, $Q^{k,s,\frac{\tau}{2}}$ and $Q^{l,s,\frac{\tau}{2}}$. We can then compute average values of and by summing over all sites, with the resulting $C/D(\tau)$ averaged over all pairs of lineages $\langle k, l\rangle$ with a common ancestral sequence:

$$C/D(\tau) = \langle \frac{\sum_{s} \sum_{m \neq A(k,s,\frac{\tau}{2}), A(l,s,\frac{\tau}{2})} Q^{k,s,\frac{\tau}{2}}_{A(k,s,\frac{\tau}{2})m} Q^{l,s,\frac{\tau}{2}}_{A(l,s,\frac{\tau}{2})m}}{\sum_{k} P_k \sum_{i,j} \pi^k_i [e^{Q^k \tau}]_{ij} ((\sum_{m \neq i} Q^k_{im})(\sum_{n \neq j} Q^k_{jn}) - \sum_{m \neq i,j} Q^k_{im} Q^k_{jm})} \rangle \tag{IV.7}$$

IV.8.6 Estimating the C/D ratio with Distance

We modeled each data point as the result of a Bernoulli process with C+D trials and C events. The probability that any double substitution is convergent can be derived from the overall C/D ratio for the window (Rw) using Equation S.62. Therefore the likelihood of the data point with C convergences and C+D total double substitutions in window w with a probability above can be calculated using the binomial distribution. The MCMCs

were allowed to run for 97,000 generations after 3,000 generations of burnin to achieve an effective sample size around 1,000. We integrated over phylogenetic uncertainty by sampling substitution data from PLEX every 100 generations, on average.

$$P_{converge}(w) = R_w/(R_w + 1) \tag{IV.8}$$

$$P(C, D|w) = \binom{C+D}{C} P_{converge}(w)^C (1 - P_{converge}(w))^D \tag{IV.9}$$

IV.8.7 Estimating the error on C/D ratio with Distance

In order to determine how well we could estimate the true C/D ratio using this method, we simulated sequences evolving over the mitochondrial tree using WAG, CAT, and Stokes-Fisher models. Since we knew the entire ancestral history of these sequences, we could calculate the how actual C/D changed with distance (Figure S7, blue curves). We then inferred the ancestors using PLEX and estimated the C/D using the method above. The estimated C/D curves are shown in red in Figure S7. The estimates are quite close to the true curves with the Stokes estimate being slightly low.

IV.8.8 Linear Fit Common Ancestor C/D ratio

We also wanted to quantify our confidence in the downward slope of the data in Figure IV.3:A. We estimated the slope (m) and intercept (b) of the C/D ratios (R) in Figure 2A using an MCMC determine the credible intervals (see Equation S.8). Again we modeled each data point as the result of C+D trials of a Bernoulli process and C events, however the probability used in the likelihood calculation was derived from the linear model of the C/D ratio. The probability of a data point with C convergences, D divergences, and at a distance of t resulting from a model with slope m and intercept b is shown in Equation S.10. The MCMC ran for 14,000 generations after 1,000 generations of burnin to achieve an effective sample size of 173 for the slope and 144 for the intercept.

$$R(t|m, b) = m * t + b \tag{IV.10}$$

Window	Distance Range	C/D Ratio 99 Percent Credible Interval	Effective Sample Size
1	0 - 0.105	0.47 - 0.49	1463
2	0.105 - 0.21	0.32 - 0.34	966
3	0.21 - 0.315	0.26 - 0.27	865
4	0.315 - 0.42	0.23 - 0.23	1020
5	0.42 - 0.525	0.21 - 0.22	1073
6	0.525 - 0.63	0.20 - 0.21	1025
7	0.63 - 0.735	0.19 - 0.20	839
8	0.735 - 0.84	0.18 - 0.19	936
9	0.84 - 0.945	0.18 - 0.19	965
10	0.945 - 1.05	0.18 - 0.19	979
11	1.05 - 1.155	0.18 - 0.18	937
12	1.155 - 1.26	0.17 - 0.18	939
13	1.26 - 1.365	0.17 - 0.18	914
14	1.365 - 1.47	0.17 - 0.18	1070
15	1.47 - 1.575	0.17 - 0.18	1050
16	1.575 - 1.68	0.16 - 0.17	1142
17	1.68 - 1.785	0.16 - 0.17	1132
18	1.785 - 1.89	0.16 - 0.18	1200
19	1.89 - 1.995	0.15 - 0.18	1517
20	1.995 - 2.1	0.11 - 0.19	885

Table IV.1: Posterior probability for topologies with substantial representation in the uncorrected posterior for the 10-taxon dataset. The topologies are labeled for reference in Figure VI.7.

$$P_{converge}(t|m,b) = \frac{R(t|m,b)}{R(t|m,b)+1} = \frac{m*t+b}{m*t+b+1} \tag{IV.11}$$

$$P(C,D,t|m,b) = \binom{C+D}{C} P_{converge}(t)^C (1-P_{converge}(t))^D \tag{IV.12}$$

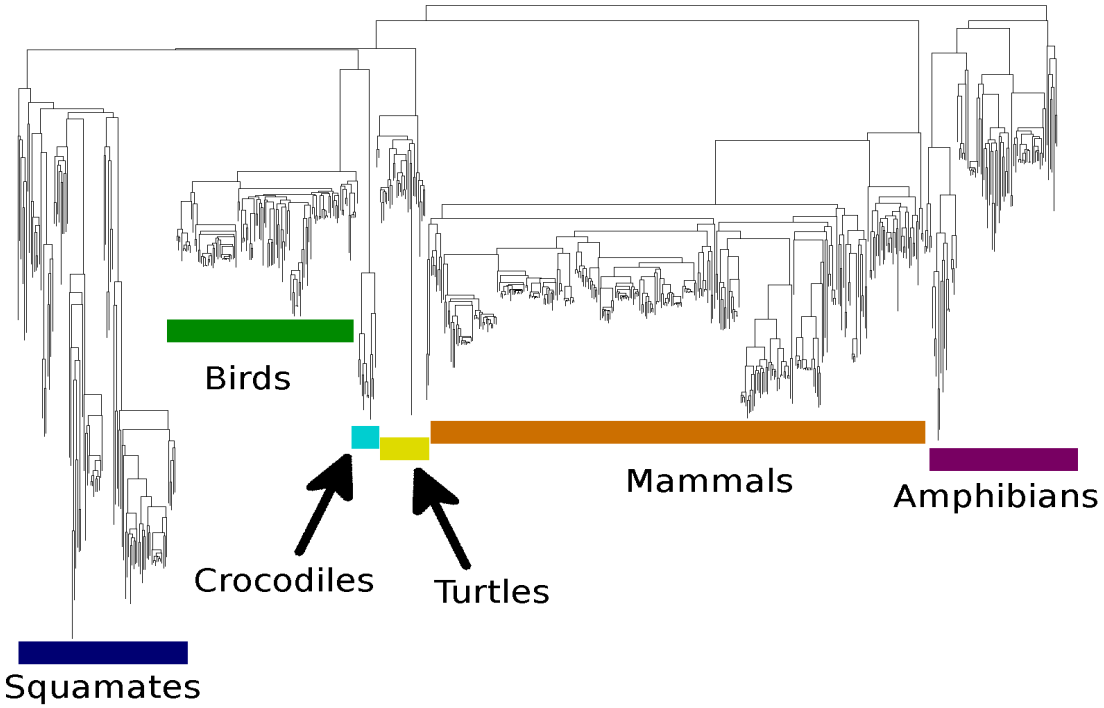
IV.8.9 References

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736.

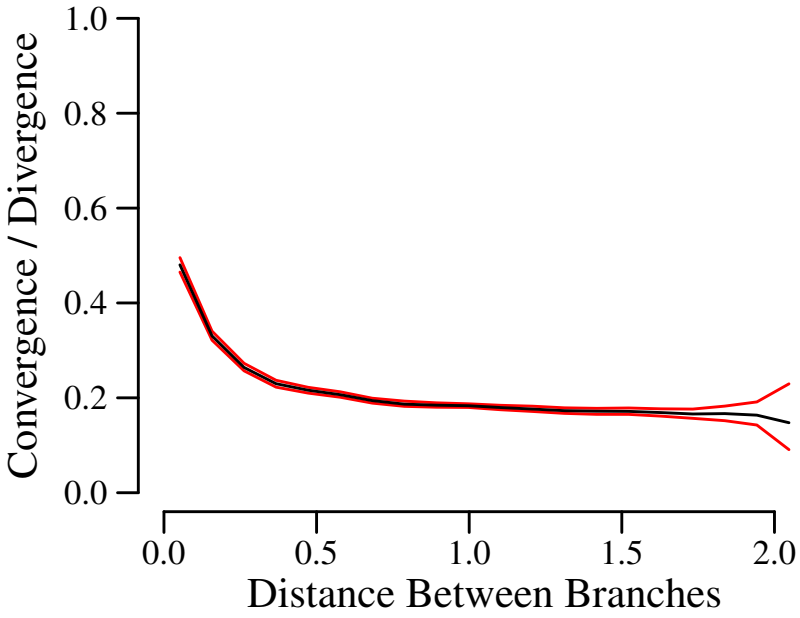
Jukes T, Cantor C. 1969. "Evolution of protein molecules.". In: Munro H, editor. Mammalian Protein Metabolism, vol. III. New York: Academic Press. p. 21-132.

Quang lS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24: 2317-2323. doi: 10.1093/bioinformatics/btn445.

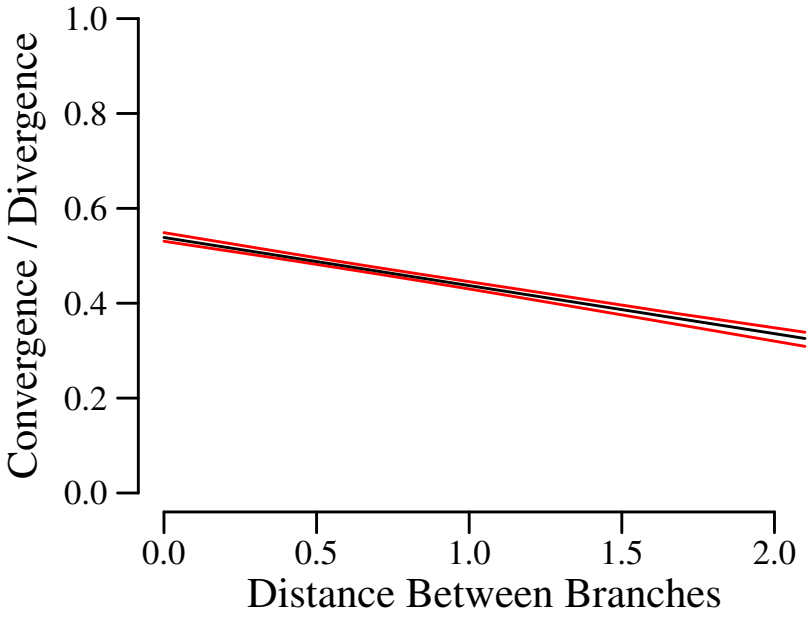
Yang Z. 2014. Molecular Evolution: A Statistical Approach: Oxford University Press.



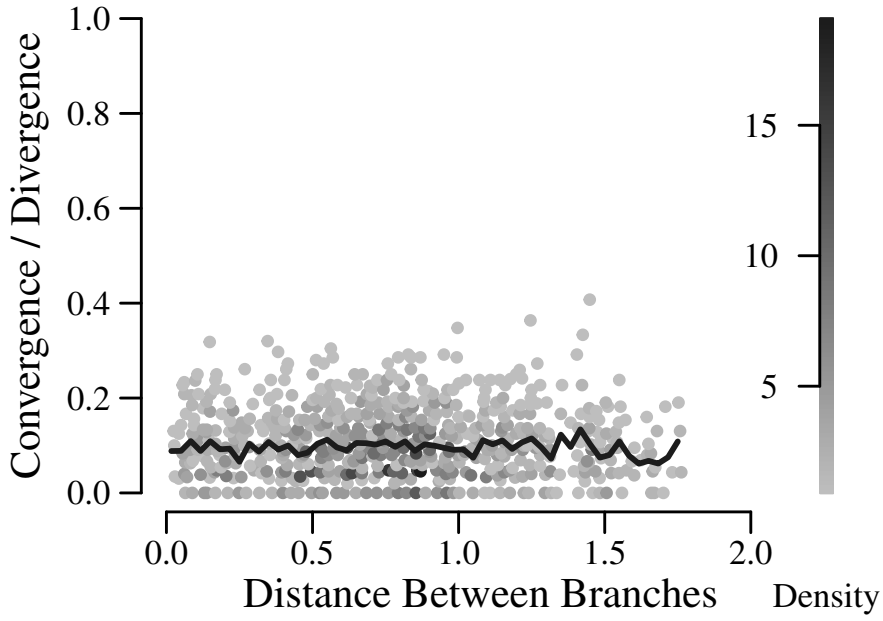
Supplementary Figure IV.1: The mitochondrial COI phylogenetic tree. The tetrapod mitochondrial tree shown was derived from COI sequences as described in the methods. This is the tree that was used in all analyses. Due to the large number of taxa, each individual taxon is not labeled, but clades that are labeled are squamates (blue), birds (green), crocodiles (cyan), turtles (yellow), mammals (orange), and amphibians (purple).



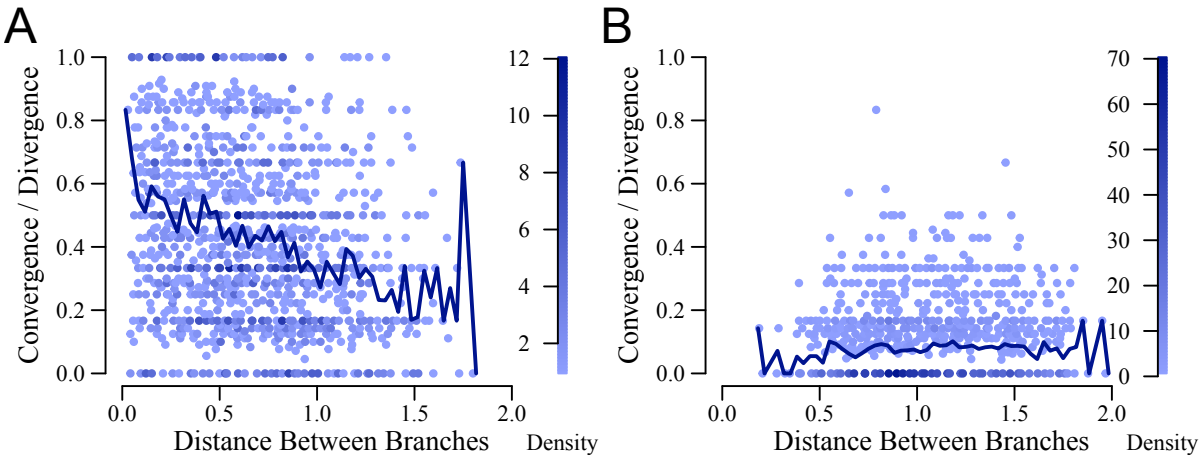
Supplementary Figure IV.2: The 99% credible region of convergence ratio in mitochondrial proteins. The convergence ratios were estimated for the data in Figure 1, divided into 20 windows. The 99% credible regions (red) are extremely tight showing that we can estimate the C/D ratios for each window very well. The mean C/D is shown in black. The C/D for first few windows are clearly decreasing, while the rest of the windows slowly decrease.



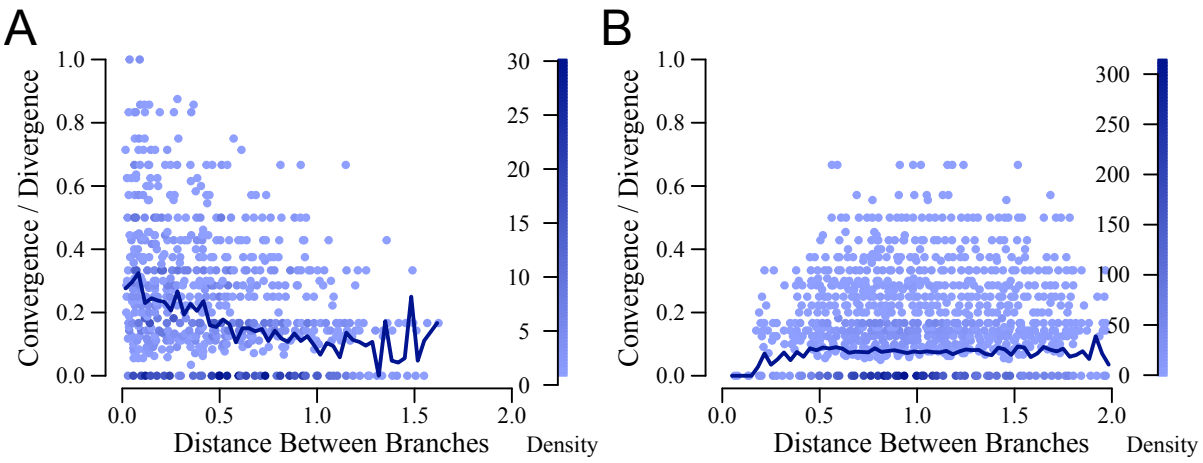
Supplementary Figure IV.3: The 99% credible linear fits of the common ancestor convergence ratio. The slope of a linear model fit to the mitochondrial data in Figure 2A was estimated. The slope never approaches zero and is always negative with a p-value of less than 0.01. The mean is shown in black and the 99% credible region is shown in red.



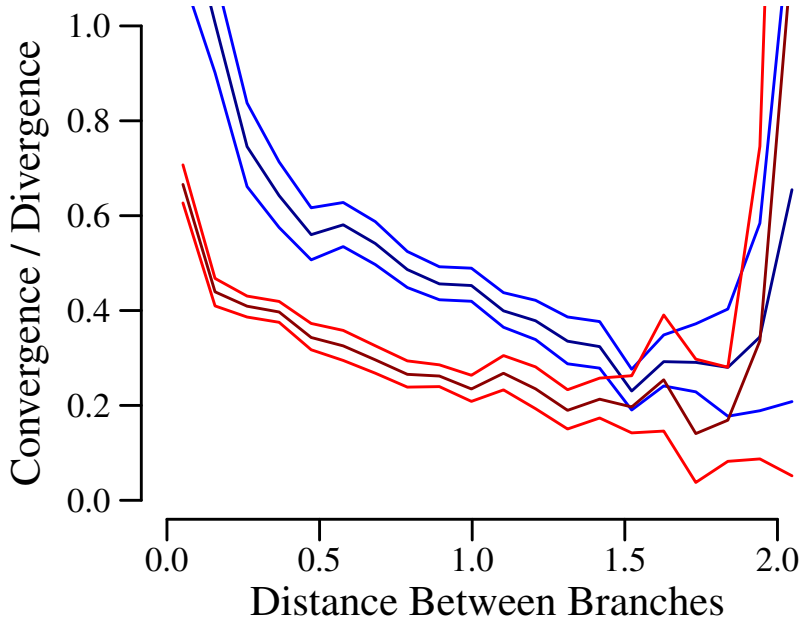
Supplementary Figure IV.4: Randomized convergence ratios. The data and visualization are the same as in Figure IV.1, except that convergence events were randomized across sites.



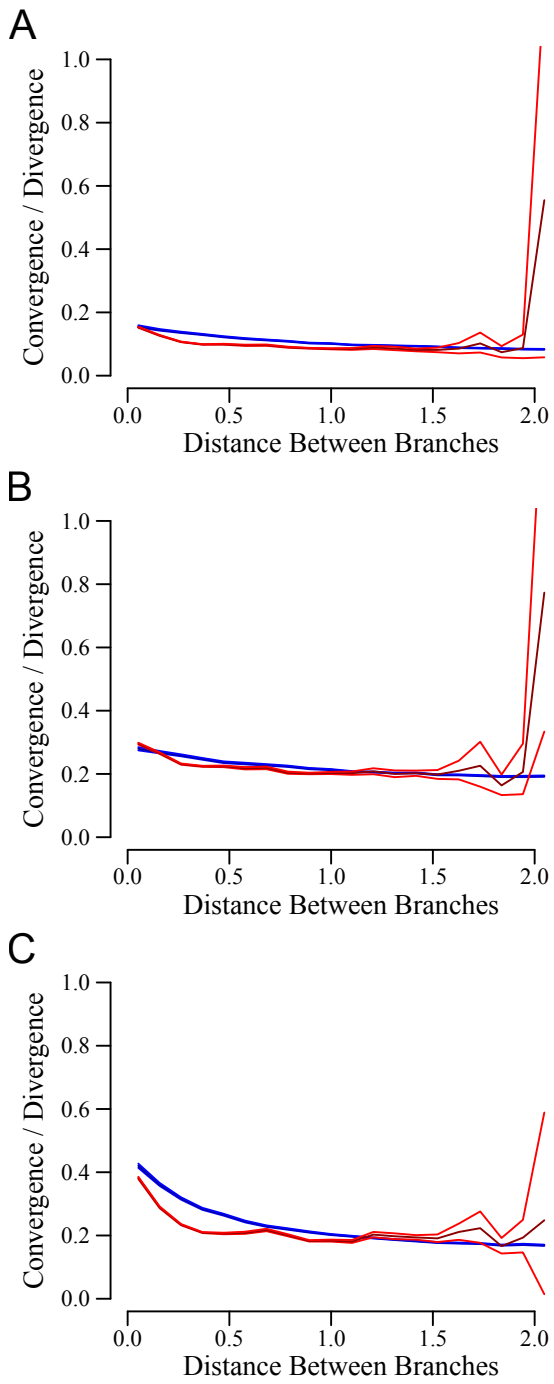
Supplementary Figure IV.5: Convergence ratios at conserved sites for common and different amino acids. The data and visualization are the same as in Figure IV.3A, except that events were separated into two categories depending on whether the ancestral amino acid at a site was the same (A) or different (B).



Supplementary Figure IV.6: Convergence ratios at variable sites for common and different amino acids. The data and visualization are the same as in Figure IV.3B, except that events were separated into two categories depending on whether the ancestral amino acid at a site was the same (A) or different (B).



Supplementary Figure IV.7: Comparing alignment methods. We compared two different alignment methods to show that the downward sloping curve is robust against alignment methods. The 99% credible region of the C/D ratios using ClustalX (blue) and PRANK (red). The means of the C/D curves for ClustalX and PRANK are in dark blue and dark red, respectively.



Supplementary Figure IV.8: Estimating the error of C/D estimates. We generated sequences evolving over the mitochondrial tree using different models: WAG (A), CAT (B), and Stokes Fisher (C).The 99% credible region of the C/D ratios using the known ancestors (blue) and the inferred ancestors (red) are very close, showing that we can estimate the C/D curves very well. The means of the C/D curves for known and inferred ancestors are in dark blue and dark red, respectively.

CHAPTER V

DETECTING AMINO ACID PROPENSITY CHANGES OVER TIME

V.1 Abstract

V.1.1 Background

Some models concerned with protein evolution included the assumption that the evolution process is homogeneous in both time and sites, suggesting that substitution rates may be the same across time and at every site. Few proposed models are compatible with allowing site-specific amino acid fitnesses to fluctuate over time. I propose a novel and general method to determine how amino acid propensities shift over time and across sites. I hypothesize that substitutions cause larger shifts at adjacent sites, and search for substitutions at adjacent sites which may have caused the shift in amino acid propensities.

V.1.2 Results

First I estimate the amino acid propensities at all sites for each ancestral branch in the phylogenetic tree, then I calculate the differences between the amino acid propensities at the ancestral and descendant sides of all branches. I compare the distributions of shifts of amino acid propensities between with and without adjacent substitutions to determine that substitutions cause larger propensity shifts at adjacent sites.

V.1.3 Conclusions

The Acceptability model has been shown to be effective at estimating the propensities at sites even as they change over time, as demonstrated by the simulation results. In estimating how amino acid propensities change over time, I have shown how substitutions at adjacent sites correlate with large shifts in amino acid propensities.

Keywords: Substitution models, Time heterogeneous, amino acid propensities

V.2 Background

Protein evolution proceeds in many different ways including through nucleotide substitutions, nucleotide insertions and deletions, and speciation events. Substitutions occur

when a mutation from one nucleotide to another at a specific site in the genome becomes fixed in a species, meaning nearly all individuals in the species have the mutation. Insertions and deletions are mutations which add or remove nucleotides from the genome, but in order to be preserved in the species history, they also must be fixed in the species. A speciation event is defined to be when one species splits into two or more species. Speciations often cause the history of a set of extant species to be a tree structure, meaning that there is a species which is the ancestor of all the species in the set, called the root of the tree, and this root splits over and over again until its branches connect to the extant species. Extant species are species that are currently alive, as opposed to extinct. Each speciation event is represented by a node on the tree and the independent evolution of each species is represented by a branch between nodes. Substitutions within proteins occur along branches, and the length of each branch is in terms of the expected number of substitutions between the ancestral node of the branch and the descendant node.

Usually each extant species has a sequence associated with it. This sequence can be nucleotides or amino acids, and the sequences for all the extant species have been aligned such that each site in the sequences is considered to be related by substitutions alone. Sites exhibiting evidence of insertions or deletions are sometimes removed, leaving only substitutions to connect the sequences. In order to estimate the probabilities of each substitution on the tree and the sequences of each ancestral node from the extant species sequences, a phylogenetic model of sequence evolution is used. Many different types of evolutionary models have been developed for phylogenetics use, see [143, 158]. Some of these models depend on the amino acid propensities at individual sites. Amino acid propensities Π have been defined in many different ways [40, 48, 26, 159], however in this study I define amino acid propensities to be the relative fitnesses of each amino acid at a site. Propensities are defined in terms of the fitness of each amino acid at a site along a branch by:

$$\Pi_{i,s}^{x} = \frac{\omega_{i,s}^{x}}{\sum_{y} \omega_{i,s}^{y}} \tag{V.1}$$

where x is the amino acid at site i, s is the branch where the propensities are defined, ω_i^x is the fitness of amino acid x at site i, and $\sum_y \omega_{i,s}^y$ is the sum of the fitnesses of all amino acids at site i on branch s [26].

Some models concerned with protein evolution included the assumption that the evolution process is homogeneous in both time and sites, suggesting that substitution rates may be the same across time and at every site [20, 24, 160, 21, 142]. This assumption has been applied to the substitution probabilities among the different residues and the average residue frequencies in these models [20, 21, 22]. Site-homogeneity has been challenged by allowing sites to vary in their overall substitution rate by assigning them to different categories or having the rates change along a sequence continuously [146, 147]. Sites can also be split up into different categories based on characteristics such as inferred average substitution rate or their location in the 3D structure [29, 30]. Models allowing the amino acid frequencies or matrices to vary across sites have also been proposed [23, 161]. For example, the CAT model is an infinite mixture model which assigns sites into categories and infers the number of site categories from the sequence data itself [23].

Strong evidence has been found that amino acid propensities at specific sites can change over time as well [41, 42, 47, 48, 49, 50, 162, 163, 26, 164]. The decreasing probability of reversions after substitution also supports changing amino acid propensities [100]. Some of this change over time has been attributed to epistasis [35, 114, 165]. Some sites have exhibited very little change in amino acid propensities over long time periods. If time heterogeneity is considered in model building, it is often evaluated within the time homogeneous framework by breaking up the tree into smaller trees of time homogeneous evolution [43, 44, 45, 54, 163]. Other models of time-heterogeneous evolution processes have been utilized including only allowing the G+C content to change over the tree [51, 52]. These methods of introducing time-dependence frequently start with the assumption

of time homogeneity and time-dependence is added as a secondary consideration.

Recent results about the rate of convergence cannot be explained by time homogeneous models [48]. Convergence is when two branches on a phylogenetic tree substitute to the same residue apparently independently. In Goldstein et al. 2015 we showed that the instantaneous convergence rate is high, indicating high constraint at any given time and site. The expected number of acceptable amino acids at a specific site varied between two to four amino acids at a time. Traditional models are averages over time and sites and so their expected constraint is generally low. Convergence levels start off high and decrease as the branches diverge. This trend holds even when considering branches which start off having the same residue, indicating that the substitution probabilities are changing. The distribution of substitutions from the mitochondrial data show that substitutions tend to be clustered on the tree and there tend to be large areas of the tree where amino acids become completely fixed, again indicating a time-dependence of the process, as seen in Figure V.1.

Given the theoretical work on the Stokes-Fisher and evolutionary mechanics models, and the supporting evidence for these kinds of models which include structure, I now would like to find more evidence that amino acid propensities are changing over time [35, 116]. If we move from an average over sites then to an average over time (site-specific) then to an average over a time segment (site- and time-specific), we lose the ability to estimate the average well. These averages do not predict the convergence we see [48], but the change in constraints or fluctuating constraints can account for the convergence observations.

Few proposed models are compatible with allowing site-specific amino acid fitnesses to fluctuate over time. One reason for this is that fitting even fixed site specific amino acid fitnesses introduces many parameters, while allowing those fitnesses to vary increases the number of parameters even more [88, 23]. One alternative is to reduce the possible values of the fitnesses to a simple binary fitness [166, 167, 168, 169, 57]. In this study, I propose

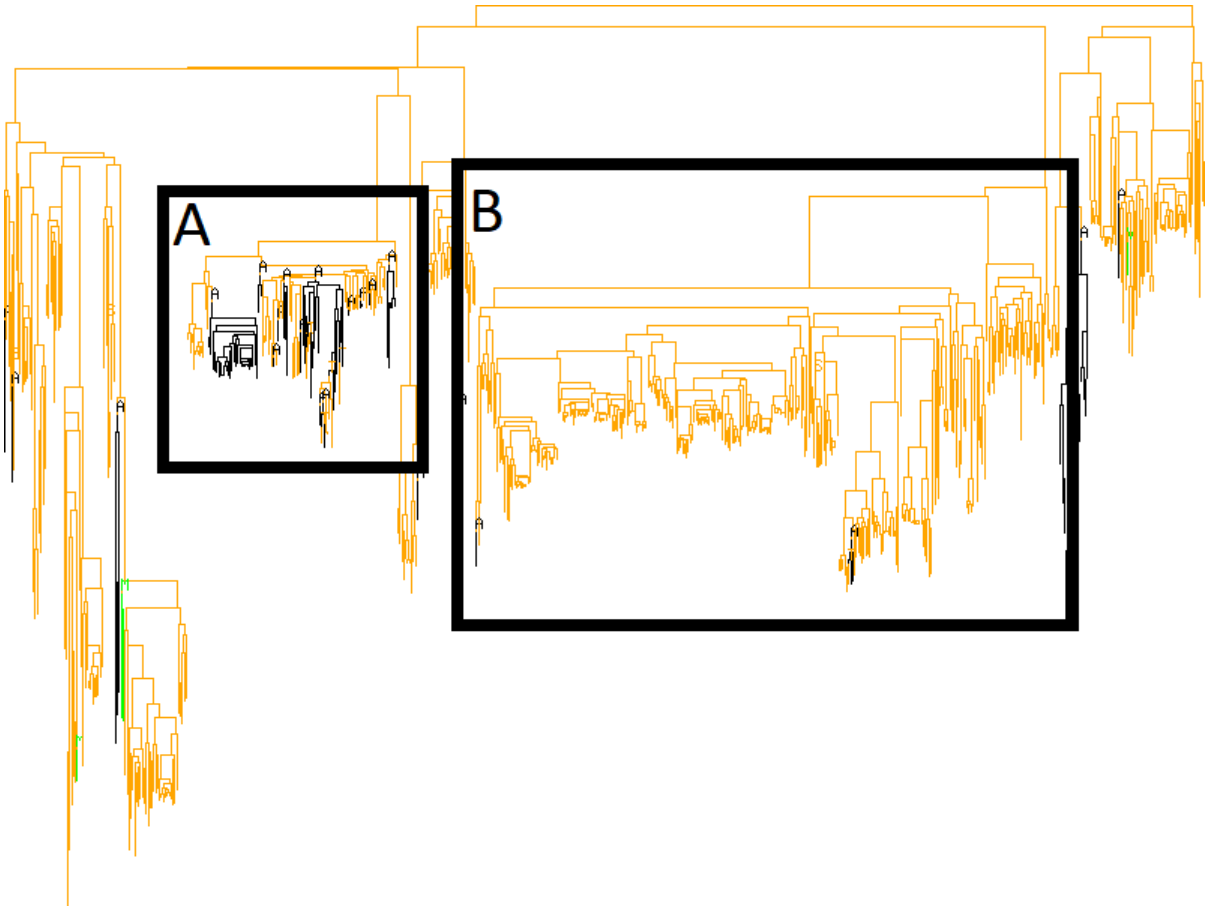


Figure V.1: The inferred substitution history of site 30 in Cytochrome C Oxidase I on a vertebrate mitochondrial tree. Box A surrounds the clade corresponding to birds and Box B surrounds the clade corresponding to mammals. The color orange on the phylogenetic tree represents where the resident amino acid at site 30 is threonine, while the black color on the tree is where alanine is resident at site 30. There are multiple substitutions from threonine to alanine and in some cases, vice versa, in the birds, while in the mammals, threonine seems to have become largely fixed and there are far fewer substitutions to alanine, showing a that alanine became much more acceptable in the birds. From these observations, one can expect threonine alone to be acceptable for most of the mammal clade, while both alanine and threonine are acceptable for the bird clade. The substitutions in the bird clade suggest that the acceptable amino acids, which are alanine and threonine, are different from the mammal clade, which is almost entirely threonine.

a new method to estimate when amino acids were high or low fitness at individual sites in a protein across an entire tree. The fitness of a particular amino acid is allowed to change over time and over sites. Amino acids are considered high fitness at a site and time if the protein continues to function with that amino acid at that site. After observing in the mitochondrial ancestral reconstructions that substitutions are often clustered together on the tree and that there are many convergences, I apply this method to a large tree of mitochondrial sequences and to a collection of nuclear glycolysis proteins.

In this study, I attempt to determine how amino acid propensities shift over time. First I estimate the amino acid propensities at all sites for each ancestral branch in the phylogenetic tree, then I calculate the differences between the amino acid propensities at the ancestral and descendant sides of all branches. I hypothesize that substitutions cause larger shifts at adjacent sites either on the immediate N- or C-side of an amino acid, and search for substitutions at adjacent sites which may have caused shifts in amino acid propensities at adjacent sites. I compare the distributions of shifts of amino acid propensities at sites with or without adjacent substitutions to determine whether substitutions cause larger propensity shifts at adjacent sites.

V.3 Glossary

Extant species - A species that is currently alive, as opposed to extinct. Extant species are often the leaves on a phylogenetic tree and the residue sequences are often considered known.

Speciation - The splitting of one species into two. These events are usually represented by nodes on a phylogenetic which have two descendant branches.

Phylogenetic tree - A history of speciation events from the root (most recent common ancestor) to the extant species on the leaves.

Ancestral species - A historical species on a phylogenetic tree which is closer to the root.

Descendant species - A species on a phylogenetic tree which is closer to the extant

species.

Branch - A connection between speciation events on a phylogenetic tree. Branches often have a length expressed in terms of the expected number of substitutions per site.

Mutation - Any immediate change in the DNA often from parent to daughter cell. Point mutations are changes in a single nucleotide at a single site.

Substitution - The fixation of a mutation in a population or species, such that nearly all organisms in the population or species have the mutation.

Resident amino acid - The amino acid at a specific site in a protein at a specific time.

Amino acid propensity - Amino acid propensities Π are defined in terms of the fitness of each amino acid at a site on a branch by:

$$\Pi_{i,s}^{x} = \frac{\omega_{i,s}^{x}}{\sum_{y} \omega_{i,s}^{y}} \tag{V.2}$$

where x is the amino acid at site i, s is the branch where the propensities are defined, ω_i^x is the fitness of amino acid x at site i, and $\sum_y \omega_{i,s}^y$ is the sum of the fitnesses of all amino acids at site i on branch s [26]. Sometimes called amino acid profiles or preferences.

Inferred amino acid propensity - Amino acid propensities on ancestral nodes based on the extant sequences of the clade below the node.

Fitness - The ability of an organism to live and reproduce, often defined relative to other organisms. This can be influenced by the organism's genotype, phenotype, and environment.

Multiple Sequence Alignment (MSA) - The sequences of the same region of a genome (DNA, RNA, protein, transposable element) from multiple species, aligned such that each row is a different species and each column is an orthologous site. Each site is presumed to connected by substitutions alone.

Adjacent site - A site in a Multiple Sequence Alignment which is immediately before or after the site of interest. For example, sites 4 and 6 are adjacent to site 5.

V.4 Methods

V.4.1 The Protein Data

The mitochondrial protein sequences analyzed come from the vertebrate mitochondrial data set used in Goldstein et al. [48]. They are 629 tetrapod mitochondrial sequences from GenBank and aligned using ClustalX [144] into a multiple sequence alignment (MSA). ClustalX was selected because it is fast with large sequences and many taxa. In this alignment, each individual amino acid is referred to as $x_{s,i}$ with two different indices: s, a string referring to the row and corresponding species name, and i, an integer referring to the column. For example, the amino acid at site 30 in the sequence of the ancestral node Node_432 would be represented as $x_{Node\_432,30}$. The entire multiple sequence alignment taken together is represented by a capital X.

The nuclear glycolysis protein sequences were collected from OrthoDB.org and selecting only vertebrate sequences. Proteins were selected based on the number of sequences and species represented in the database. The proteins chosen were Acyl-CoA synthetase short chain family member 3, ATP-dependent 6-phosphofructokinase, fructose-2,6-bisphosphatase TIGAR, Fructose-bisphosphate aldolase, GTP-binding protein, hexokinase-2, L-lactate dehydrogenase, Phosphoglycerate mutase, and Pyruvate kinase. These sequences were then aligned using Clustal Omega (the successor to ClustalX) and the alignment was analyzed using MaxAlign to remove sequences, species, and columns with large numbers of gaps [170, 171]. One sequence per species was chosen for each protein.

In order to perform the phylogenetic analysis, a tree was built using the multiple sequence alignment. The Cytochrome C Oxidase subunit I sequences were selected from the MSA and used in MrBayes version 3.0b4 to build a consensus tree [151, 152]. This tree was used throughout the analysis for the mitochondrial sequences. For the nuclear proteins, a tree was built using MrBayes with the hexokinase-2 sequences. MrBayes was used because it handles large data sets well.

The most likely ancestral sequences and substitution history was sampled by PLEX using the mtMam model [87, 141]. The mtMam is a fixed 20x20 amino acid matrix model which was generated by fitting the substitution probabilities and amino acid exchangeabilities to an alignment of the mitochondrial proteins. It was selected for this analysis because it was easy to implement into PLEX and I am using mitochondrial sequences. When I infer the most likely substitution history, I can calculate the most likely sequences at each ancestral species. When the ancestral species are inferred, the sequences are added to the multiple sequence alignment with the ancestral species name indicating the row. Using the most likely substitution history requires less computational cost, as compared to integrating over the uncertainty in the ancestral state reconstruction. I expect there to be minor differences between the substitution history inferred using the mtMam model, and the possible substitution history inferred using Acceptability Model. In order to demonstrate the method's robustness to ancestral sequence reconstruction methods, the nuclear protein ancestors were reconstructed using PAML with the default settings [172].

In order to expedite the likelihood calculations and to allow the substitution process to change along branches, long branches were broken up into smaller segments of a predetermined maximum length, according to the precision required. This corresponds with the B1xM approximation used in de Koning et al. [65]. Using longer branch segments will result in fewer segments and faster calculations, however shorter segments will give more accuracy. In order to avoid double substitutions, a maximum branch segment length of 0.08 is recommended [65]. Ideally, this length would be short enough to disallow double fitness switches, however this will depend on the data set used. A max branch length of 0.05 was used throughout this study. When the branches of the tree are subdivided, the sequences of the ancestral species corresponding to the branch segments are also added to the multiple sequence alignment.

The B1u method was also proposed and tested in de Koning et al. [65] as applied

to phylogenetic contiuous time Markov chains. The major difference between B1u and B1xM is that in the B1u method, the instantaneous rates away from every state (in our case amino acid state) are made the same (uniformized) by adding a 'virtual substitution' rate. Virtual substitutions are virtual because they are changes from one amino acid to itself. This modification allows for the likelihood of the entire tree to be calculated faster when updating the branch lengths or substitution rates. When the substitution matrix is fixed over the whole tree, this can lead to a large acceleration, however if the matrix changes continuously over time B1u adds complication. The B1u method was not used in this study because it would require calculating an entire rate matrix for every branch in the tree and for every site. The current implementation of the Acceptability model only calculates the necessary terms in the substitution rate matrix and not the entire matrix. For example, when calculating the probability of an I to L substitution when I, L, and M are acceptable, one needs to consider only the probabilities of the descendant residue being I, L, of M given that the ancestral residue is I, rather than the 400 possibilities of the full 20x20 substitution rate matrix.

V.4.2 Amino Acid Propensity Estimation

In order to detect large shifts in amino acid propensities at a specific site on a specific branch, one must first infer the amino acid propensities. For this task, I introduce a new method based on binary amino acid fitnesses which are allowed to change across the protein and through time. The two amino acid fitness states are high fitness, or "acceptable", and low fitness, or "unacceptable". The amino acid propensities for each site are calculated using the probabilities of each amino acid being in the high fitness state. When all the resident amino acids in the protein are high fitness, then the protein functions properly. The definition of the low fitness state for an amino acid at a site is that if that amino acid were resident, the protein would not function properly. Every amino acid at every site at every point in time has been in either the high fitness state or the low fitness state. Because of natural selection, the resident amino acid at a site at a

point in time is considered to be in the high fitness state.

The goal of this method is to infer the set of high fitness amino acids for every site and every branch on the tree. All of the sets for every site and every branch segment are represented by the variable A, where $a_{s,i}$ is the set of high fitness amino acids at species s and at site i. The high fitness set for each site is allowed to change continuously across time within the same site. I do not directly model any forces that define which amino acids are acceptable at any site at any point on the tree. I simply acknowledge that at every time in history and for every site, more amino acids may have been acceptable than only the resident amino acid. It is possible for a site to have any number of high fitness amino acids at a time.

This method relies on a protein evolution model which allows for amino acids to switch between high and low fitness over time and to substitute among high fitness amino acids. The parameters of the model include the amino acid fitness switching rate (ν) and the substitution rate among amino acids of high fitness (λ). The fitness switching rate defines the instantaneous rate of an amino acid switching between the high fitness and low fitness states. The instantaneous switching rate (ν) is used for both rates from acceptable to unacceptable and from unacceptable to acceptable. Although a constant substitution probability for all amino acids is used here in order to simplify the computation, any substitution probability matrix could be used for the different substitution probabilities among acceptable amino acids. The substitution rates among acceptable amino acids can be defined by any instantaneous substitution model, such as a Jukes-Cantor or WAG model [20, 142]. P_λ and P_ν are the probabilities of a substitution or amino acid fitness switch given a branch length of 0.05 and substitution rate λ and switch rate ν. Since the branch segments are so short, we can safely ignore the possibility of double substitutions or double fitness switches. These can be calculated using the equations:

$$P_\lambda = 1 - e^{-\lambda * 0.05} \tag{V.3}$$

$$P_\nu = 1 - e^{-\nu * 0.05} \tag{V.4}$$

The sets of acceptable amino acids are not considered proper parameters of the model, similar to how inferred ancestral amino acids on a phylogenetic tree are not considered part of phylogenetic models. Instead the acceptable sets are considered nuisance variables, which must be inferred and integrated away. The acceptable set with the resident amino acid constitutes the entirety of the state of the evolutionary process. From this state and the substitution rate, one could construct the rate matrix for any branch on the tree in order to calculate the substitution probabilities along that branch. Transitions of an amino acid between acceptable and unacceptable are also treated as a continuous time Markov process.

An extension to this model is to include the acceptable set size prior (ρ). The acceptable set size prior is an array of probabilities summing to one which denotes the probabilities of different numbers of amino acids being acceptable at the same time. One can use acceptable set size prior ρ to influence the amount of instantaneous constraint at all sites to be relatively high, in light of previous evidence that only a few amino acids are acceptable at any point in time. I can use two different forms of prior to limit the number of acceptable amino acids. The first form of prior is to choose a prior probability that a particular amino acid is acceptable at any time, e.g. $\frac{1}{10}$. This leads to a prior of having N acceptable amino acids of

$$\rho(N) = C(\frac{1}{10})^N \tag{V.5}$$

where C is a normalization constant such that the priors sum to 1. The second form of prior is any discrete probability distribution with 20 different values corresponding to having N acceptable amino acids. One could use any distribution for this prior, such as a Poisson distribution or discrete gamma distribution.

V.4.3 Estimating the Acceptable Sets

In order to estimate the propensities, all of the acceptable sets A must be fit to sequence data X and phylogenetic tree T. The likelihood of the model M given a multiple sequence alignment X, a phylogenetic tree T, and acceptable sets A is:

$$L(M|X,T,A) = \prod^{s\in T}\prod^{i\in X} e^{-\lambda j t_s}(1-e^{-\lambda t_s})^{1-j} e^{-\nu k t_s}(1-e^{-\nu t_s)^{20-k}\rho(a_{s,i})} \quad \text{(V.6)}$$

where j is the number of substitutions (either 0 or 1) at site i along the branch segment s, k is the number of amino acids which switched from high fitness to low fitness or vice versa along the branch segment, t_s is the length of the branch segment in terms of average substitutions per site, s is a branch segment in the tree T, and i is a site in sequence alignment X. A substitution at site i occurs when $x_{s,i}$ is not equal to $x_{s-1,i}$, where $s-1$ is the ancestral species to species s. The number of amino acids k which switched fitness states at site i is determined by comparing the fitness states of each amino acid in the ancestral species $a_{s-1,i}$ and the acceptable amino acids of the descendant species associated with the branch $a_{s,i}$.

This form of the likelihood equation is computationally expensive to calculate over the entire tree especially when sampling the amino acid switch rate ν and the substitution rate λ. Breaking long branches down into segments shorter than 0.05 using the B1xM approximation allows for a substantial simplification of the likelihood equation. The approximate likelihood of the model M given a multiple sequence alignment X, a phylogenetic tree T, and acceptable sets A is:

$$L(M|X,T,A) = \prod^{s\in T}\prod^{i\in X} P_\lambda^j (1-P_\lambda)^{1-j} P_\nu^k (1-P_\nu)^{20-k}\rho(a_{s,i}) \quad \text{(V.7)}$$

where j is the number of substitutions (either 0 or 1) at site i along the branch segment, and k is the number of amino acids which switched from high fitness to low

fitness or vice versa along the branch segment, s is a species in the tree T, and i is a site in sequence alignment X. A substitution at site i occurs when $x_{s,i}$ is not equal to $x_{s-1,i}$, where $s-1$ is the ancestral species to species s. The number of amino acids k which switched fitness states at site i is determined by comparing the fitness states of each amino acid in the ancestral species $a_{s-1,i}$ and the acceptable amino acids of the descendant species associated with the branch $a_{s,i}$. P_λ and P_ν are defined by equations V.3 and V.4.

Since this model is different from the typical evolution models used in phylogenetics currently, some methods for model fitting and likelihood calculation are not appropriate for this model. Felsenstein's pruning algorithm cannot be applied because the substitution probabilities change over time. For this study, I used a Markov chain Monte Carlo method to fit the Acceptability model to the mitochondrial data. The Markov chain Monte Carlo method is used to fit models to data by sampling the posterior distribution of the model parameters. It does so by starting with random values of the model parameters then proposing new values to the parameters, according to the Metropolis-Hastings method. These updates are accepted if the likelihood of the model given the new parameter values is higher than the likelihood given the old parameter values. If the new likelihood is lower than the old likelihood, then the new parameters are accepted with a probability equal to $\frac{L_{new}}{L_{old}}$.

For some parameters, one can calculate what the posterior distribution of the values should be, fixing all other parameter values. These parameters can be Gibbs sampled by directly drawing a new parameter from the calculated posterior distribution. This method of parameter sampling can save time since the new value is always accepted, as compared to the Metropolis-Hasting method above where some updates are rejected.

The Markov chain Monte Carlo method is a Bayesian method, however other methods such as Maximum Likelihood could have been used. Maximum Likelihood methods attempt to fit a model to data by finding the values of parameters of a model that

produce the maximum likelihood given the data. Maximum Likelihood methods are generally computationally easier to implement and less costly in terms of time to run, however they have some drawbacks. Biases have been observed in Maximum Likelihood ancestral reconstructions, whereas Bayesian inference was much less biased in important reconstructions [60]. For these reasons, the Markov chain Monte Carlo method was used.

The phylogenetic tree was fixed and the most likely substitution history was sampled with a simpler model, mtMam, using PLEX [87, 141]. The propensities are estimated using this sampled substitution history. The benefits of using this method are that a new ancestral state sampler does not need to be written to take into account the more complex model parameters. The substitution history can be produced quickly using the simpler model. In order to integrate over uncertainty in the substitution history, one could randomly select a substitution history from the set produced by the simpler model, rather than use the most likely substitution history, and fit the more complex model parameters to the selected substitution history. Given that the ancestral states are considered known for this analysis, the only unknowns are which amino acids are acceptable on each branch and the model parameters.

The fitness switching rate and substitution rate are sampled using a normal distribution with a fixed standard deviation and accepted according to a Metropolis-Hastings algorithm [173, 174]. The standard deviation of sampling could be tuned by hand or automatically to optimize mixing. A value of 0.01 was found to work in practice for both sampling standard deviations. The set of acceptable amino acids along each lineage was inferred using a Gibbs sampler [175].

The prior probabilities of different sized acceptable sets (ρ) were fixed as well (See table V.1). The chosen values are approximately proportional to the relative numbers of sizes of acceptabilities found in the mitochondrial data set using flat priors.

Acceptable Set Size	Prior Probability
0	0.00001
1	0.5
2	0.05
3	0.001
>3	0.0001

Table V.1: The unnormalized values of the prior probabilities of the size of acceptable sets ρ used to fit the Acceptability model and acceptable amino acid sets in the simulations and mitochondrial analysis. The left column indicates the different acceptable set sizes and the right column is the prior probability of that size of acceptable set. For each size greater than three up to a size of 20, the prior probability is 0.0001. These values were derived from all sites in the complete mitochondrial protein data set.

V.4.4 Propensity Shift Calculation

Detecting large shifts in propensities requires first calculating the propensities. The amino acid propensities for each site and branch are calculated from the posterior probabilities of each amino acid being in the high fitness state. The vector of posterior probabilities is normalized such that the sum of the posteriors over all amino acids equals one. Then the propensity difference is calculated from the propensities on the ancestor node of the branch to the propensities on the descendant node of the branch. The Euclidean distance (D) was calculated using the two vectors of propensities. The Euclidean distance is the square root of the sum of squares of the differences between propensities of each amino acid in the two vectors.

$$D_{i,s} = \sqrt{\sum_{y} (\Pi^{y}_{i,s-1} - \Pi^{y}_{i,s})^2} \tag{V.8}$$

where i is the site, s is the species node, $s-1$ is the node ancestral to node s, y is an amino acid, $\Pi^{y}_{i,s}$ is the propensity of amino acid y in species s at site i, $\Pi^{y}_{i,s-1}$ is the propensity of the same amino acid y in the ancestral species $s-1$ at site i.

These distances were collected then divided into two different populations based on whether a substitution was found on the same branch segment at either adjacent site. Distances with adjacent substitutions for the mitochondrial data are shown in Figure

V.3 and distances without adjacent substitutions are in Figure V.2. A histogram of all differences independent of whether a substitution occurred nearby or not is shown in Supplemental Figure V.1. Propensity shifts for the nuclear encoded glycolysis protein data are shown in Supplementary Figures V.16 and V.17.

The R code for calculating the population sizes with and without substitutions and running the Kolmogorov–Smirnov test on data sets where propensity shifts are possibly influenced by sites which are 1, 2, 3, 10, and 100 positions away is shown below.

```
directories = c("../data/mito_2609_2/jumps1/",
                "../data/mito_2609_2/jumps2_2/",
                "../data/mito_2609_2/jumps3_2/",
                "../data/mito_2609_2/jumps10/",
                "../data/mito_2609_2/jumps100/")

for (directory in directories){
    substitution_dataframe = read.table(paste0(directory,
                                        "sub_differences.data"))
    print(paste("Propensity shifts with substitutions population size",
          nrow(substitution_dataframe)))
    no_substition_dataframe = read.table(paste0(directory,
                                         "no_sub_differences.data"))
    print(paste("Propensity shifts without substitutions population size",
          nrow(no_substition_dataframe)))
    # Run the Kolmogorov{Smirnov test
    ks.test(substitution_dataframe$V1, no_substition_dataframe$V1)
}
```

V.4.5 Comparison with Previous Methods

The evolution model used in this method shares some aspects in common with the covarion models [176, 177, 46, 178]. In a covarion model, sites switch between being able to substitute (variable) and not being able to substitute (fixed). This fixed or variable hidden state is allowed to change over the tree and across sites, affecting when sites are allowed to substitute or not. The Acceptability model similarly changes across sites and the tree and effectively restricts when substitutions occur to specific sites and times on the tree: when only one amino acid is acceptable, substitutions are not allowed. The covarion model does not model directly the amino acid propensities changing over time, as the Acceptability model does.

A similar model was proposed by Usmanova in 2015 and the model used in this study

can be seen as a more realistic alternative to it [57]. Usmanova has proposed a model that uses the same binary fitnesses and allows fitnesses for each amino acid to change over time [57]. Amino acids move among six different hidden states: "Current" or the resident amino acid acid at a site on a branch, "Allowed" or high fitness, "Blocked" or low fitness, and "Forbidden" or always low fitness. The allowed and blocked amino acids are further split into "near" and "far" based on whether they are a single nucleotide substitution away from the current amino acid. Non-current amino acids can switch between allowed and blocked freely along the tree and across sites, while changes in near and far only occur when there is a substitution [57]. The rate of switching between allowed and blocked was estimated on quadruplet sets of aligned sequences.

The model proposed here has a few advantages over the model proposed by Usmanova. It is designed to be used in phylogenetic analysis of full tree and multiple sequence alignment data sets. Rather than the hidden state having six different possible states (current, forbidden, allowed near, allowed far, blocked near, blocked far), the Acceptability model only allows two states corresponding to a binary fitness (acceptable or unacceptable) and completely separates the state into the resident amino acid and the acceptable set of amino acids. The current amino acid is not considered a hidden state of the model. The substitution rates and fitness switching rates are distinct processes from each other. The Acceptability model also is flexible with respect to the substitution process used. Any process which defines the substitution among amino acids of equal fitness may be used. Since this model uses a binary fitness and does not allow substitutions across different fitness levels, the substitution process provided must be among equal fitness amino acids.

The most similar model previously studied was proposed by Usmanova in 2015 [57], however there are several important distinctions between this study and the previous. First, the model proposed by Usmanova distinguishes between an amino acid at a site at a time in the past as being temporarily low fitness ("Blocked") and permanently low fitness ("Forbidden"). Since all of the amino acids seen in the extant sequences

are the majority alleles, every observed amino acid observed is high fitness. Therefore, there is no information to distinguish temporarily and permanently low fitness residues. Secondly, since fitness acts on the amino acid encoded by a codon and many third position transitions encode for the same amino acid, it is impossible to truly distinguish between A_n ("Allowed near") and A_f ("Allowed far") for many codons for much of the ancestral history.

Finally, the Acceptability model is simpler, because it works only on amino acids and does not attempt to infer codons about which we have very limited information. One amino acid can be encoded by more than one nucleotide triplet codon, thus the fitness of different codons that encode for the same amino acid are most likely identical if fitness is based on amino acid fitness. If one does take into account the codon structure, then one should incorporate known information about codon biases [179, 180, 181], which Usmanova did not do. There may even be a distribution of codon usage within each species all translating to the same amino acid, further complicating the usage of a single codon for each species, while the amino acid analysis remains simple.

Another study recently was published with a goal similar to this study: to identify large changes in propensities across a phylogenetic tree [163]. The CAT-BP model was used to identify break points where support for categories of specific sites changed from one category to another. Apart from the model used to calculate the propensities, these studies have many differences. This study was run on all 13 protein coding genes in the mitochondrial genome, while the former study only considered the Influenza PB2 gene.

The CAT-BP method has some advantages over the method proposed here and some disadvantages. The CAT-BP method uses nucleotide alignments and codon structure as part of the mutation model, while here I use only amino acid information for each site. Codon structure may be an important factor in evolutionary models, offering the CAT-BP model a potential advantage over the Acceptability model. Since the CAT-BP method is a combination of previously described and studied methods, the tools for

evaluating the parts of the method already had been developed, making evaluation more straightforward. The method used here allowed for the propensities to change anywhere on the phylogenetic tree, while the former study only searched for propensities changing at the division between human and avian hosts. This allows for the Acceptability model to be more flexible at the expense of evaluating time.

The CAT-BP model could recover a reasonable tree given the mitochondrial sequences from 20 arthropods in a previous study, despite the CAT and BP models individually producing erroneous trees [52]. This demonstrates that the CAT-BP model is a robust model to reconstruct trees from and may be a more apt description of the evolutionary process, since it allows for time- and site-heterogeneity. The results from both methods generally agree that the amino acid propensities may differ dramatically across sites and time, however this study also investigates whether substitutions are highly correlated with larger propensity changes at adjacent sites, providing a rationale for why propensities change in the first place.

V.4.6 Simulations

In order to test whether acceptabilities can be correctly inferred from sequences evolved under changing acceptabilities, I ran two different simulations. The first simulation produces its own changing acceptabilities de novo and the second simulation is based on the acceptabilities derived from the mitochondrial data set.

In the first simulation, the acceptabilities are simulated across a tree using the parameters provided in the model. A random 100 taxon tree was generated using T-REX, and the branch lengths were normalized so that the average is similar to the mitochondrial branch length average of 0.045 [182]. A random amino acid is chosen for each site in the sequence at the root of the tree, and that amino acid is added to the acceptable set for the site. The tree is segmented using the same B1xM method above using a maximum branch length of 0.05. Then the acceptabilities are allowed to change along each branch and the amino acid at the end of each branch is sampled given the ancestral amino acid

and the acceptabilities at the ancestor and descendant. The process is repeated for every site and for every branch segment, up to 100 sites. The switch rate ν and the substitution rate λ were chosen such that P_ν and P_λ were 0.01 and 0.01 respectively.

I wanted to simulate sequences that were similar to the mitochondrial sequences and had similar sized acceptable sets, therefore I introduced a prior on the size of the acceptable sets, ρ. A prior against having many acceptable amino acids was chosen, using the same values as described in table V.1. Without this prior, the number of acceptable amino acids at any point in time increased to around 10 amino acids, which is much higher than the size of the acceptabilities inferred from the mitochondrial data.

In the second simulation, sequences were evolved using the inferred mitochondrial acceptabilities, in order to capture the complexity of how the mitochondrial acceptabilities changed over time. The first 100 sites of the mitochondrial alignment were used. Since the acceptability fitting produced posterior probabilities of amino acids being acceptable at every branch segment, I thresholded each posterior in order to define the acceptabilities. The posterior probabilities for each amino acid being acceptable at any branch segment were thresholded such that if the posterior was above 0.3, the amino acid was considered acceptable at that branch segment. The threshold of 0.3 was chosen because it is far above the noise of the amino acids which have low support for being acceptable (see Figure V.11) and it reduces the possibility of having zero acceptable amino acids at any site or branch segment on the tree. For each branch segment and site, a random amino acid was chosen from the acceptable amino acids. Nearly all of the branch segments had only one acceptable amino acid at each site, however occasionally two or three amino acids were acceptable. In this way, I did not have to artificially specify the values of the rate parameters to simulate under, but rather allow the values to be inferred alongside the acceptabilities.

In order to compare the simulated acceptabilities and the fitted acceptabilities for both simulations, the acceptabilities of the simulated sequences and model parameter values

were estimated using the above method. Then the inferred acceptabilities at each site and branch segment were compared against the correct acceptabilities that the sequences were simulated under. The correct and inferred acceptabilities are compared by considering each amino acid in the correct set of acceptabilities and checking if it is in the inferred set of acceptabilities. Any amino acid that was not inferred correctly was counted as a false negative or missed acceptability. Then any extra amino acids in the inferred set that were not in the correct set are also counted and considered a false positive.

V.5 Results

V.5.1 Detecting Propensity Shifts in the Mitochondrial and Nuclear Data

In the mitochondrial data, evidence was found that a substitution at a site causes the shifts in amino acid propensities at adjacent sites to be larger than if there were no substitution at the site. Sites with at least one substitution at an adjacent site in the alignment had statistically significantly larger propensities shifts on average than sites with zero substitutions at adjacent sites. A histogram of the propensity shifts given no adjacent substitutions is shown in Figure V.2 and a histogram of the propensity shifts given at least one substitution at adjacent sites is shown in Figure V.3. A histogram of all differences independent of whether a substitution occurred nearby or not is shown in Supplemental Figure V.1.

Large propensity shifts were also detected in the nuclear sequences. Propensity shifts for the nuclear encoded glycolysis protein data are shown in Supplementary Figures V.16 and V.17, indicating that large propensity shifts are not a mitochondrial genome specific phenomenon.

Substitutions increase the amino acid propensity shift size of adjacent sites. The average shift in propensities given an adjacent substitution is 0.171, while the average shift in propensities without an adjacent substitution is 0.082. The overall average shift independent of whether a substitution occurred at an adjacent site is 0.085.

The size of a propensity shift from 100% of one amino acid to 100% of another amino

acid is $\sqrt{2}$ or 1.41. If the propensities shift from 100% of one amino acid to 50% of the same amino acid and 50% of a new amino acid, then the propensity shift would be around 0.71.

A simple way to detect the difference of the propensity shifts between distributions is to consider what percent of the propensity shifts are larger than 0.5. This indicates how many of the shifts are large and how many are close to zero. One can calculate this percentage for the propensity shifts given a substitution and compare that number against the percentage above 0.5 for the propensity shifts given no substitution in order to see the effects of a nearby substitution has on the propensity shifts.

In order to determine if the distributions of propensity shifts with and without adjacent substitutions are drawn from the same distribution, the Kolmogorov–Smirnov test was applied with the propensity differences given an adjacent substitution as one population and the propensity differences given no adjacent substitution as the other population. The probability of calculating mean values at least this extreme given the two populations derive from the same distribution is less than $2.2\text{x}10^{-16}$, indicating that there is a statistically significant difference between the distribution of propensity shifts with and without adjacent substitutions. Since the mean of the shifts with a substitution is substantially higher than the mean without a substitution, we can conclude that there is a significant increase in the average propensity shift when there are nearby substitutions.

To determine whether close substitutions correlated with larger shifts in propensities more than distant substitutions, the above analysis was repeated considering substitutions at 2, 3, 10, and 100 positions away from the site with propensity shifts. The propensity shifts were separated into two categories, large and small, depending on whether the shift was greater than 0.5. The threshold of 0.5 was chosen because it cleanly separates the distribution of shifts near zero and the shifts around 1.4. The fraction of the propensity shifts above 0.5 given a substitution was calculated for the various number of sites between propensity shifts and substitutions and is shown in Figure V.4. There is a clear decrease in

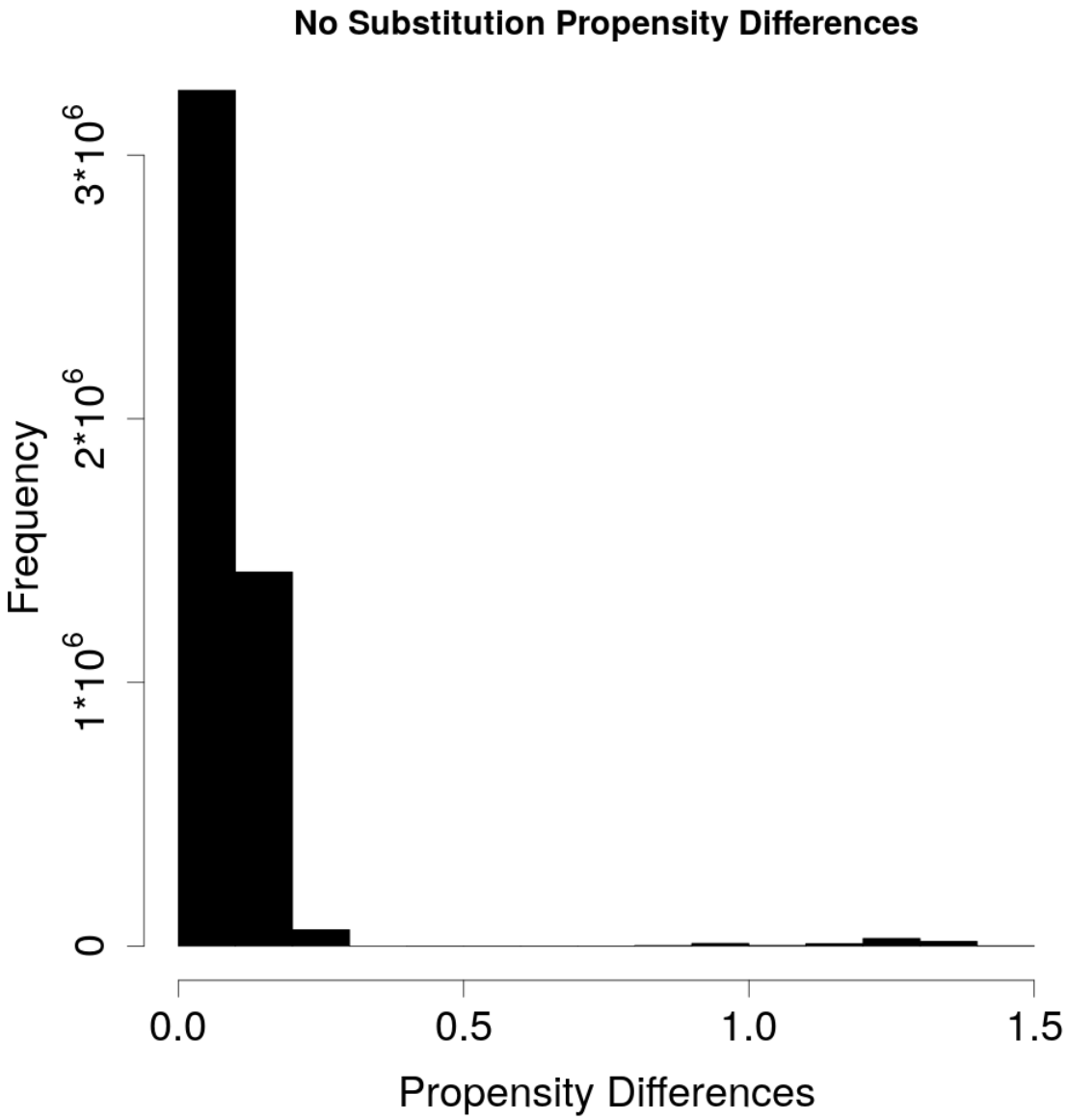


Figure V.2: A histogram showing the distribution of amino acid propensity shifts given that no substitutions were detected at adjacent sites in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41.

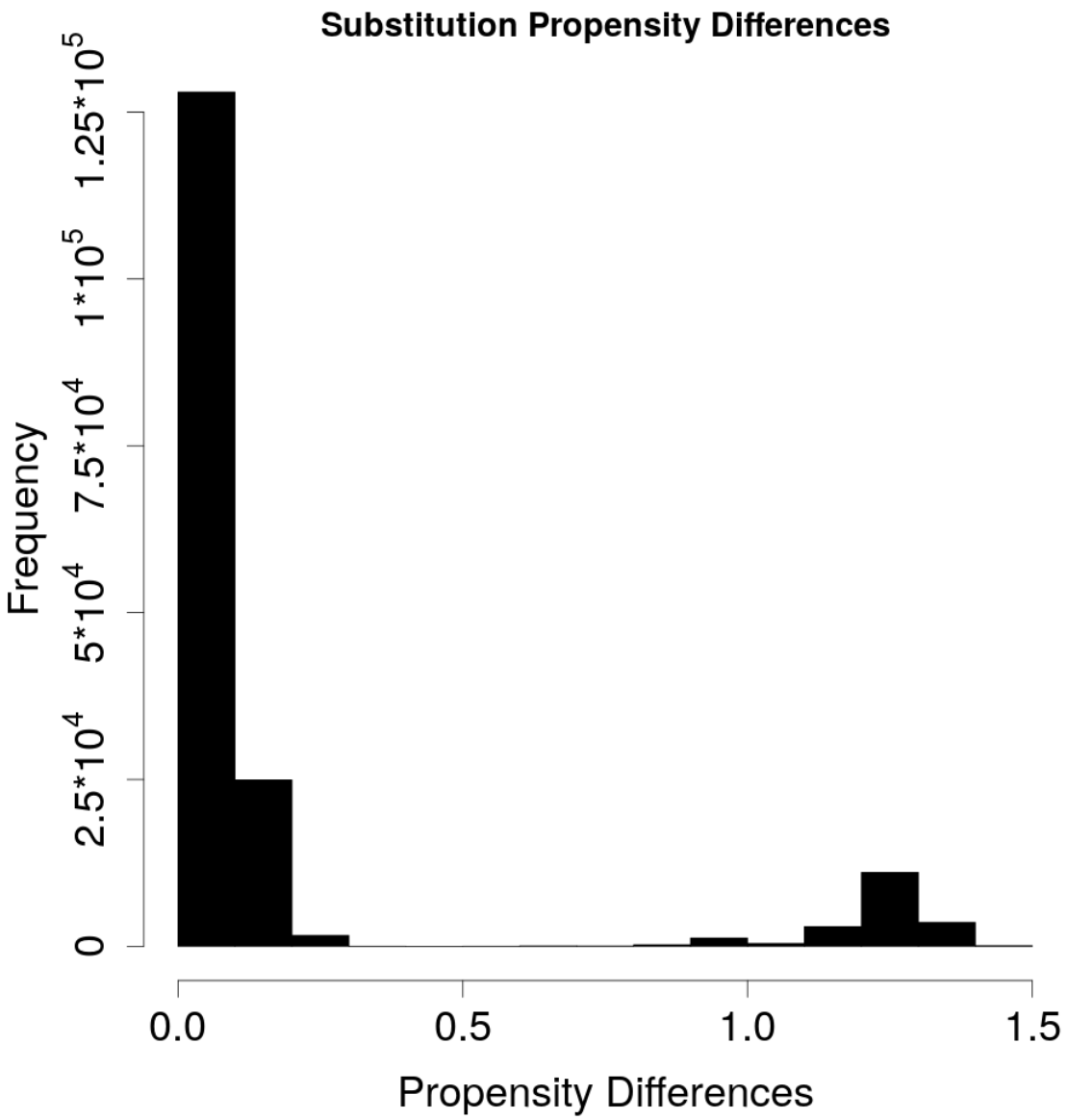


Figure V.3: A histogram showing the distribution of amino acid propensity shifts given that at least one substitution was detected at adjacent sites in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8, as in Figure V.2. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution is compared against the distribution given zero substitutions at adjacent sites using the Kolmogorov–Smirnov test and the probability that the two distributions resulted from the same underlying distribution is less than $2.2\text{x}10^{-16}$, indicating that amino acid substitutions substantially increase the propensity shifts at adjacent sites in the multiple sequence alignment.

Sites between propensity and substitution	Fraction > 0.5 given substitution	Fraction > 0.5 given no substitution	Sub / No Sub Ratio
1	0.1135	0.015	7.54
2	0.0976	0.0156	6.24
3	0.1129	0.0151	7.48
10	0.0884	0.0159	5.55
100	0.0672	0.0168	4.01

Table V.2: The fraction of propensity shifts above 0.5 split by the number of steps away from a site and whether a substitution occurred or not.

the proportion of large shifts between substitutions at adjacent sites (bar 1 in Figure V.4) and substitutions at sites 100 positions away (bar 5 Figure V.4). A table summarizing the fractions of the propensity distributions above 0.5 is shown in Table V.2. This suggests that closer substitutions correlate with larger propensity shifts than do further substitutions. It is surprising that effects of substitutions were seen in propensities at sites as far away as 10 or 100 sites away.

When calculating the propensity shifts above and below 0.5, many of the shifts were around 0. This corresponds with our understanding that the mitochondrial genome is strongly conserved and many sites are unable to substitute to any other amino acid, at least over the time period covered by our vertebrate data set. After applying the Kolmogorov–Smirnov test, the distributions considering substitutions at 1, 2, 3, 10, and 100 sites away from the propensity calculation all had p-values of less than $2.2\text{x}10^{-16}$ of being from the same underlying distribution.

Specific examples of sites in which an adjacent substitution correlates with major shifts in the amino acid propensities of a site are numerous. One example is a large shift in propensities at node 432 and site 1030 which is correlated with a substitution at the adjacent site 1031. Node 432 corresponds to a species that is ancestral to the birds. The mitochondrial phylogenetic tree colored by amino acid for sites 1030 and 1031 are shown in Figure V.5 and Supplemental Figure V.2. Figure V.6 shows the resident amino acids on the tree for both sites 1030 and 1031 for the clade below node 432. The substitution

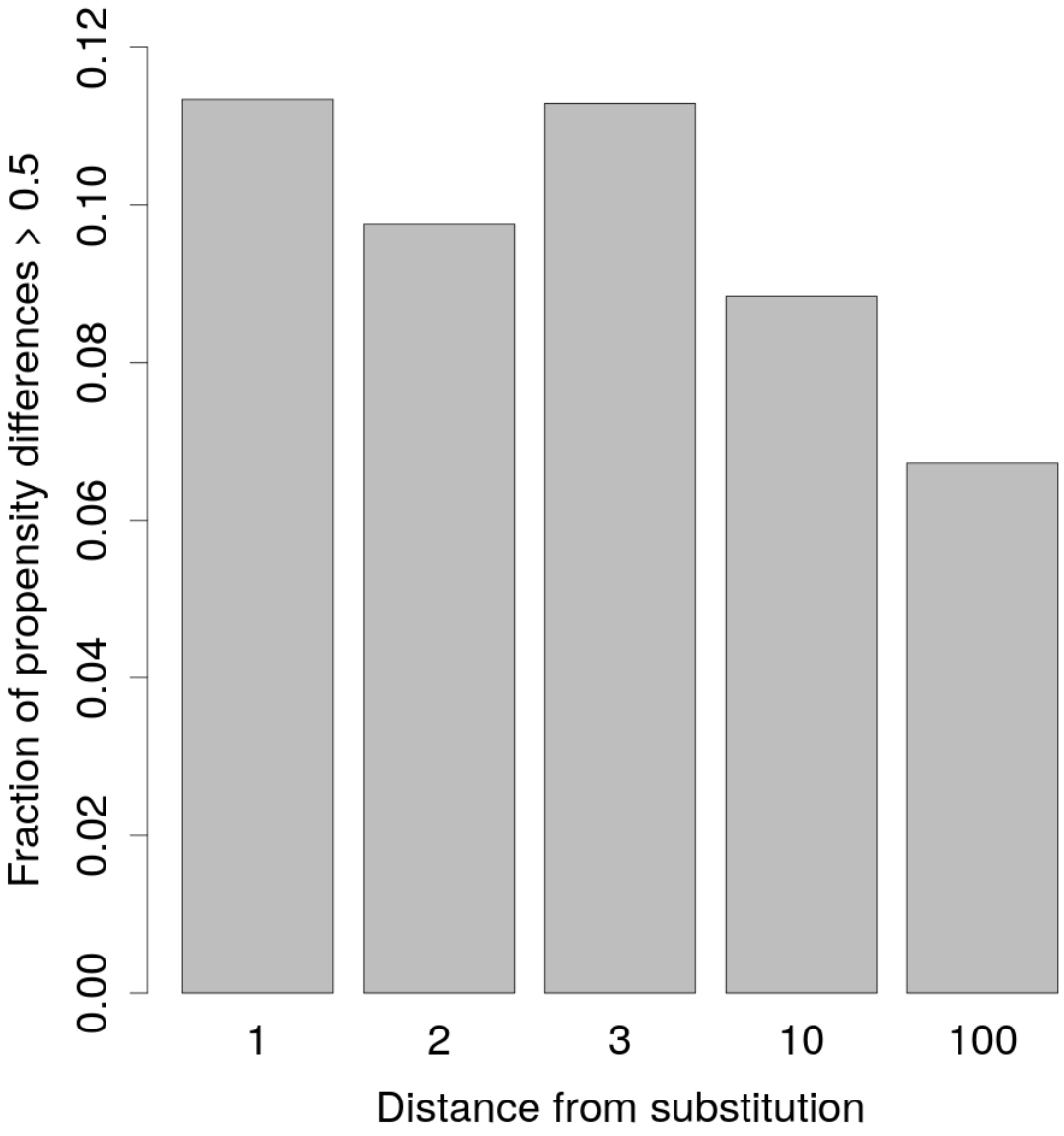


Figure V.4: A histogram showing the proportion of the amino acid propensity shift distribution given a substitution that is above 0.5 for multiple different steps away from a site.

from isoleucine to leucine at node 432 at site 1031 correlates on the tree with a amino acid propensity shift of magnitude 0.60 at site 1030 at the same node 432. This shift is primarily due to changes in the propensities of methionine and threonine. In the ancestor of node 432, methionine is strongly favored 99% compared to threonine's 1%. The node 432 has a 57%-43% split between methionine and threonine, indicating that the propensity for methionine has decreased and the propensity for threonine has increased drastically at site 1030. The shift resulted in substitutions at site 1030 at the direct descendants of node 432 from methionine to threonine, however the 11 reversions to methionine further down in the clade support that methionine is probably acceptable throughout the clade.

Site 1251 shows a similarly large propensity shift at node 432 with a substitution at the adjacent site 1250. A substitution from isoleucine to leucine at site 1250 correlates with the change in propensities at site 1251 from 99% isoleucine and 1% threonine to 52% isoleucine, 46% threonine, and 2% phenylalanine. The size of the amino acid propensity shift along the branch above node 432 is 0.65. Figure V.7 shows the phylogenetic tree of site 1251 colored by resident amino acid. The substitution at node 432 at site 1250 can be seen in Supplemental Figure V.3 surrounded by a black box. The subsequent substitutions from isoleucine to threonine at site 1251 at the descendants of node 432 could only occur if the fitness of threonine increased to near or above the fitness of isoleucine. The substitution at site 1250 probably had at least some influence on the change in propensities of site 1251 at node 432.

Node 841 is an ancestral mammal which has a large shift in the amino acid propensities at site 1376 and a substitution at the adjacent site 1375. The resident amino acids are shown on the mitochondrial phylogenetic tree in Figure V.8 for sites 1376 and Supplemental Figure V.4 for sites 1375. The size of the amino acid propensity shift is 0.74, shifting from 100% valine at the ancestor to node 841 to 48% valine and 52% isoleucine at node 841. The substitution from methionine to isoleucine at ancestral node 841 in site 1375 correlates with the shift in the propensities of the amino acids at site 1375 at ancestral

node 841. Multiple substitutions from isoleucine to valine and vice versa are observed in the clade directly below node 841, giving evidence that both amino acids are acceptable after the propensity shift.

V.5.2 Acceptable Set Inference in Mitochondrial Data

The Markov chain Monte Carlo fitting the Acceptability model to the mitochondrial data can infer when amino acids are acceptable and when they are unacceptable for each site and on every branch of the vertebrate mitochondrial tree. For most evolution along each site, only a single amino acid is acceptable, corresponding to the resident amino acid being so stable relative to the other amino acids that no substitutions can occur. When a substitution occurs, both the ancestral and descendant are acceptable along the branch with the substitution, but interestingly the posterior probability of the descendant amino acid being acceptable increases in the branch segments prior to the substitution (see Figure V.9). Correspondingly, I can see that the probability of the ancestral amino acid being acceptable decreases in simulations that allow propensities to change over time [48].

This model could possibly detect when an amino acid becomes acceptable before convergences to that amino acid are seen on the tree. There are many examples of sites in the mitochondrial data set where a clade has many convergences to the same amino acid. One real example from threonine to alanine is shown in Figure V.1. In this example, alanine may have become acceptable at the root of the clade in box A, and remain acceptable throughout the whole clade, resulting in a few substitutions to it. This result makes sense with our intuition that it is more likely that an amino acid becomes acceptable once at the ancestor of all the convergences, rather than the same amino acid becoming acceptable many times independently.

The Acceptability model can correctly infer when an amino acid becomes acceptable prior to a few convergence events to that amino acid, as shown in Figure V.10. Using the model, I can reasonably infer that methionine (green) became acceptable early in the

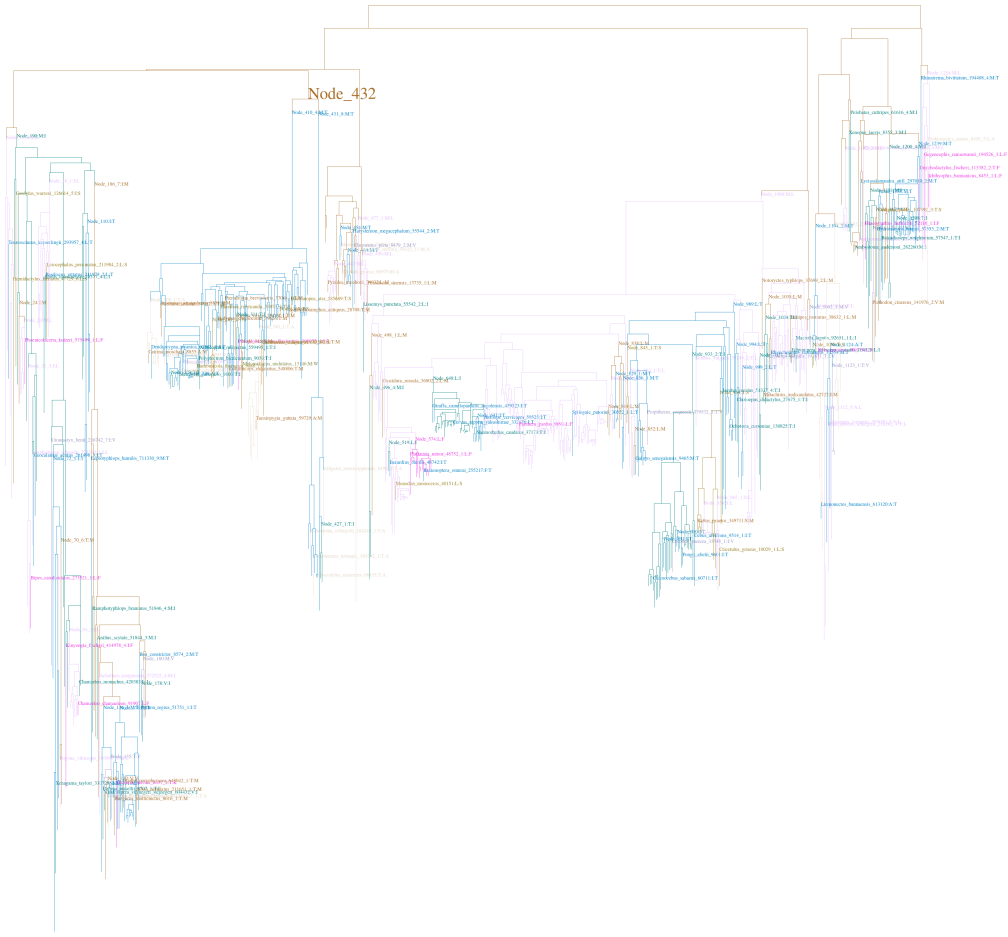


Figure V.5: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1030 of the amino acid multiple sequence alignment. The branches where methionine is resident are colored orange, and the branches where threonine is resident are colored light blue. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. For example the label Node_410_4:M:T indicates that a substitution from methionine to threonine occurred at the branch directly ancestral to Node_410_4. The amino acid propensities shift along the branch directly above ancestral node 432 from heavily favoring methionine (99% methionine and 1% threonine) at the ancestral node to a 57%-43% split between methionine and threonine at node 432. The shift resulted in substitutions at site 1030 at the direct descendants of node 432 from methionine to threonine, however the 11 reversions to methionine further down in the clade support that methionine is probably acceptable throughout the clade. A substitution at ancestral node 432 in site 1031 from isoleucine to leucine correlates with the shift in the propensities of the amino acids at site 1030 at ancestral node 432. The colored phylogenetic tree of site 1031 is shown in Supplemental Figure V.2. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.

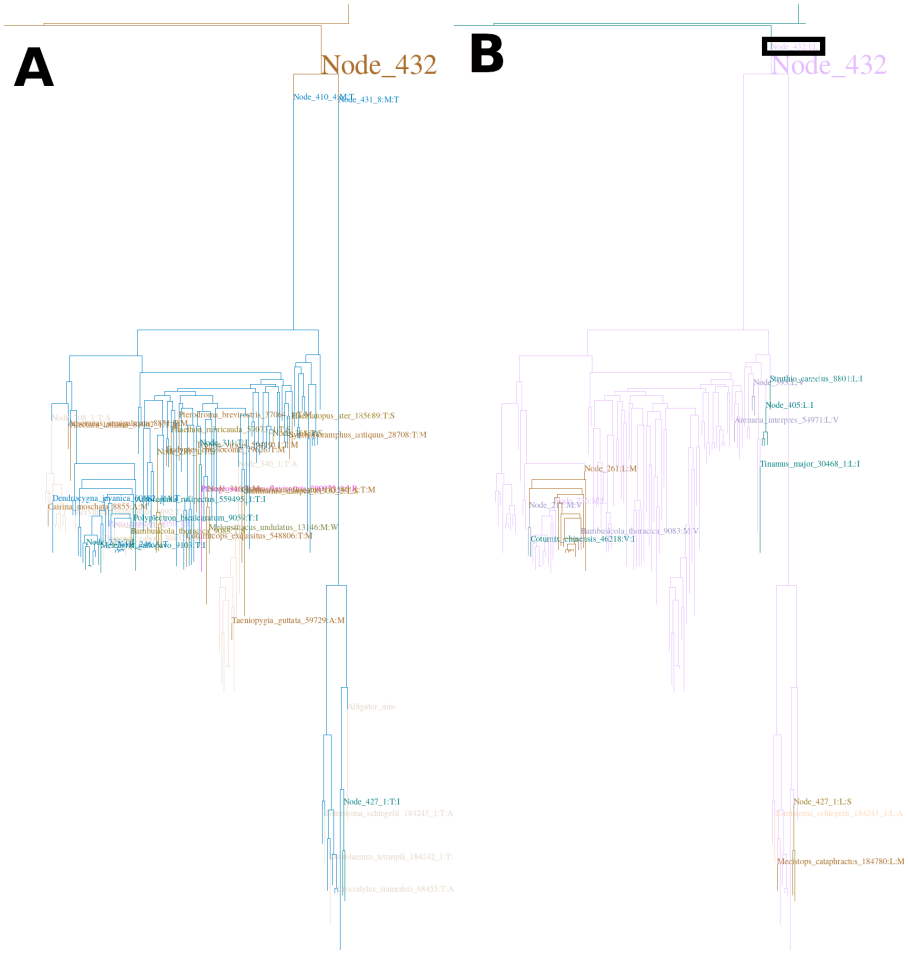


Figure V.6: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1030 (part A) and 1031 (part B) of the amino acid multiple sequence alignment. Site 1031 is adjacent to site 1030 in the amino acid alignment. The branches are colored based on which amino acid is resident at that branch: methionine is orange, threonine is light blue, isoleucine is dark blue, and leucine is pink. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. For example the label Node_410_4:M:T indicates that a substitution from methionine to threonine occurred at the branch directly ancestral to Node_410_4. The amino acid propensities shift along the branch directly above ancestral node 432 from heavily favoring methionine (99% methionine and 1% threonine) at the ancestral node to a 57%-43% split between methionine and threonine at node 432. The shift resulted in substitutions at site 1030 at the direct descendants of node 432 from methionine to threonine, however the reversions to methionine further down in the clade support that methionine is probably acceptable throughout the clade. The substitution from isoleucine to leucine at ancestral node 432 in site 1031 (shown in a black box) correlates with the shift in the propensities of the amino acids at site 1030 at ancestral node 432. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.

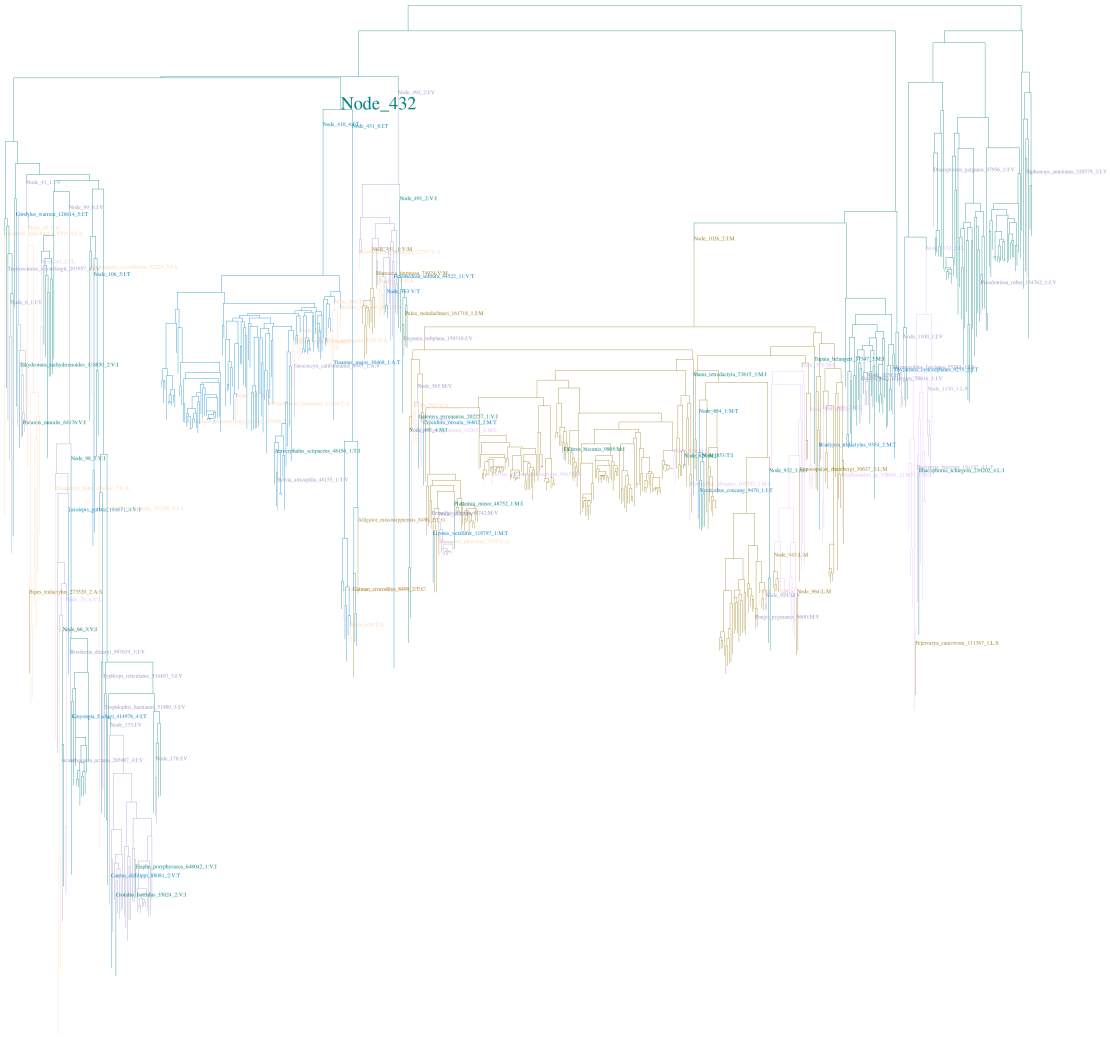


Figure V.7: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1251 of the amino acid multiple sequence alignment. The branches where isoleucine is resident are colored dark blue, and the branches where leucine is resident are colored pink. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. The label Node_432:I:L indicates that a substitution from isoleucine to leucine occurred at the branch directly ancestral to Node_432. A substitution from isoleucine to leucine at ancestral node 432 in site 1250 correlates with the shift in the propensities of the amino acids at site 1251 at ancestral node 432 from 99% isoleucine and 1% threonine to 52% isoleucine, 46% threonine, and 2% phenylalanine. Site 1251 is adjacent to site 1250 in the amino acid alignment. The colored phylogenetic tree of site 1250 is shown in Supplemental Figure V.3. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.

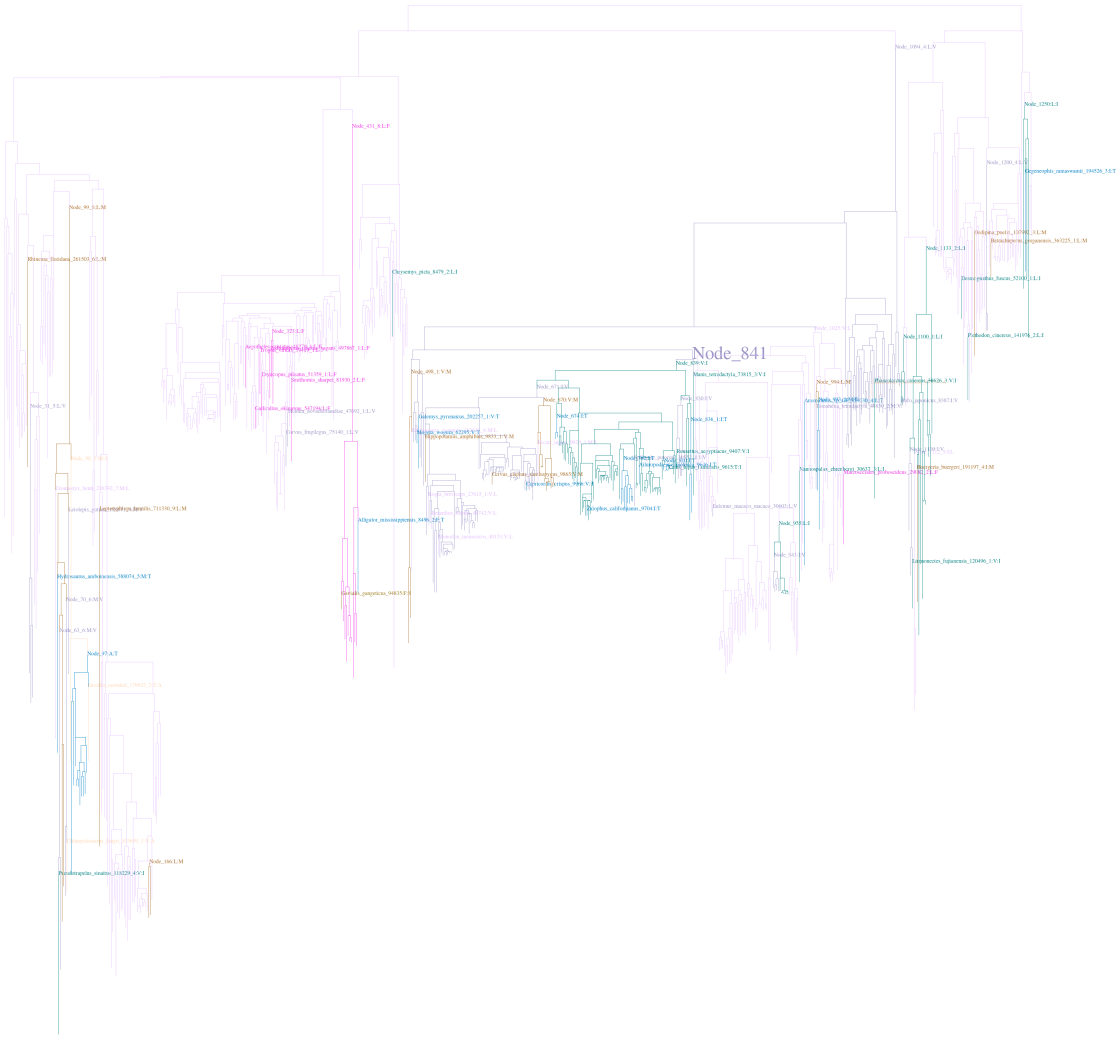


Figure V.8: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1376 of the amino acid multiple sequence alignment. Site 1376 is adjacent to site 1375 in the amino acid alignment. The branches where valine is resident are colored purple and the branches where isoleucine is resident are dark blue. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. The label Node_432:I:L indicates that a substitution from isoleucine to leucine occurred at the branch directly ancestral to Node_432. The substitution from methionine to isoleucine at ancestral node 841 in site 1375 correlates with the shift in the propensities of the amino acids at site 1376 at ancestral node 841. The propensities shift from 100% valine to 48% valine and 52% isoleucine. The colored phylogenetic tree of site 1375 is shown in Supplemental Figure V.4. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.

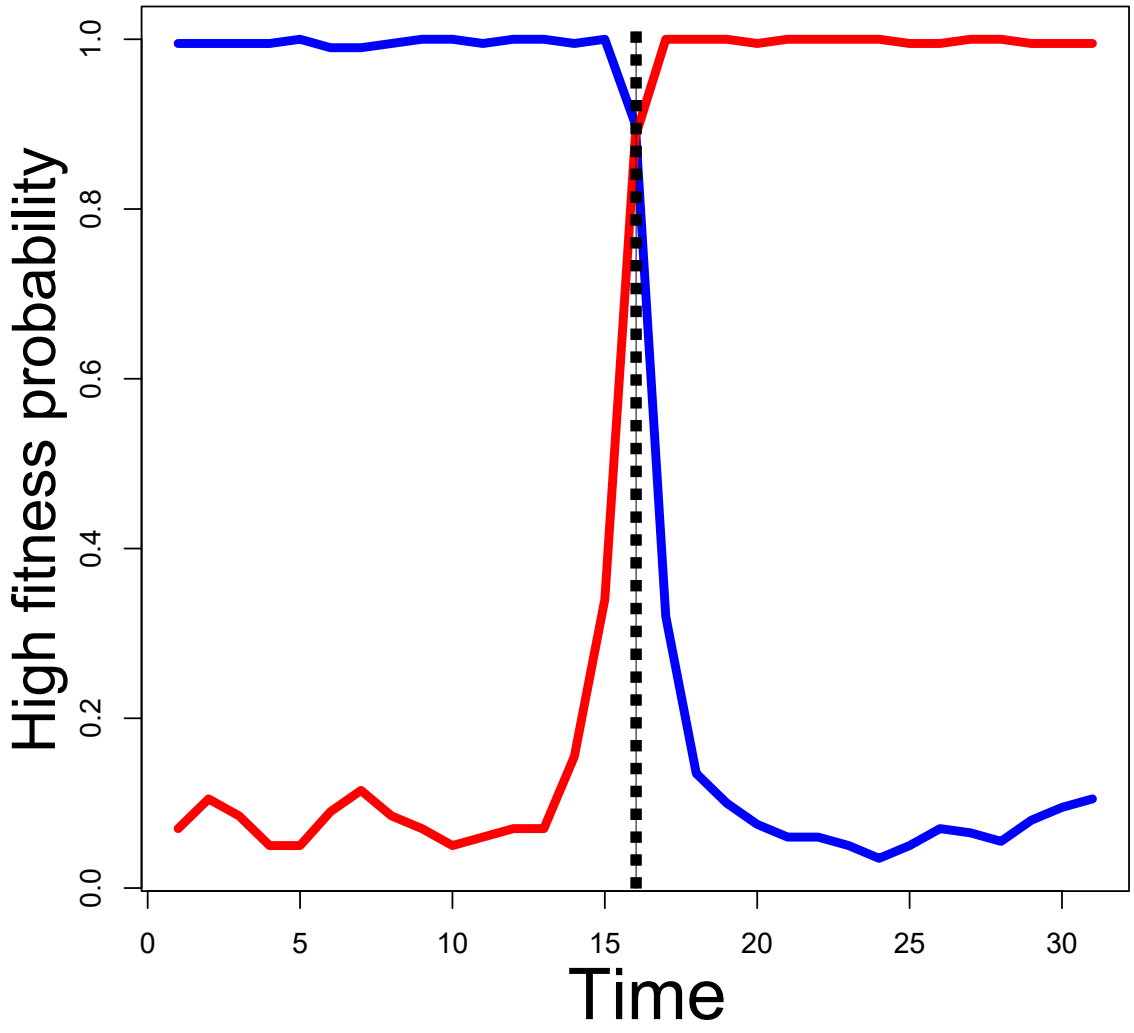


Figure V.9: The posterior probabilities of high fitness of the ancestral and descendant amino acids around a substitution in blue and red respectively. The time of substitution is shown with a dotted vertical line.

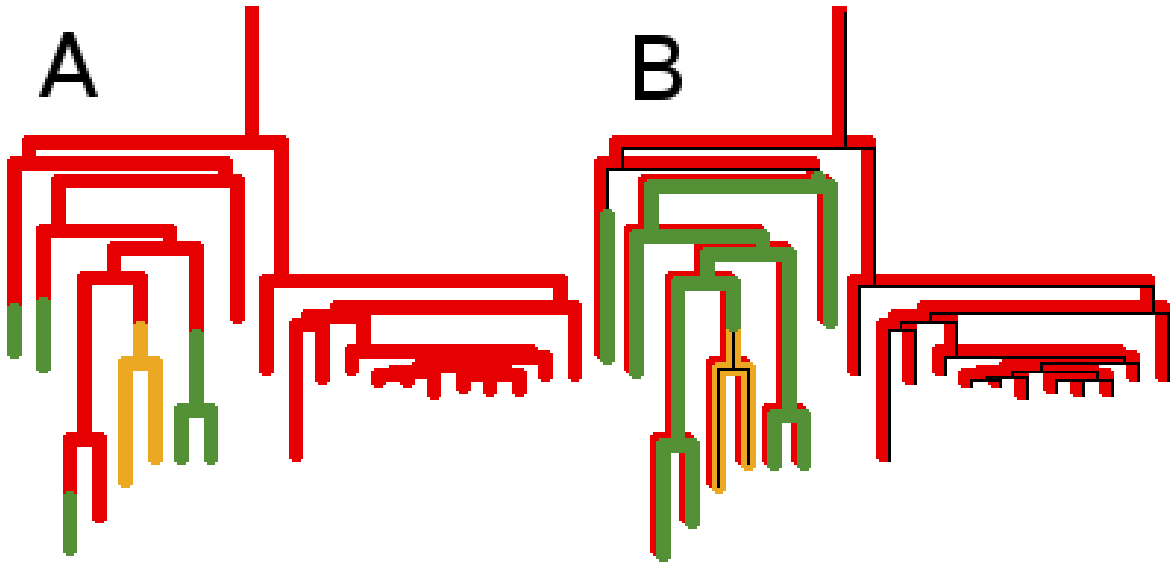


Figure V.10: A real example of preadaptation before a number of convergences from site 549 from the sequence alignment in a squamate clade of the mitochondrial data set. Part A shows the ancestral reconstruction for the site with threonine, methionine, and alanine in red, green, and orange respectively. Part B shows when each amino acid was inferred to be acceptable by the model. Where multiple colors overlap along the tree, multiple amino acids are acceptable.

left-most clade, which resulted in multiple convergences later. It is also probable that threonine (red) did not become unacceptable in the left clade, given that a few leaves still had threonine at this site.

In order to speed up the computation, the 2609 site data set was split into 26 separate groups of 100 sites each, except for the last group which had 109 sites. The substitution and switch probabilities (P_λ and P_ν) were fixed at 0.1 for testing purposes. Each group was fit with the model for 260,000 acceptability set updates. An example likelihood trace is shown in Supplemental Figure V.5. The chain burns in after around 26,000 updates, resulting in 90 independent samples from the posterior for each acceptability. Each sample is independent because every amino acid acceptability set for every branch segment is sampled every generation, adding to the computational cost substantially.

I can probe how well the model fits different levels of constraint to the mitochondrial data. A bar plot showing how often an acceptable amino acid with a certain size

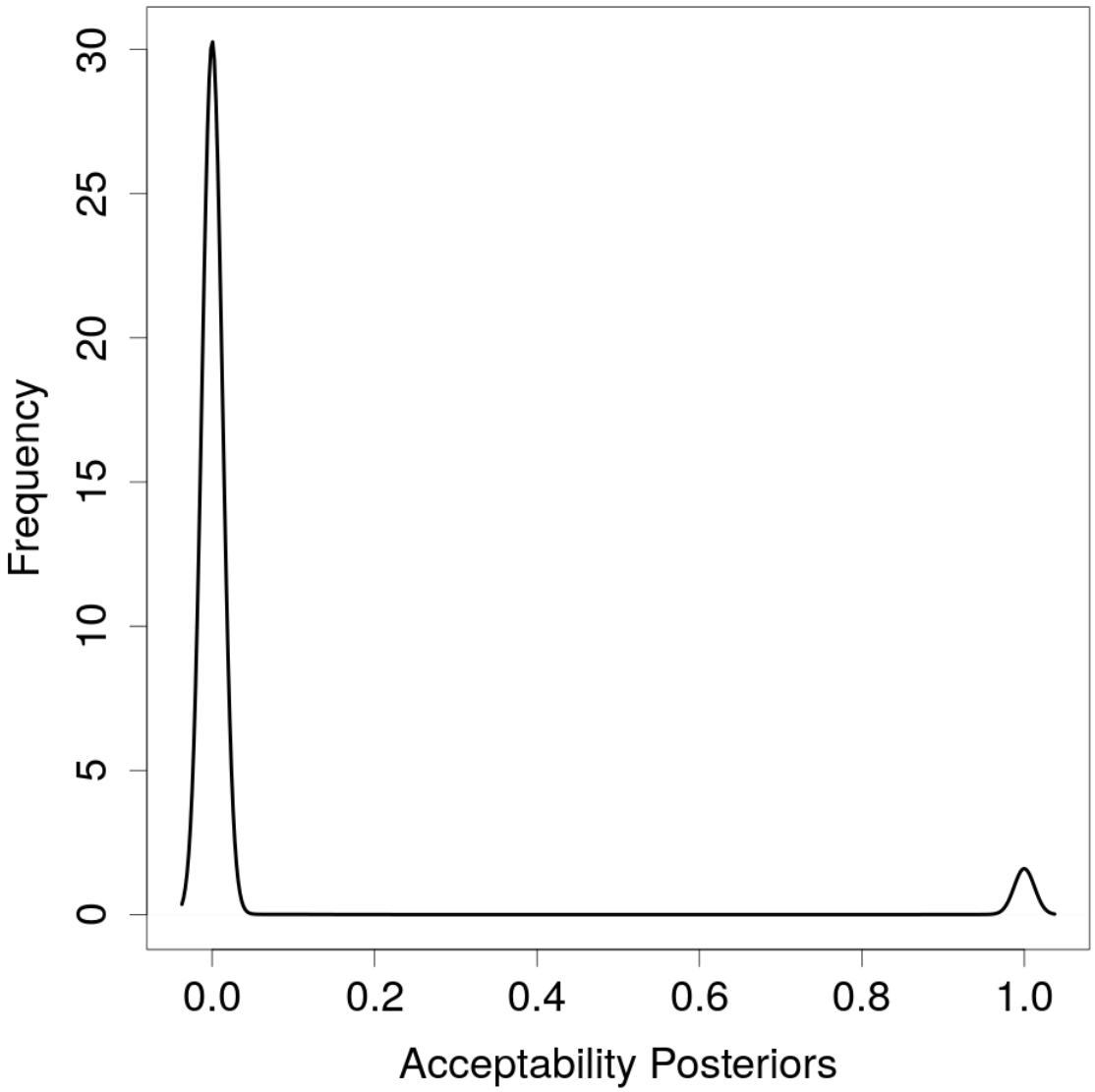


Figure V.11: The distribution of posteriors of amino acids being acceptable averaged across all sites and branch segments. Many amino acids have around a zero propensity for many sites and times along the tree. This observation agrees with the expectation that only a few amino acids are acceptable at any specific site and point in time. When an amino acid becomes completely fixed for some period of time, the propensity for that amino acid can be very high, around 100%. This is a fairly frequent phenomenon, as shown by the high density around 100%.

was inferred is shown in Figure V.12. The instantaneous constraint averaged across sites and the tree is around 1.57 acceptable amino acids at a time. This shows that the mitochondrial genome is indeed constrained, which agrees with intuition about how essential mitochondrial function is and other sources which estimate the average instantaneous constraint to be very few amino acids [48]. The variation across sites in the average instantaneous constraint is low.

Perhaps the most important result is that the fitting of this kind of model, which allows the substitution process to change over time based on the fitnesses of the amino acids, was successful. It could correctly infer when multiple amino acids were acceptable from nearby substitutions. It produces a reasonable description of the substitution process as it changes over time.

V.5.3 Method Validation via Simulations

De Novo Simulations

The de novo simulations where sequences were generated directly from the model (Simulation 1) shows that the correct acceptabilities can be inferred from only the sequences. The model was fit to the 100 site and 100 taxon data set for 10,000 generations. It was completely burned in after 3,000 generations as shown in Supplemental Figure V.6. The likelihood converged, and the chain continued to explore the high likelihood space as shown in Supplemental Figure V.7. There were a total of 73,248 amino acids acceptable on branch segments across the whole tree. If one steps through every branch segment in the tree, and counts all acceptable amino acids at all sites at that branch segment, you would count 73,248 acceptable amino acids. Out of those 73,248 instances of amino acids being acceptable, the method described above inferred 50,692 (69.2%) correctly. This leaves 30.8% or 22,556 amino acids of the 73,248 that the method did not infer to be acceptable when they actually were. Only 205 (0.3%) amino acid acceptabilities were incorrectly added to the inferred set, indicating that false positives are very rare and amino acids are only inferred to be acceptable if there is sufficient support.

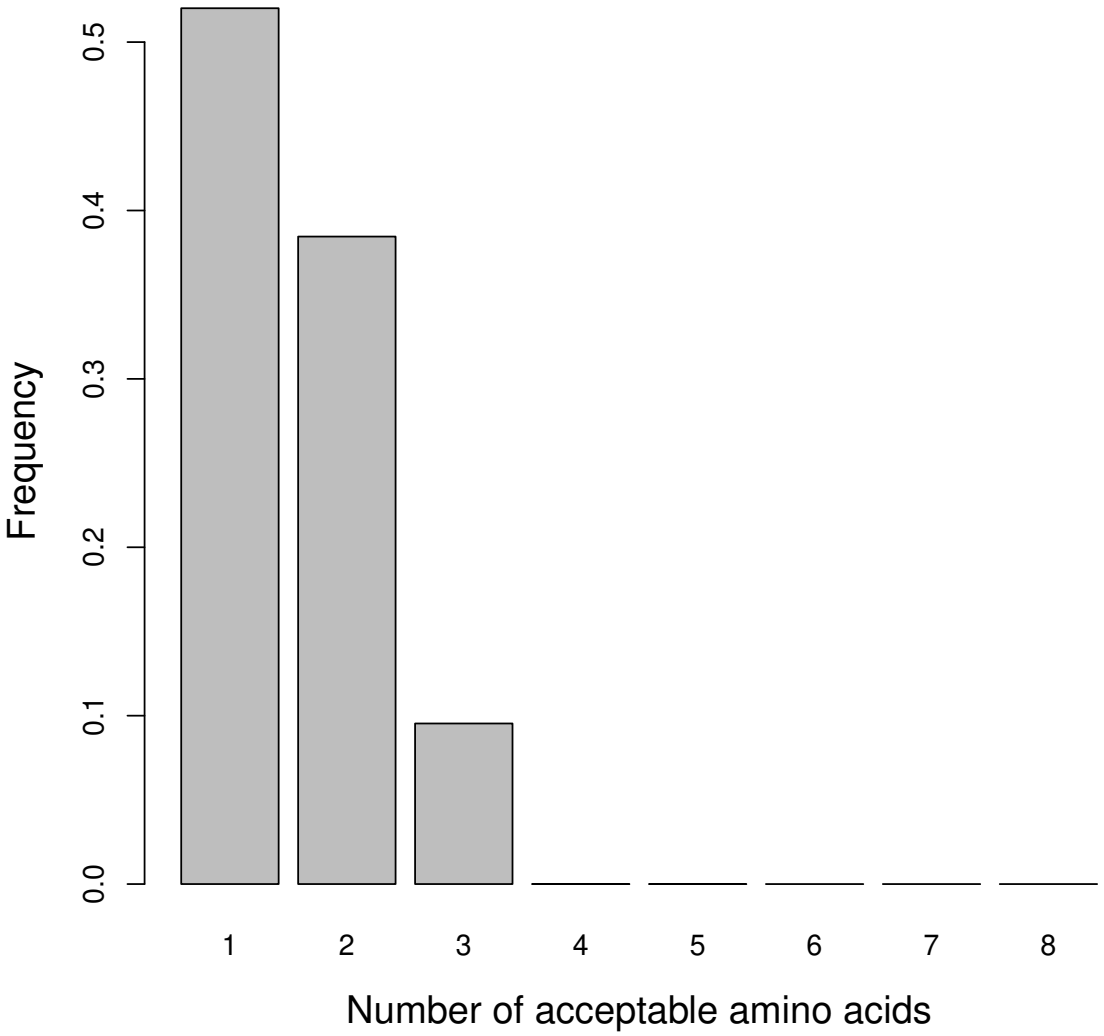


Figure V.12: The frequency distribution of the amount of constraint at each site for mitochondrial sites 101-200 integrated across the tree. The probability of having 4 or more amino acids acceptable is about $3 * 10^{-5}$. This agrees with the estimation that most sites have very few acceptable amino acids at any point in time.

Another de novo simulation was run to test the parameter fitting abilities of the Markov chain Monte Carlo. The simulated values for the switch probability P_ν and the substitution probability P_λ were 0.01 and 0.01. The chain had convergence and correctly estimated both the substitution rate and the switch rate as shown in Figures V.13 and V.14 after only 100 generations. This shows that while it may take many generations to infer the acceptabilities correctly, the model parameters can be estimated rapidly.

Simulation 2

For Simulation 2 where the inferred amino acid acceptabilities from the mitochondrial data are used to simulate sequences, most of the acceptabilities were inferred correctly. The posterior probability of the substitution probability (P_λ) was inferred to be around 0.12, which is very close to the number of substitutions divided by the number of branch segments 0.13.

The total number of a amino acid acceptabilities was calculated by stepping through every branch segment and counting how many amino acids were acceptable at each site, resulting in 127,764 actual acceptabilities. Out of those, the above method inferred 113,070 (88%) correctly. The actual acceptable amino acids which the method deemed incorrectly as unacceptable were 14,694 (12%). Amino acids which were added incorrectly to the acceptable sets accounted for 4,768 (22%) of the total. The results from Simulations 1 and 2 show that the inference procedure can capture many of the correct acceptabilities.

V.6 Conclusions and Discussion

In estimating how amino acid propensities change over time, I have shown how substitutions at adjacent sites correlate with large shifts in amino acid propensities. A novel method for determining the amino acid propensities of sites at specific times on the tree was proposed and tested via simulation. This propensity estimation method was then applied to the mitochondrial genome of 629 vertebrates. Substitutions were found to increase the average propensity shifts of adjacent sites at the same time as the substitution occurred. Many examples of large propensity shifts which correlated with

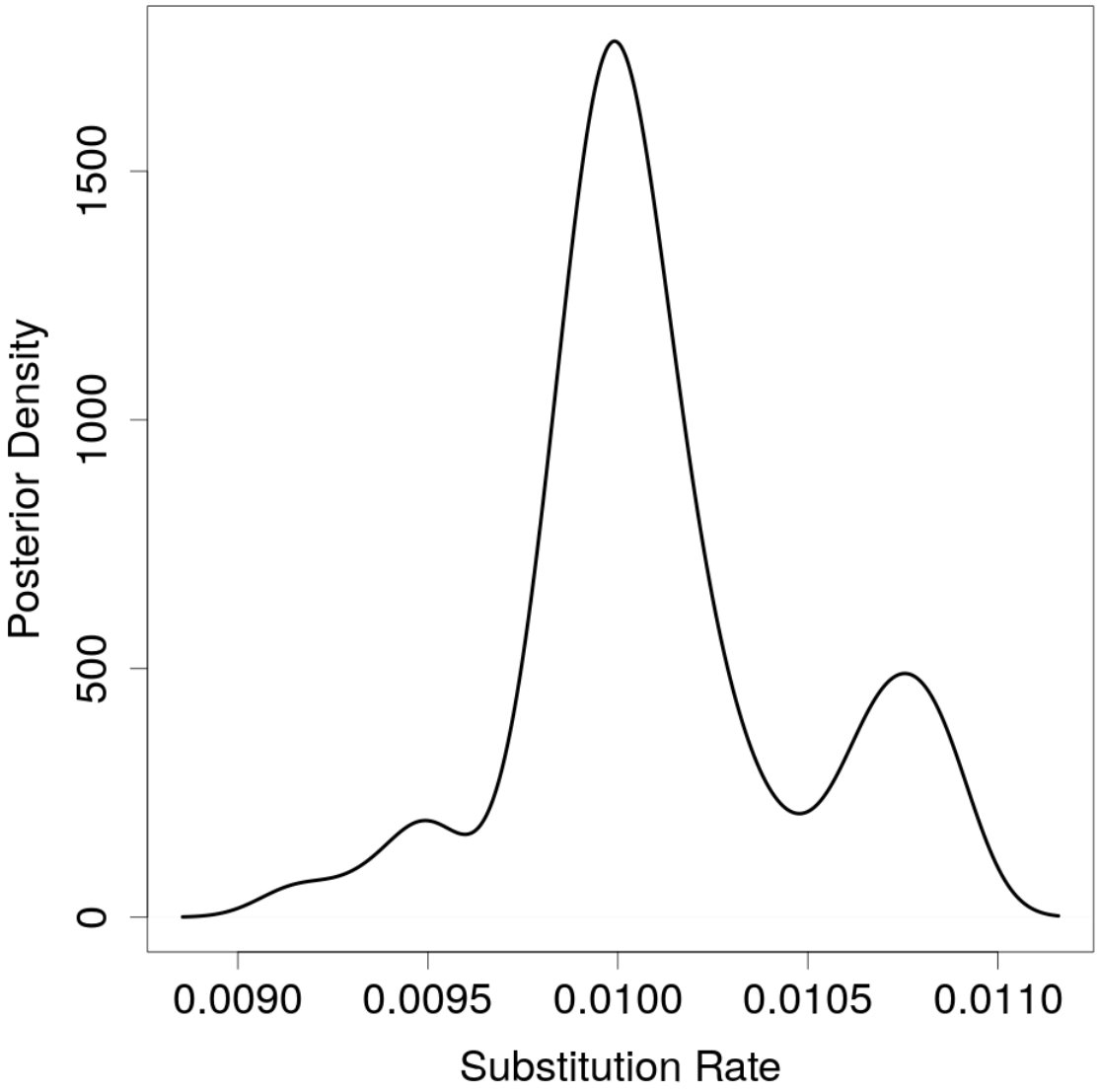


Figure V.13: The posterior distribution of substitution probability for a de novo simulation. The correct value is 0.01, which is clearly within the credible region. This shows that the inference method can estimate the substitution probability well given sequences which were simulated under the Acceptability Model.

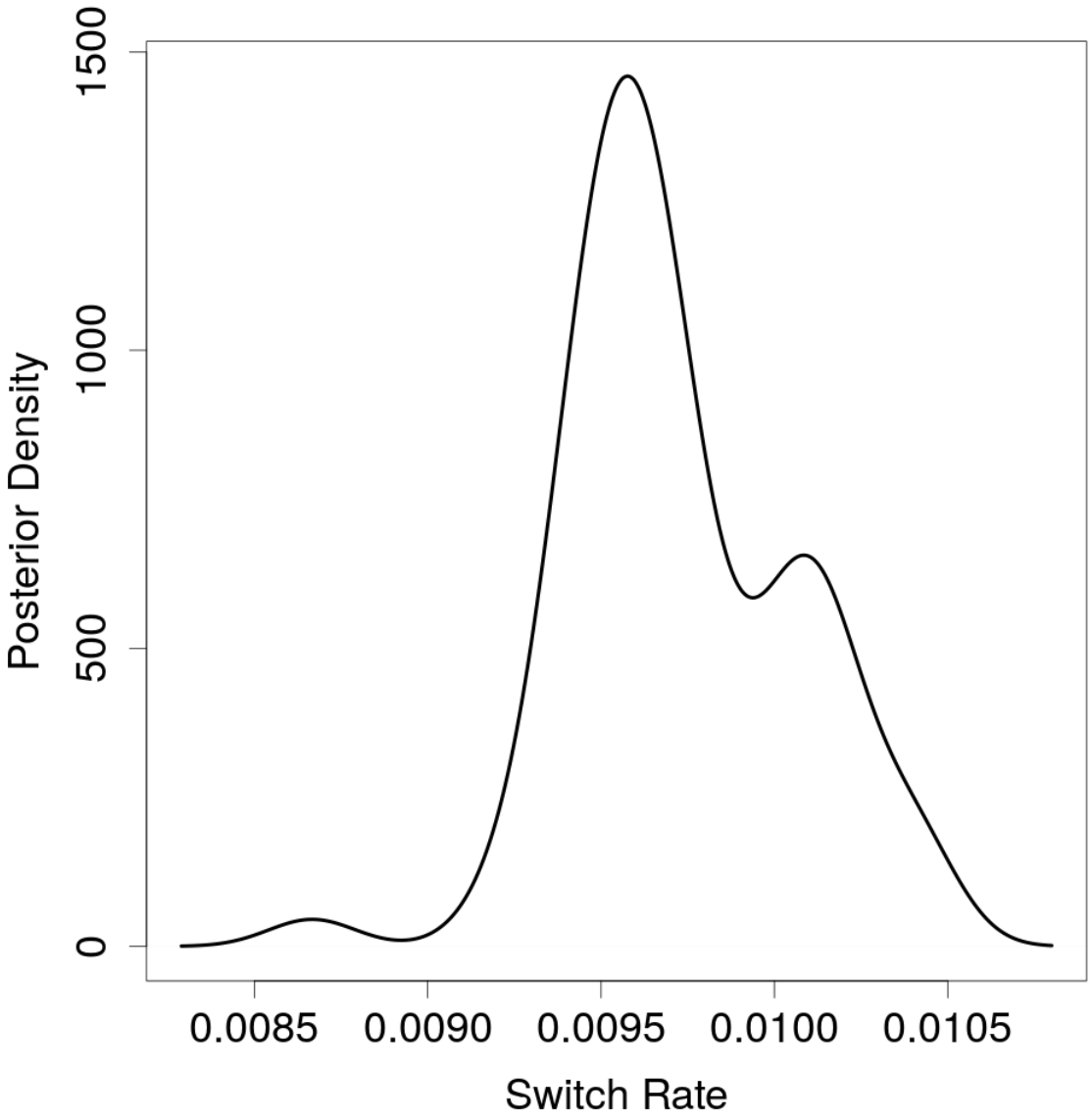


Figure V.14: The posterior distribution of the switch rate for a de novo simulation. The simulated value is 0.01. This shows that the inference method can estimate the amino acid fitness switching probability well given sequences which were simulated under the Acceptability Model.

adjacent substitutions were found and a few are shown in figures above. These results are novel because no model allows for a straightforward assessment of changes of amino acid propensities over time.

These results are corroborated with previous studies that found fluctuating substitution rates [29, 183, 163]. In Robinson et al. 2013, evidence is found that substitution rates at sites change depending on the amino acids in the sites surrounding the site. Substitution rates to amino acids that are high fitness given the surrounding amino acids are higher than rates to low fitness amino acids. Rodrigue found that a mutation selection model could detect shifts in the amino acid fitness landscape at a site. This new method could then be used to detect molecular adaptation better than the overall non-synonymous over the synonymous substitution rates (dN/dS) statistic alone. In a recent paper, Kazmi proposed another model allowing the amino acid propensities to change over time using site categories and breakpoints [163]. The results showed again that the amino propensities can shift and may differ dramatically across sites and time.

The Acceptability model has been shown to be effective at estimating the propensities at sites even as they change over time, as demonstrated by the simulation results. When comparing to the model proposed by Usmanova et al., the Acceptability model is simpler and easier to fit to data [57]. The proposed model can be applied to real, full sized trees and sequences, as opposed to only quadruplets. It also does not make assumptions about what amino acids are forbidden from a site altogether, allowing any amino acid to be acceptable at any site, and is general enough to allow any amino acid substitution probability matrix to be used.

There are some computational improvements yet to be made for inferring the acceptabilities as they change across sites and time. Most of the computational work involved in fitting the Acceptability model is in calculating the likelihood and sampling the hidden states. At present the likelihood is recalculated after the hidden states are Gibbs sampled. A possible improvement in calculating the likelihood would be simply update the likelihood

during the Gibbs sampling and only recalculating the full likelihood occasionally. One could test if the likelihood updating method calculates the likelihood exactly correctly by comparing the updated likelihood with the fully calculated. If the updating method introduces slight errors, then there may be a discrepancy in the calculated likelihoods. Since each site is completely independent, they can be sampled in parallel. Using parallel threads for each Gibbs sampling of the hidden states for each site should improve the sampling speed substantially, making the speedup roughly equal to the number of sites in the multiple sequence alignment. If the two improvements above are implemented, then each site would update the likelihood independently, possibly causing race conditions. This problem could be solved by either decomposing the overall likelihood into site-specific likelihoods or using a data structure that allows for concurrent access to the likelihood value.

The Acceptability model is designed to be simple and yet allow the evolution process to vary over sites and time. Possible extensions to this model could relax some assumptions inherent to this model. First there is a single parameter which governs both the switch on and switch off rates ν, which could be split into two different rates: switch on rate ν_{on} and switch off rate ν_{off}. This would eliminate the assumption that the on and off rates are equal. It could also reduce or remove the need for the prior on the acceptable set size ρ, since the ratio of $\frac{\nu_{on}}{\nu_{off}}$ would determine the steady state equilibrium probability that any amino acid is on at any point in time.

It is also possible that the resident amino acid might increase the switch on rate for other amino acids with similar physicochemical properties. The switch on rates would then be described by a matrix with each row corresponding to the resident amino acid and each cell is the switch on rate for each amino acid in the column. A model like this might be sufficient to explain the substitution rates seen in current matrix substitution rate models, however the explanation for these rates would be different. The resident amino acid would not randomly substitute to different amino acids at different rates, rather

the resident amino acid would change the probability of similar amino acids becoming acceptable. An explanation like this brings us closer to a mechanistic model instead of a descriptive model.

One could also use more fitness levels than the binary fitness used here, if the data support having more levels. One complication is that the complexity of the Gibbs sampling increases linearly with the number of levels added. Adding more levels may add to the computational cost without learning more about the evolution process. The substitution process among high fitness amino acids could also be expanded above the simple one parameter model used here. If the nucleotide information is available, one could incorporate a model which allows for transition/transversion rate differences, such as Kimura's model [21]. One could use an entire matrix to describe the substitution process such as WAG or mtMam [142, 141].

Few models of evolution take into account that there may be a few different amino acids at a site in different individuals within a species, a phenomenon known as genetic polymorphism. The sequencing the genomes of hundreds of humans has shown the prevalence of polymorphism in the human population. Although genetic polymorphism has been demonstrated in many species, the genetic information for other species in most sequence databases does not include polymorphic information, perhaps because the sequences were derived from a single individual or because information about polymorphism and minor allele frequencies was removed in order to produce a single residue per site. Allele frequencies within extant species could be used to estimate allele fitness within the species, and thus expand the information at each site in an extant species sequence from a single residue to a fitness profile across all residues. The fitness profiles could then be used by models which allow the fitness profiles of amino acids at each site to change over time, such as the Acceptability model.

Genetic linkage among variants also can misinform analyses of shifting amino acid propensities. It is possible that a slightly deleterious or neutral mutation could be

genetically linked to high fitness variants at nearby sites and then the mutation could be carried to higher frequencies simply from this genetic association. In the light of this possibility, the assertion that every inferred ancestral amino acid is a high fitness amino acid may be generally correct, but perhaps incorrect in specific instances.

Now that we can estimate the propensities at all the sites in an alignment and at the ancestral species, we can ask what other factors impact propensity shifts. There are many questions about how the structure of a protein influences its evolution that we might be able to answer. Do substitutions at one site cause large shifts in the propensities of sites which interact in the 3D structure with that site? We can probe whether there are sites which influence each other's propensities strongly but are in fact far away in the 3D structure of a protein. If we observe a substitution from a neutral amino acid to a charged amino acid, do the propensities for the oppositely charged amino acids increase at sites nearby in 3D space? Once we can calculate how the propensities have changed over time, we can start answering questions like these.

V.7 Declarations

V.7.1 Ethics approval and consent to participate

Not applicable

V.7.2 Consent for publication

Not applicable

V.7.3 Availability of data and material

The datasets generated and/or analyzed during the current study are available in the Propensity Changes repository, https://github.com/spollard/PropensityChanges.

V.7.4 Competing interests

The authors declare that they have no competing interests.

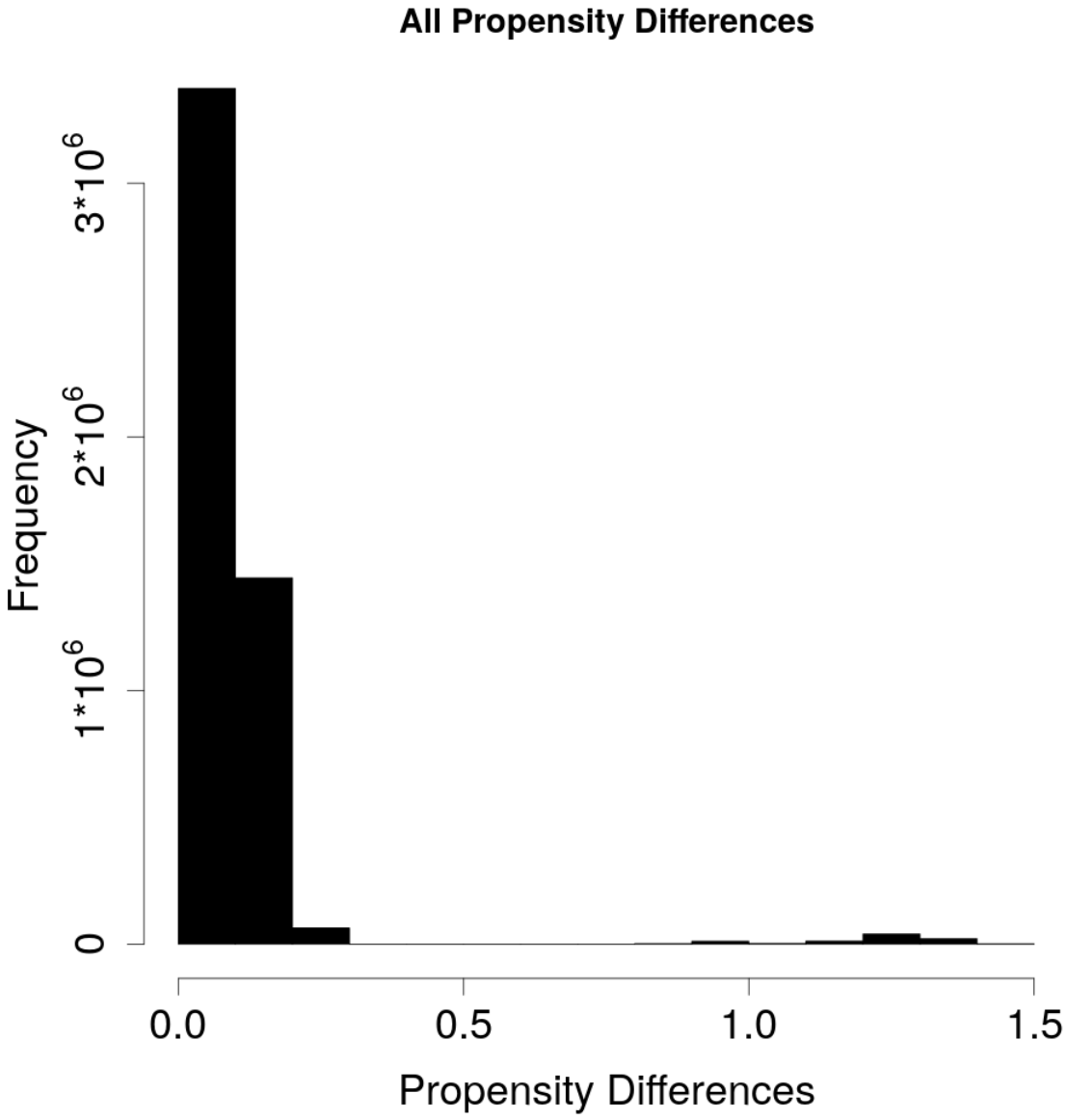
V.7.5 Funding

We acknowledge the support of the National Institutes of Health (NIH; GM083127 and GM097251) to David D. Pollock.

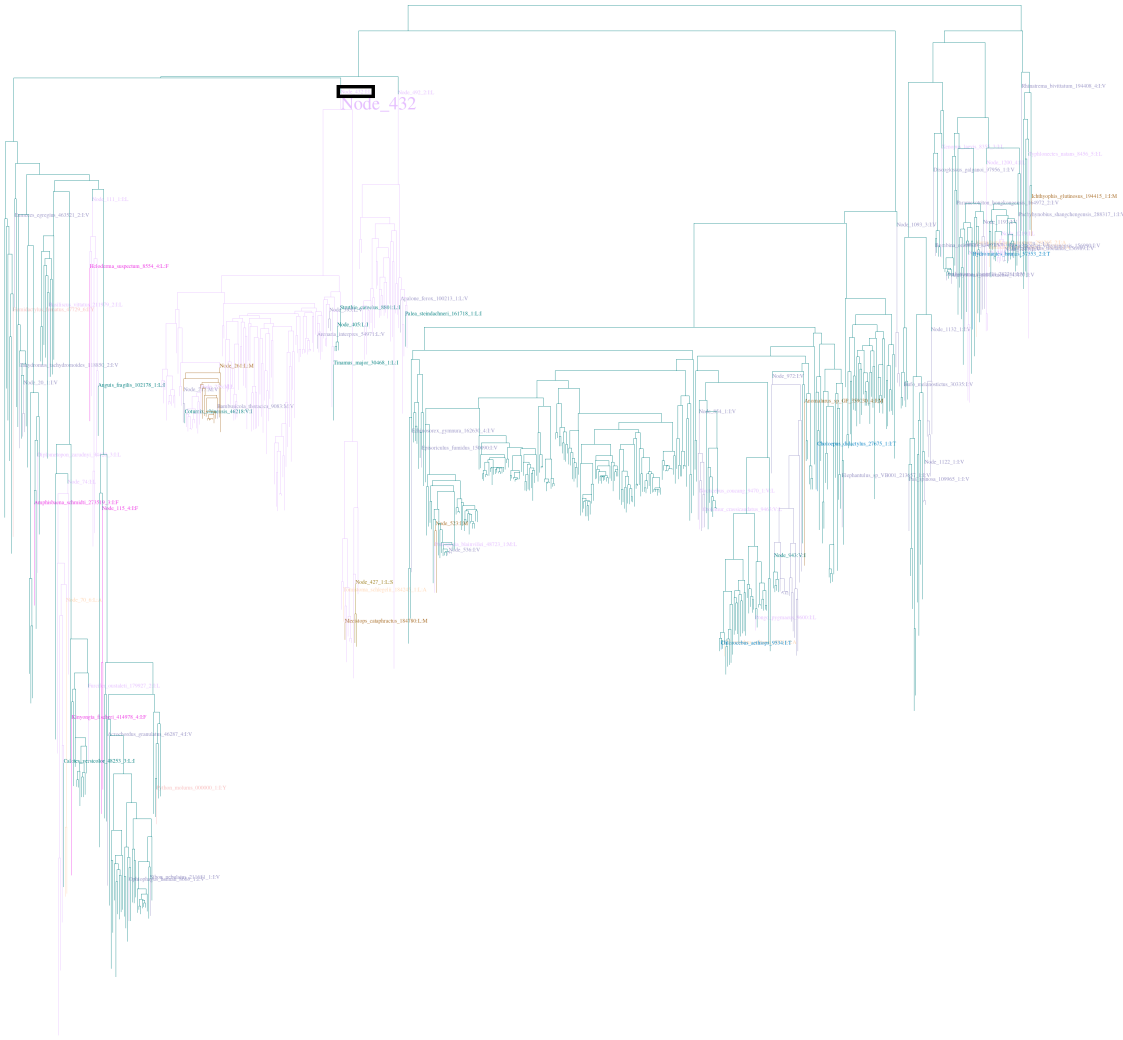
V.7.6 Authors' contributions

Stephen Pollard developed the software, performed the research, and wrote the manuscript. All authors read and approved the final manuscript.

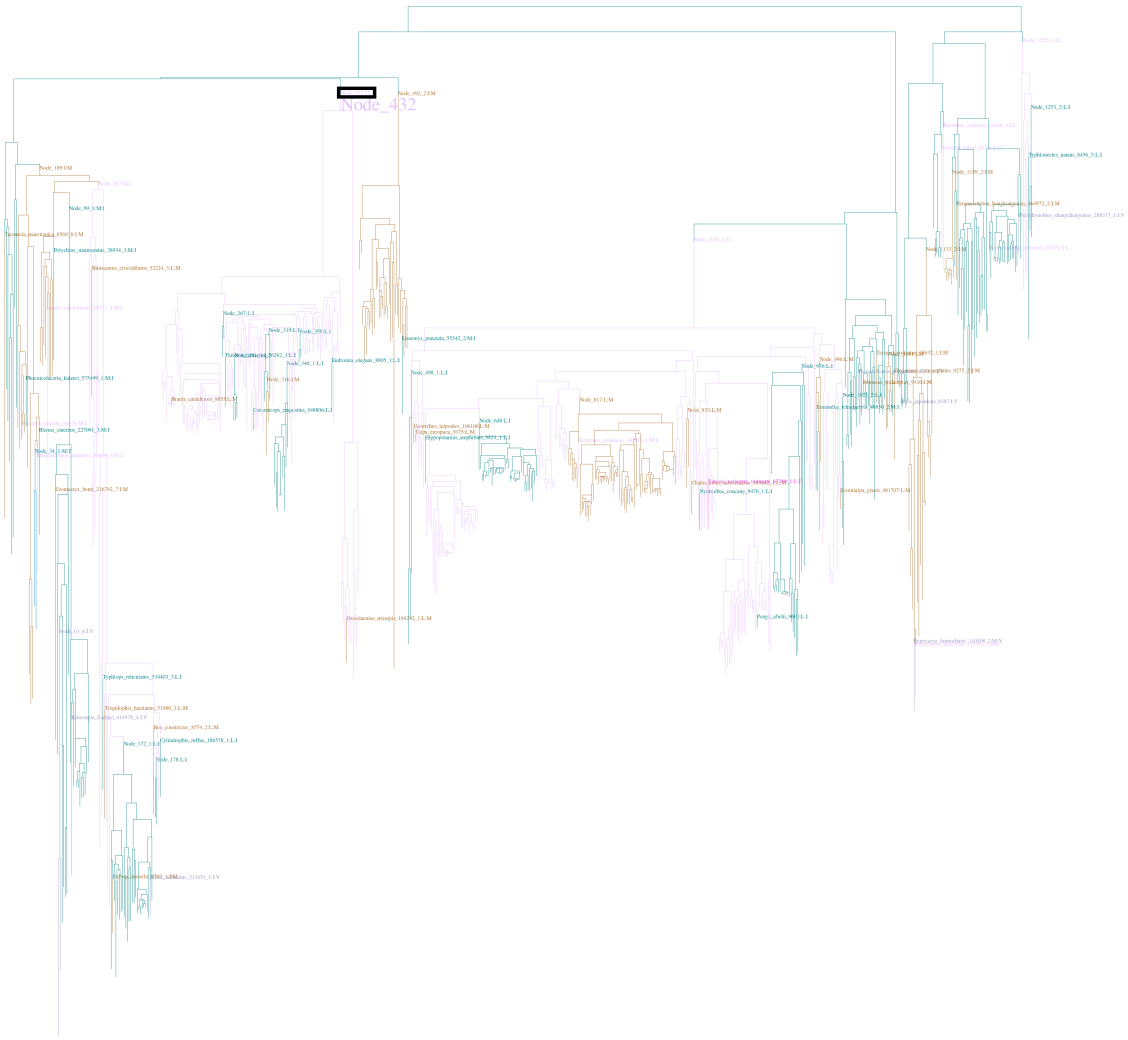
V.8 Supplemental Figures



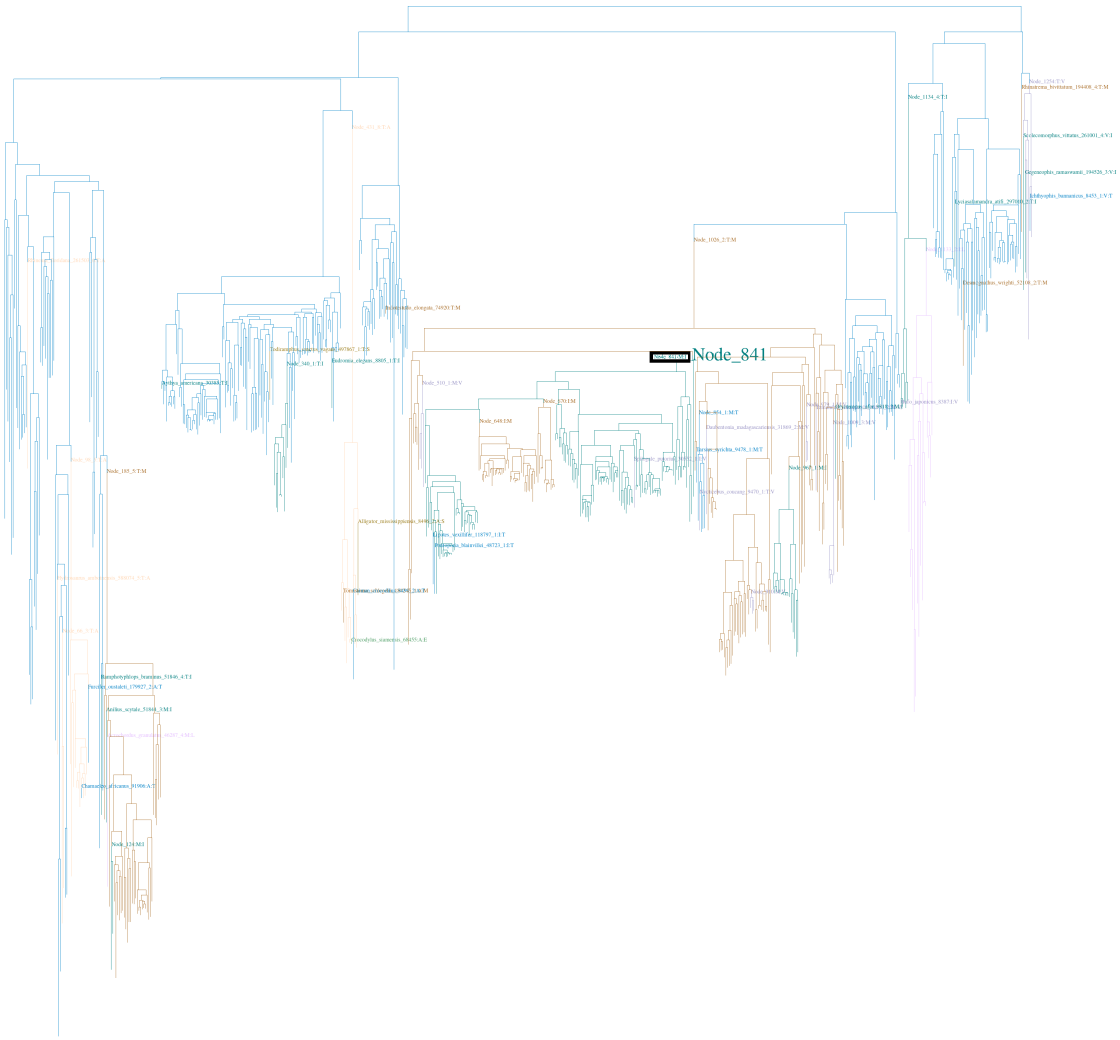
Supplementary Figure V.1: A histogram showing the distribution of amino acid propensity shifts for all sites and all branches. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution results in a mean shift size of 0.085.



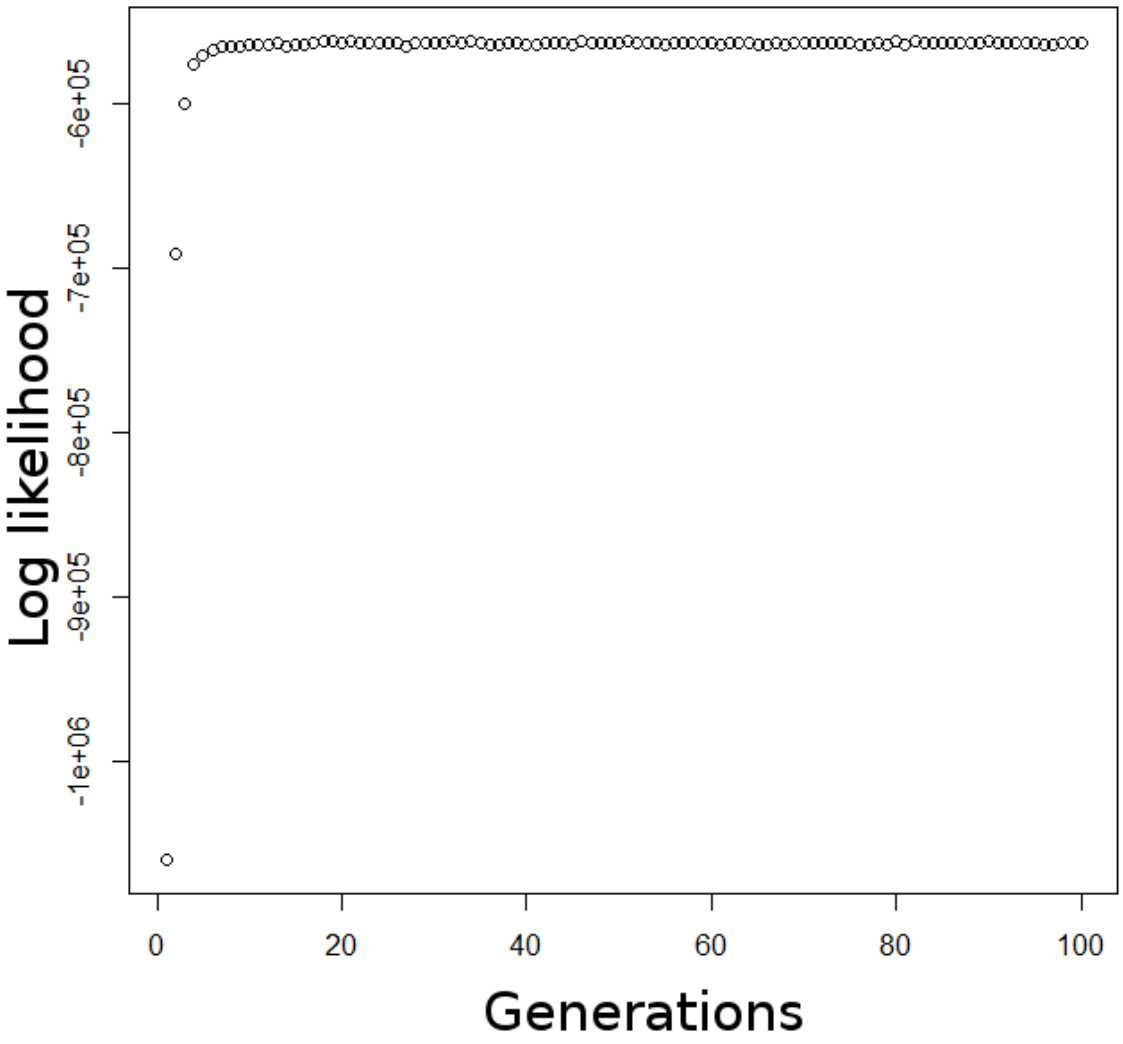
Supplementary Figure V.2: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1031 of the amino acid multiple sequence alignment. Site 1031 is adjacent to site 1030 in the amino acid alignment. The branches where isoleucine is resident are colored dark blue, and the branches where leucine is resident are colored pink. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. The label Node_432:I:L indicates that a substitution from isoleucine to leucine occurred at the branch directly ancestral to Node_432. This substitution from isoleucine to leucine at ancestral node 432 in site 1031 (shown in a black box) correlates with the large shift in the propensities of the amino acids at site 1030 at ancestral node 432. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.



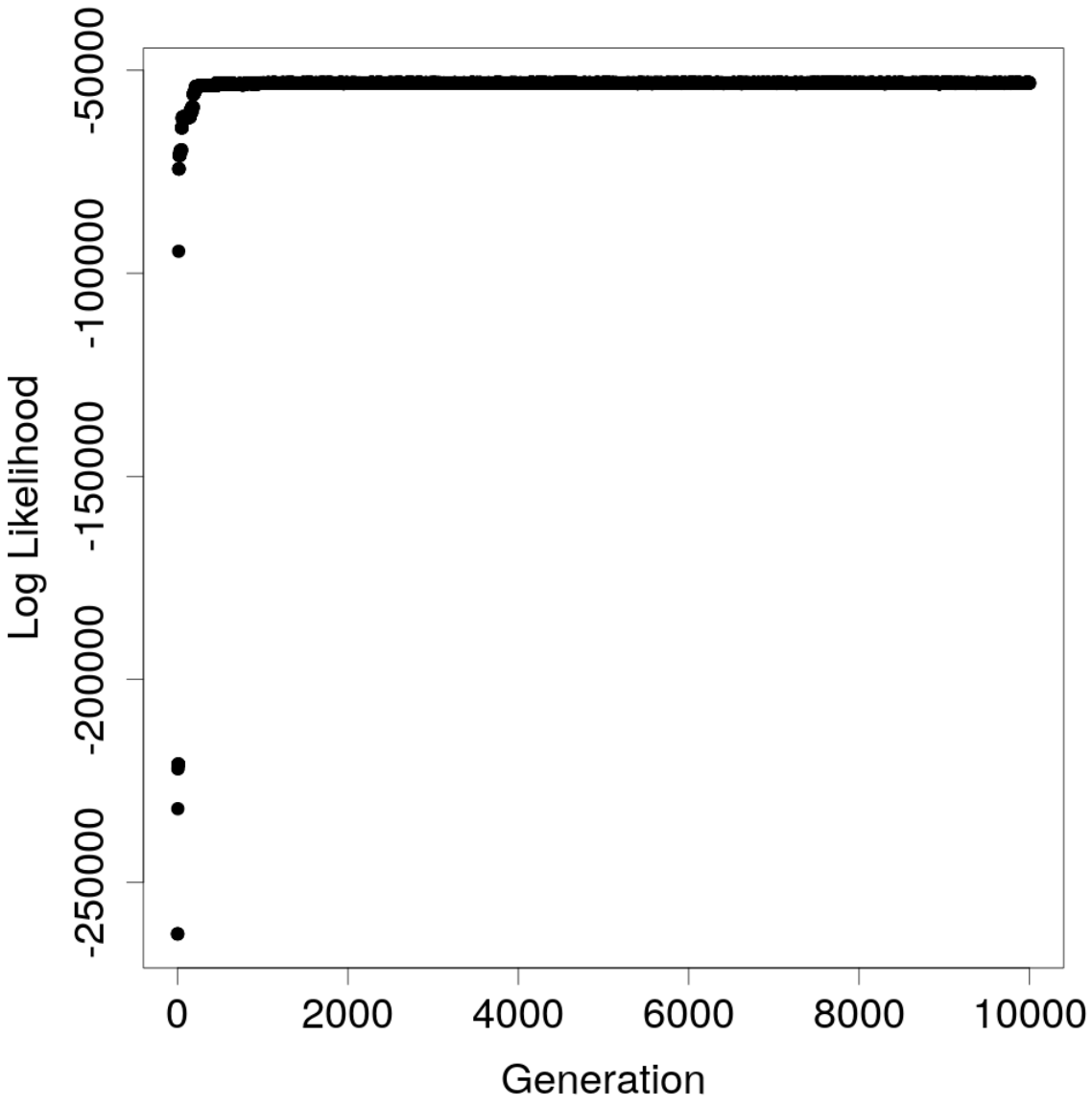
Supplementary Figure V.3: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1250 of the amino acid multiple sequence alignment. The branches where methionine is resident are colored orange, and the branches where threonine is resident are colored light blue. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. For example the label Node_410_4:M:T indicates that a substitution from methionine to threonine occurred at the branch directly ancestral to Node_410_4. The substitution from isoleucine to leucine at ancestral node 432 in site 1250 (shown in a black box) correlates with the large shift in the propensities of the amino acids at site 1251 at ancestral node 432. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.



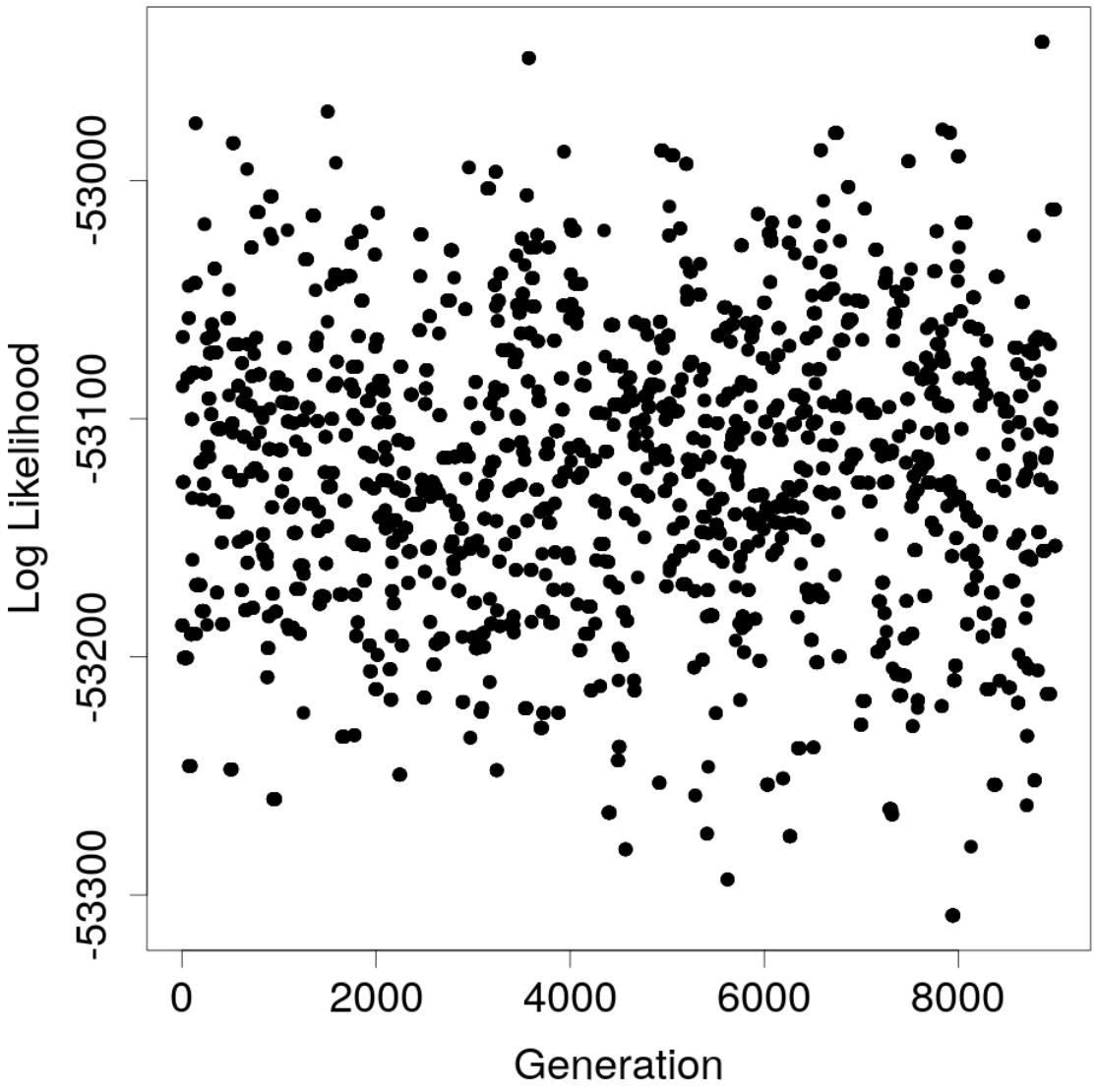
Supplementary Figure V.4: The mitochondrial phylogenetic tree colored by the resident amino acid at site 1375 of the amino acid multiple sequence alignment. The branches where methionine is resident are colored orange, the branches where threonine is resident are colored light blue, and the branches where isoleucine is resident are colored dark blue. Substitutions are denoted by a change of color and a label of which node the substitution occurred on and the amino acids substituted from and to. For example the label Node_410_4:M:T indicates that a substitution from methionine to threonine occurred at the branch directly ancestral to Node_410_4. The amino acid propensities shift along the branch directly above ancestral node 481 from heavily favoring valine (100% valine) at the ancestral node to a 48%-52% split between valine and isoleucine at node 481. This shift correlates with a substitution at site 1375 and node 481 from methionine to isoleucine (boxed). The shift resulted in substitutions at site 1376 at the direct descendants of node 432 from valine to isoleucine, however the reversions to valine further down in the clade support that valine is probably acceptable throughout the clade. High resolution images can be found here: `https://www.dropbox.com/s/2annigg98nu3mme/high_resolution_figures.zip?dl=0`.



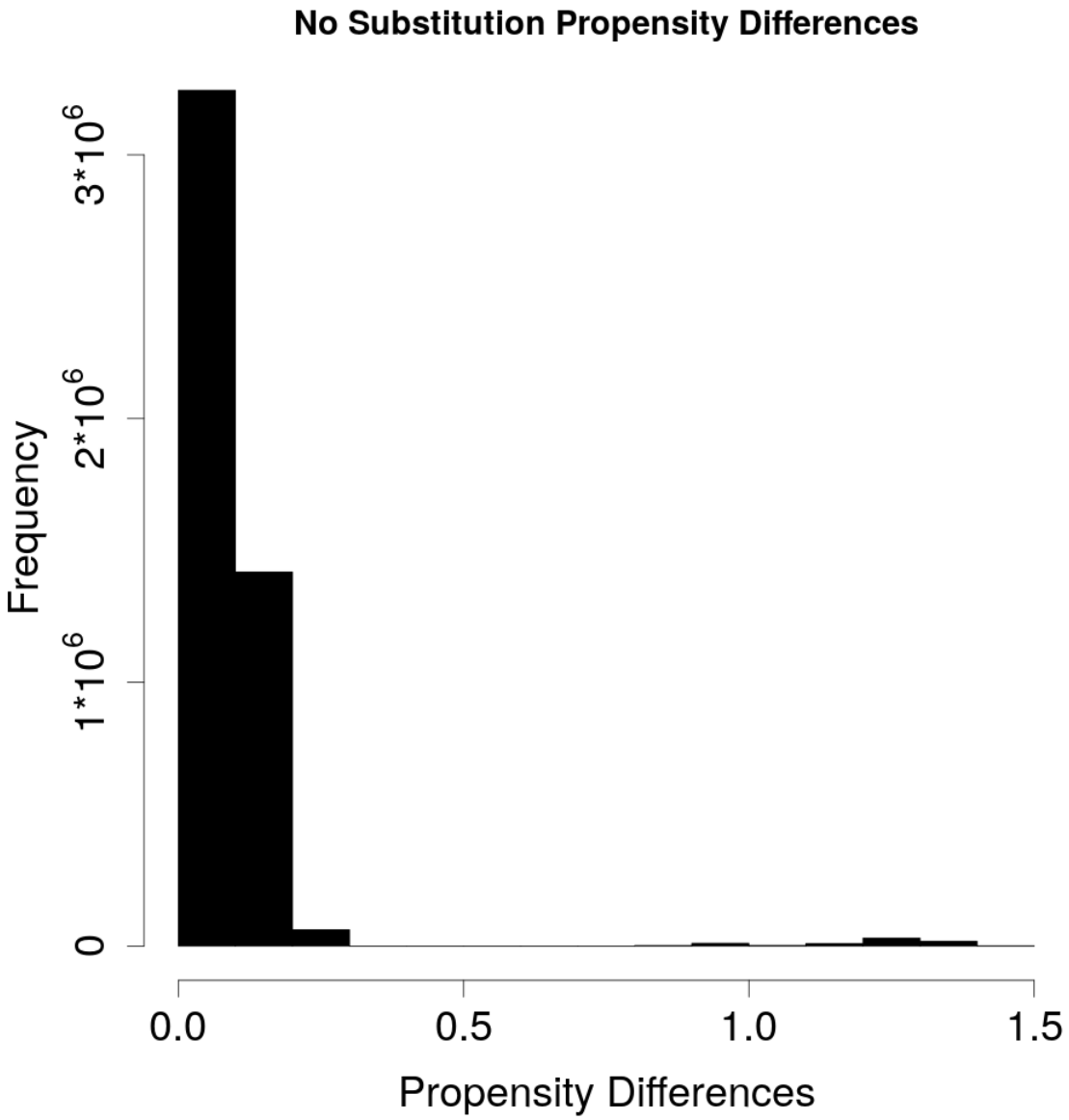
Supplementary Figure V.5: The likelihood trace from the group of sites from 101 to 200 from the mitochondrial data set. After about 10 generations the chain is completely burned in.



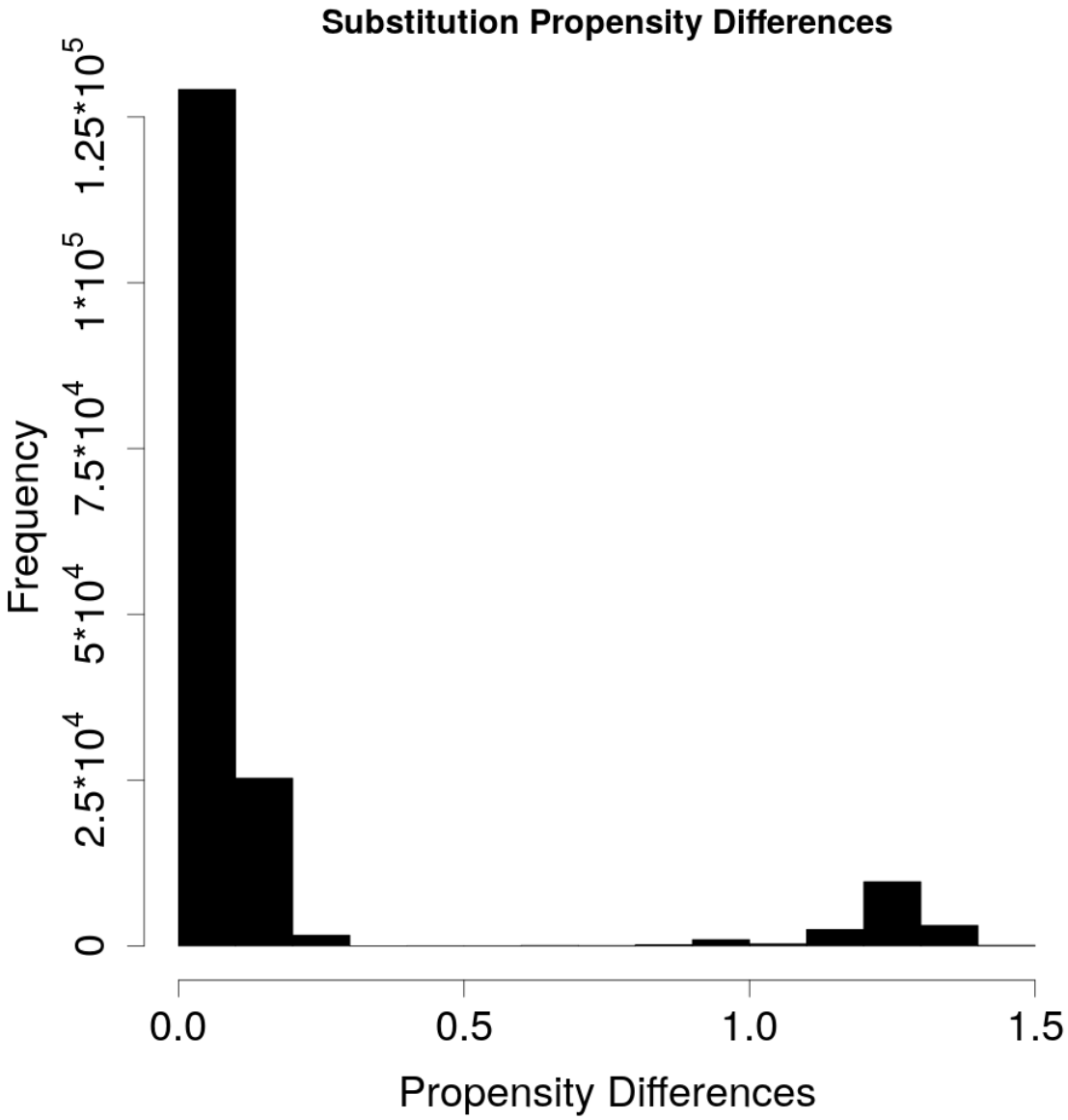
Supplementary Figure V.6: The likelihood trace for the de novo simulation, Simulation 1, showing that the fitting method is finding the high likelihood propensities. The chain is completely burned in after 3,000 generations.



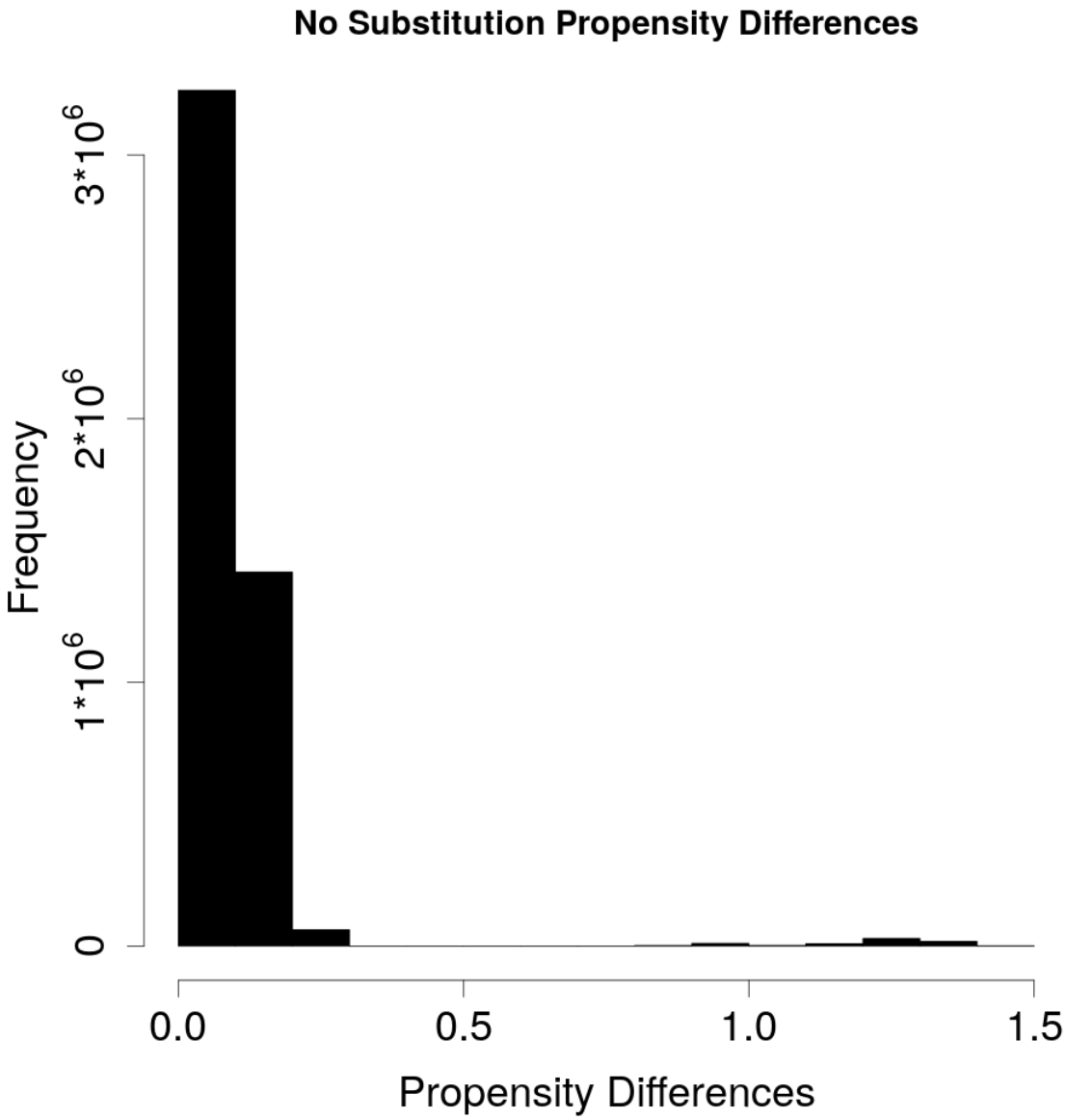
Supplementary Figure V.7: The likelihood trace for the first simulation after the burnin period, indicating that the chain is mixing well after burning in.



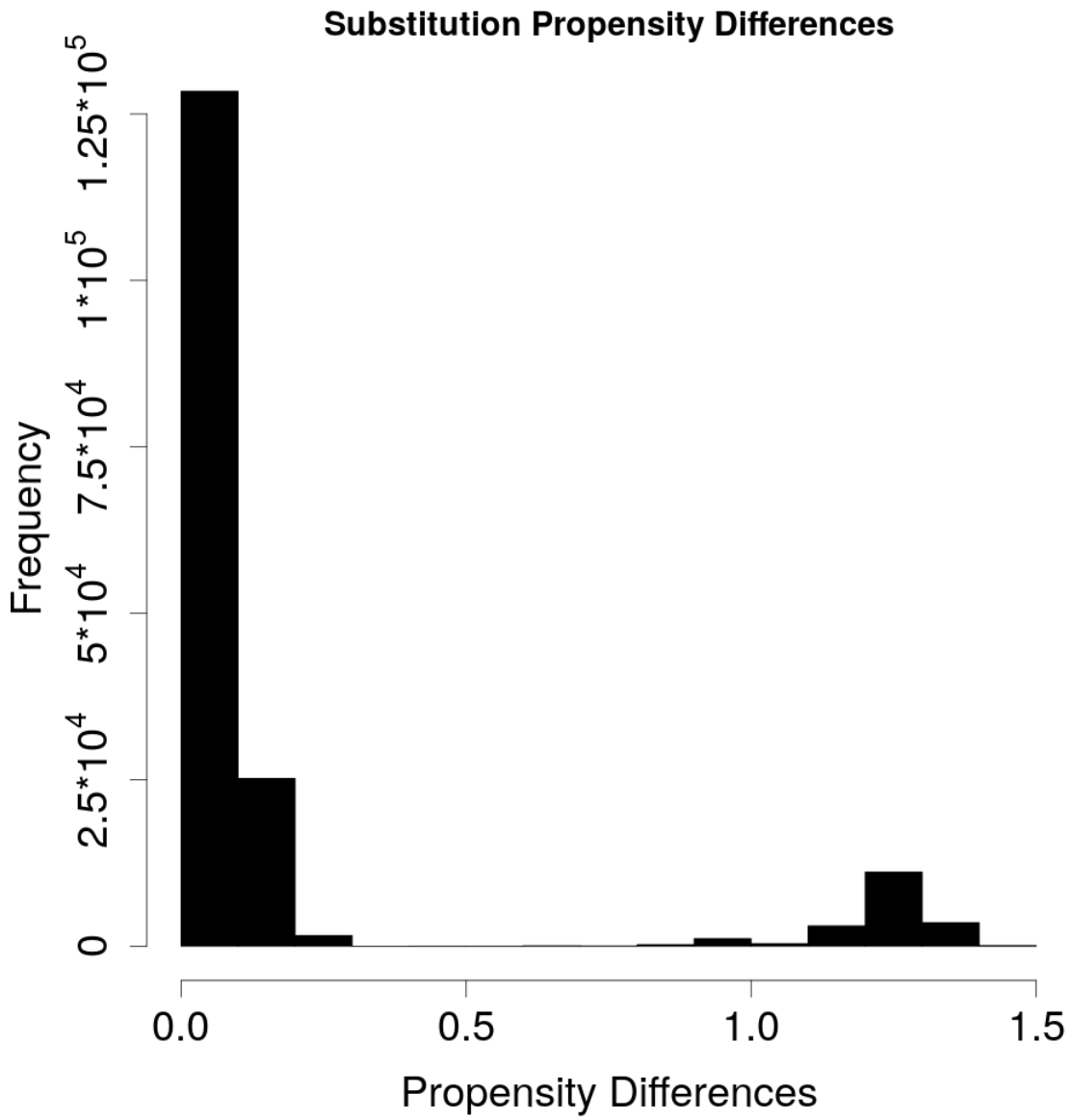
Supplementary Figure V.8: A histogram showing the distribution of amino acid propensity shifts given that no substitutions were detected at sites two positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41.



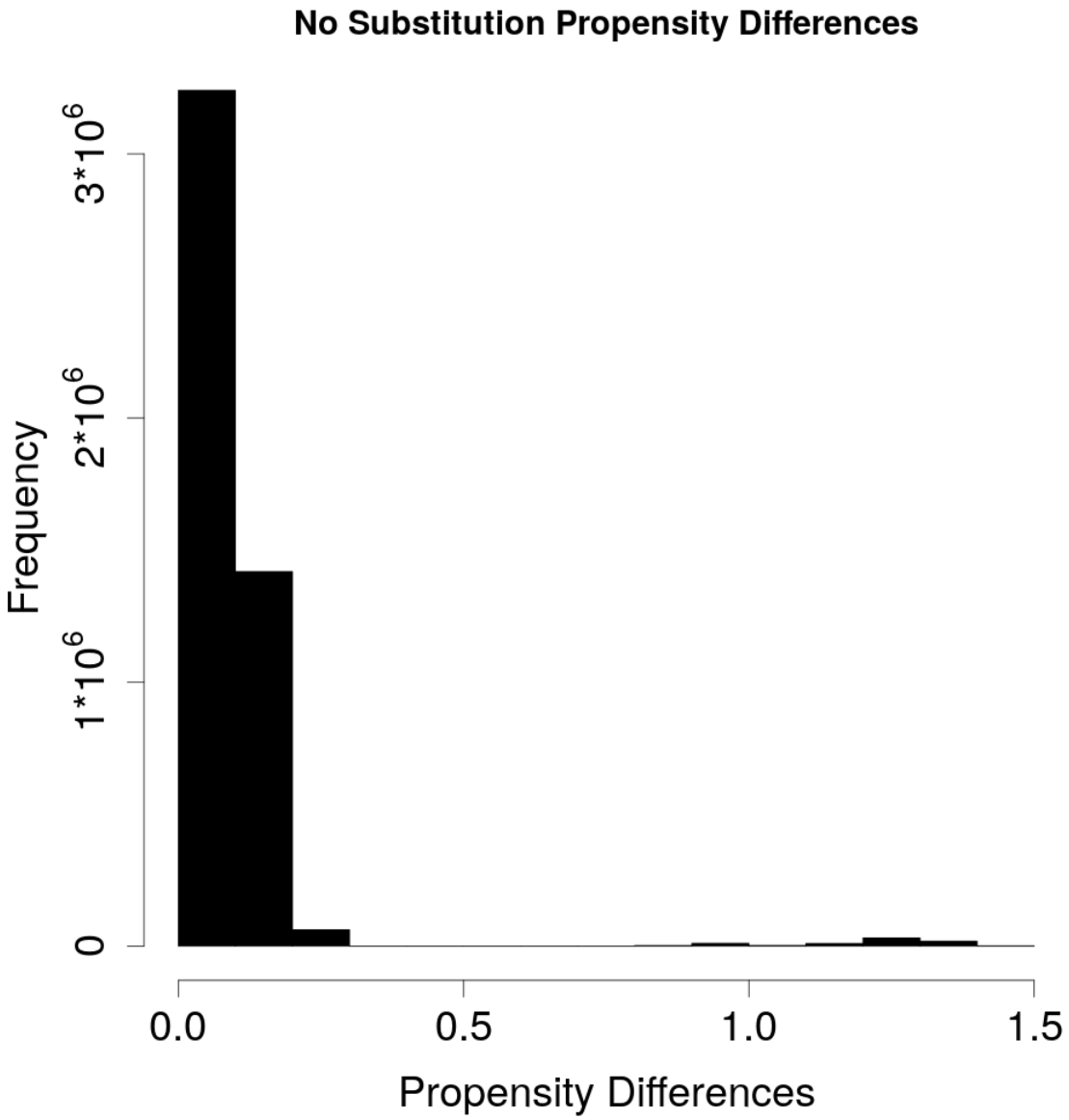
Supplementary Figure V.9: A histogram showing the distribution of amino acid propensity shifts given that at least one substitution was detected at sites two positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8, as in Figure V.2. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution is compared against the distribution given zero substitutions at sites two positions away using the Kolmogorov–Smirnov test and the probability that the two distributions resulted from the same underlying distribution is less than $2.2\text{x}10^{-16}$, indicating that amino acid substitutions substantially increase the propensity shifts at adjacent sites in the multiple sequence alignment.



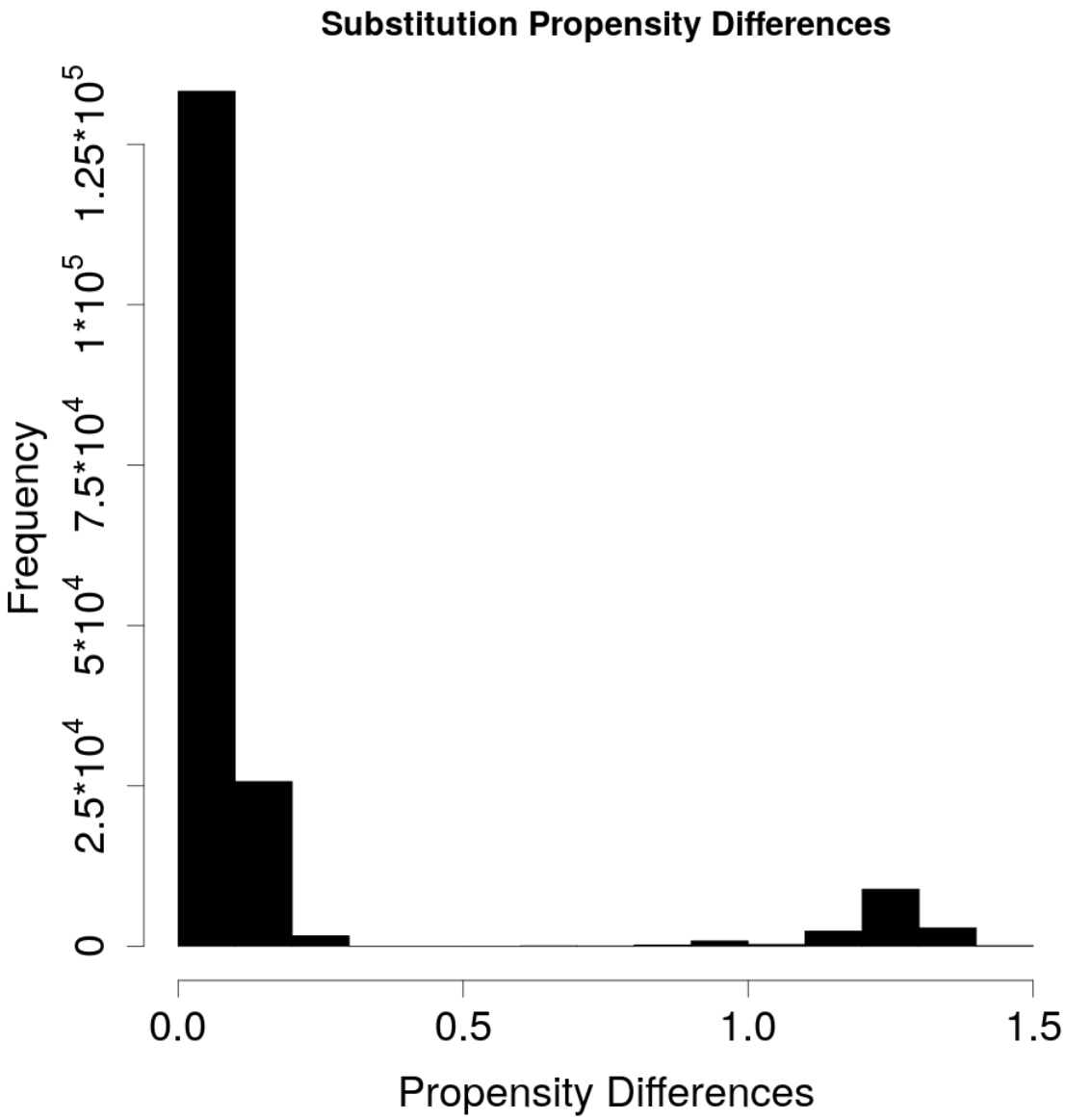
Supplementary Figure V.10: A histogram showing the distribution of amino acid propensity shifts given that no substitutions were detected at sites three positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41.



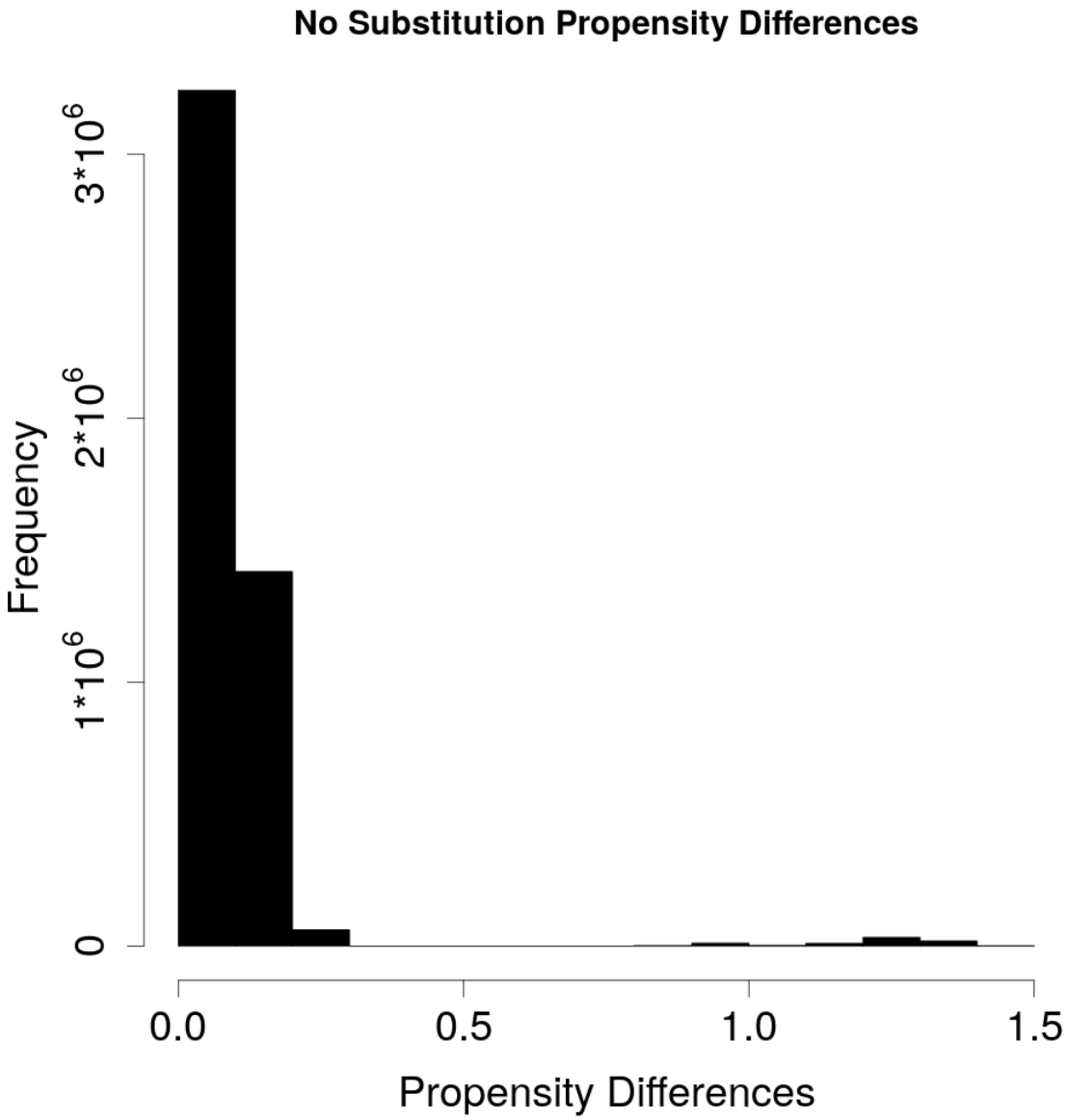
Supplementary Figure V.11: A histogram showing the distribution of amino acid propensity shifts given that at least one substitution was detected at sites three positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8, as in Figure V.2. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution is compared against the distribution given zero substitutions at three positions away using the Kolmogorov–Smirnov test and the probability that the two distributions resulted from the same underlying distribution is less than $2.2\text{x}10^{-16}$, indicating that amino acid substitutions substantially increase the propensity shifts at adjacent sites in the multiple sequence alignment.



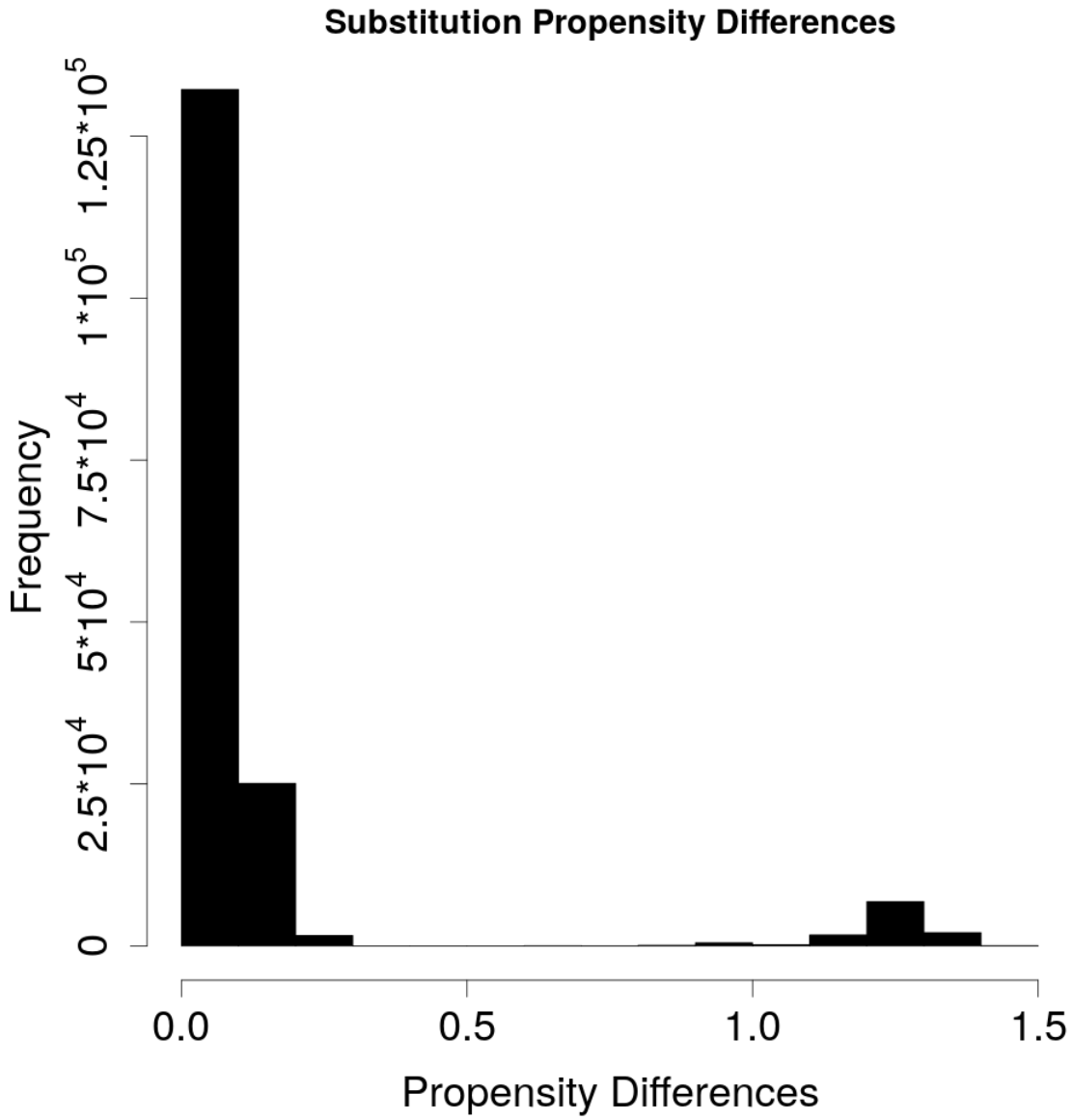
Supplementary Figure V.12: A histogram showing the distribution of amino acid propensity shifts given that no substitutions were detected at sites ten positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41.



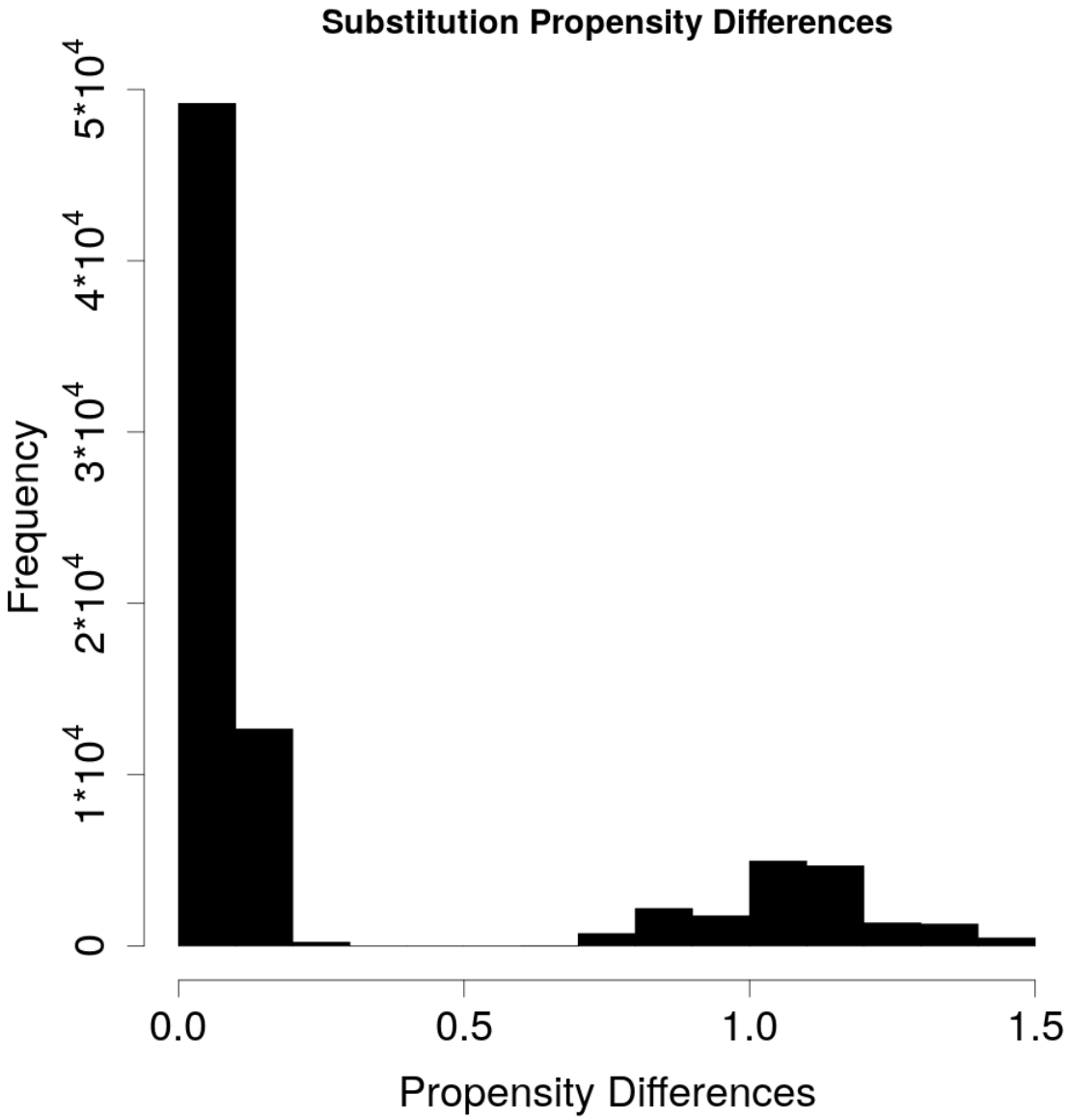
Supplementary Figure V.13: A histogram showing the distribution of amino acid propensity shifts given that at least one substitution was detected at sites ten positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8, as in Figure V.2. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution is compared against the distribution given zero substitutions at sites 10 positions away using the Kolmogorov–Smirnov test and the probability that the two distributions resulted from the same underlying distribution is less than $2.2\text{x}10^{-16}$, indicating that amino acid substitutions substantially increase the propensity shifts at adjacent sites in the multiple sequence alignment.



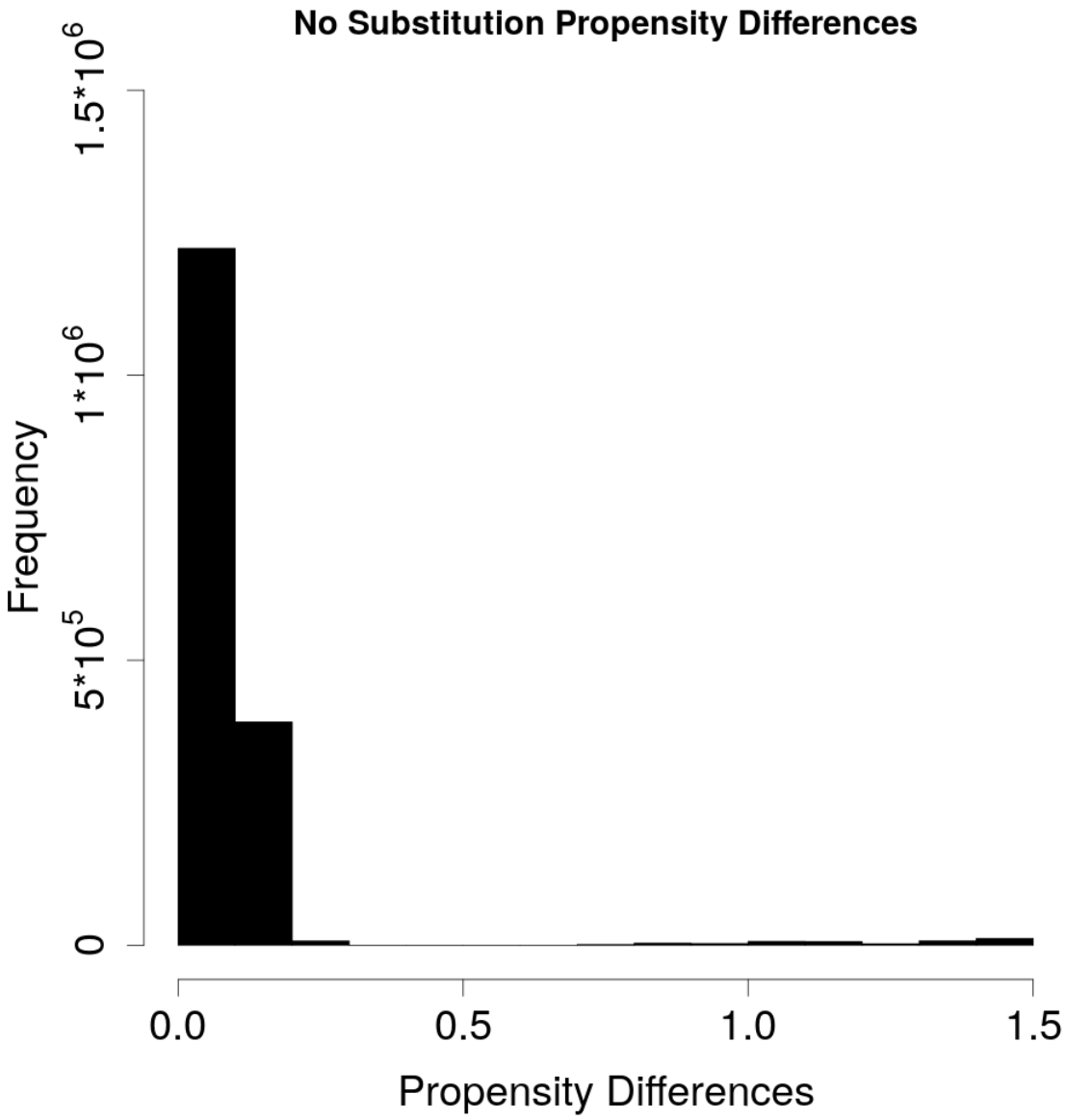
Supplementary Figure V.14: A histogram showing the distribution of amino acid propensity shifts given that no substitutions were detected at sites one hundred positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41.



Supplementary Figure V.15: A histogram showing the distribution of amino acid propensity shifts given that at least one substitution was detected at sites one hundred positions away in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8, as in Figure V.2. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution is compared against the distribution given zero substitutions at positions 100 sites away using the Kolmogorov–Smirnov test and the probability that the two distributions resulted from the same underlying distribution is less than $2.2\text{x}10^{-16}$, indicating that amino acid substitutions substantially increase the propensity shifts at adjacent sites in the multiple sequence alignment.



Supplementary Figure V.16: A histogram showing the distribution of amino acid propensity shifts from the nuclear encoded glycolysis proteins given that at least one substitution was detected at sites immediately adjacent in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8, as in Figure V.2. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41. This distribution is compared against the distribution given zero substitutions at adjacent sites using the Kolmogorov–Smirnov test and the probability that the two distributions resulted from the same underlying distribution is less than $2.2\text{x}10^{-16}$, indicating that amino acid substitutions substantially increase the propensity shifts at adjacent sites in the multiple sequence alignment.



Supplementary Figure V.17: A histogram showing the distribution of amino acid propensity shifts from the nuclear encoded glycolysis proteins given that no substitutions were detected at sites immediately adjacent in the alignment at the branch segment where the propensity differences were calculated. Propensity shifts are calculated using Euclidean distance, as described by equation V.8. Many sites show zero shifts, corresponding to a zero Euclidean distance in the propensities from ancestor to descendant. The size of a propensity shift from 100% of one amino acid to 100% of another amino acid is $\sqrt{2}$ or 1.41.

CHAPTER VI

MARKOV KATANA*

VI.1 Abstract

VI.1.1 Background

Phylogenetic inference requires a means to search phylogenetic tree space. This is usually achieved using progressive algorithms that propose and test small alterations in the current tree topology and branch lengths. Current programs search tree topology space using branch-swapping algorithms, but proposals do not discriminate well between swaps likely to succeed or fail. When applied to datasets with many taxa, the huge number of possible topologies slows these programs dramatically. To overcome this, we developed a statistical approach for proposal generation in Bayesian analysis and evaluated its applicability for the problem of searching phylogenetic tree space. The general idea of the approach, which we call "Markov Katana", is to make proposals based on a heuristic algorithm using bootstrapped subsets of the data. Such proposals induce an unintended sampling distribution that must be determined and removed to generate posterior estimates, but the cost of this extra step can in principle be small compared to the added value of more efficient parameter exploration in Markov chain Monte Carlo analyses.

VI.1.2 Results

Our prototype application uses the simple neighbor-joining distance heuristic on data subsets to propose new reasonably likely phylogenetic trees (including topologies and branch lengths). The evolutionary model used to generate distances in our prototype was far simpler than the more complex model used to evaluate the likelihood of phylogenies based on the full dataset. We demonstrate that this method can be used to efficiently estimate a Bayesian posterior.

*Authors include Stephen T. Pollard, Kenji Fukushima, Zhengyuan O. Wang, Todd A. Castoe, and David D. Pollock.

VI.1.3 Conclusions

This prototype implementation indicates that the Markov Katana approach could be easily incorporated into existing phylogenetic search programs and may prove a useful alternative in conjunction with existing methods. The general features of this statistical approach may also prove useful in disciplines other than phylogenetics.

VI.2 Keywords

Bootstrap, Bayesian, phylogenetics, tree search, heuristic

VI.3 Background

Phylogenetic inference has long played a pivotal role in molecular evolution and evolutionary genomics (e.g. [143, 184, 185]). It provides unique information about gene and protein interactions [186, 187, 188, 189] and is critical for detecting adaptive bursts and functional divergence (e.g. [75, 106]). Despite its importance, phylogenetic inference is difficult partly because searching tree space is an NP-hard problem [190, 191]. Distance-based methods such as neighbor-joining (NJ; [192]) are fast and often provide good approximate results but are considered less reliable than the computationally expensive likelihood-based methods (maximum likelihood, ML, and Bayesian or posterior probability, PP) [193, 194, 110]. While distance methods generate a single tree using heuristic approaches, likelihood methods must search tree space, generally by running an optimization scheme or Markov chain Monte Carlo (MCMC). Tree space is often searched using various forms of branch swapping [195, 196, 151, 197, 198]. A cautious approach to interpreting results from traditional branch-swapping algorithms is warranted, particularly for trees with sequences from many taxa [199].

The principle confounding effect in phylogenetic inference is that multiple substitutions may occur at the same site. Distance-based methods are inferior to likelihood-based methods in accurately inferring multiple substitutions [200, 201, 202]. Distance-based methods are also far more strongly biased by long-branch attraction and cannot fully

incorporate the advantages of site-specific models of evolution [203, 196, 204]. Another major class of phylogenetic analysis, based on the principle of maximum parsimony, will not be considered here because parsimony methods are far slower than distance methods, and they do not accurately model evolutionary processes despite having the same biases and inaccuracies as distance methods. The computational limitations of likelihood-based methods become far more severe with large amounts of sequence data from highly diverse sets of organisms [205, 206, 65]. For example there are $2.75 * 10^{76}$ possible topologies relating 50 taxa, making exhaustive approaches impossible [143]. Branch-and-bound searches can reduce the tree space to be examined for smaller trees but are insufficient for large datasets because the number of tree topologies is still too large [207]. Thus, heuristic searches must be used for large trees, evaluating trees that are proximal to reasonably likely trees that have already been found. These searches are currently often performed using branch-swapping algorithms such as nearest-neighbor interchange (NNI), subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR) (e.g. [152, 208]). The number of NNI, SPR, and TBR neighbors of any topology increase respectively as linear, quadratic, and cubic functions of the number of taxa, and the trees proposed are not necessarily of similar likelihood to the known tree. Therefore, many highly improbable trees are evaluated in branch-swapping algorithms, and the correct solution is not guaranteed due to the presence of local optima in tree space [199]. Branch length optimization (or posterior equilibration) must also be performed after branch swapping and is an additional source of computational cost.

Several heuristic approaches have been developed to release tree searches from local optima. Ratchet methods employ multiple initial trees perturbed by bootstrap resampling to ensure a less-overlapping tree space in subsequent optimizations using branch swapping [209, 210]. The partial stepwise addition (PSA) approach enables escape from local optima by removing some taxa during the topology search [211]. Simulated annealing (SA; [212]) and Metropolis-coupled Markov chain Monte Carlo (MCMCMC; [213]) manipulate a likely

range of proposed tree acceptances in a single heuristic search or in multiple interacting chains, respectively. Genetic algorithms (GAs) simulate the population dynamics of tree topologies using likelihood as a fitness parameter [214]. These methods outperform simple heuristic searches in at least some contexts.

Many methods for sampling tree space have been developed for Bayesian tree selection also, and most build off of the previously mentioned tree search methods. These methods include Narrow Exchange, Wilson-Baldwin, and Intermediate Exchange [215, 216, 217]. A good review of the efficiencies of Bayesian search methods was published by Lakner et al. in 2008 [218]. All approaches listed above employ branch swapping to explore tree space and therefore suffer from inefficiency due to the decoupling of topology proposals from the likelihoods of the topologies.

Here we consider whether the beneficial features of Bayesian analyses under relatively complex models can be profitably combined with the speed of distance methods based on relatively simple models. The key to our approach is that rather than using branch swapping to explore phylogenetic tree space, distance-based trees predicted from partially sampled sequences are used. We use Markov chain Monte Carlo and a Metropolis-Hastings algorithm in which new steps in the chain are proposed based on bootstrap resampling a proportion of the current sequence sample. Heuristic phylogenetic trees based on the new sample are created using NJ, and the likelihoods of the new trees are evaluated using the full sequence dataset and the mtMam model [141]. The unwanted sampling distribution induced by the NJ proposal mechanism is estimated by running the proposal mechanism without calculating the likelihoods of the proposed trees. The posterior is then corrected for this sampling distribution. We evaluated the effect of different site sample sizes used to generate the NJ trees (sample size) and different resample proportions (jump size).

A bootstrap sampling procedure was employed to sample sites in the alignment that were then used to calculate distance matrices [219, 220]. Although complete and partial bootstrapping has been used extensively in phylogenetic studies to evaluate branch support

and tree confidence, we used it solely to generate a broad distribution of reasonably likely trees based on the NJ heuristic [221, 222]. Note that while partial sampling is more common when employing the related jackknife approach, bootstrapping approaches such as that employed here sample with replacement, rather than without replacement as in the jackknife. Depending on the number of sites sampled (the sample size), trees produced from partial sequence samples can be quite different from the ML tree of the entire alignment and considerably less likely [106]. Evaluating the posterior distribution with an importance sampling approach using these trees is not feasible because the extreme variation in likelihoods among trees means that a few trees would dominate the weighted importance sampling average [223]. Instead, it is necessary to use a progressive Markov chain approach to evaluate the posterior, such as the Metropolis-Hastings algorithm, in which the proposed sample depends on the current sample [174]. Only a fraction of sites is resampled in each generation of the chain. The NJ tree generated from the proposed sample updates both branch lengths and topology simultaneously, and the likelihood of this proposal was then calculated on the full alignment. The number of sites resampled was uniform randomly chosen up to some maximum, which we will call the "jump size".

VI.4 Glossary

Extant species - A species that is currently alive, as opposed to extinct. Extant species are often the leaves on a phylogenetic tree and the residue sequences are often considered known.

Speciation - The splitting of one species into two. These events are usually represented by nodes on a phylogenetic which have two descendant branches.

Phylogenetic tree - A history of speciation events from the root (most recent common ancestor) to the extant species on the leaves.

Ancestral species - A historical species on a phylogenetic tree which is closer to the root.

Descendant species - A species on a phylogenetic tree which is closer to the extant

species.

Branch - A connection between speciation events on a phylogenetic tree. Branches often have a length expressed in terms of the expected number of substitutions per site.

Multiple Sequence Alignment(MSA) - The sequences of the same region of a genome (protein, RNA, transposable element) from multiple species, aligned such that each row is a different species and each column is an orthologous site. Each site is presumed to

Acceptance Probability(α) The likelihood that the proposed next step in a Markov chain is accepted. In Markov Katana it is the probability that a proposed tree topology will be accepted by the chain. Tree topologies with higher likelihood than the current topology are always accepted if proposed. Lower likelihood topologies are accepted with the probability of ratio of the proposed topology likelihood divided by the current topology likelihood. Acceptance probability is denoted by α in Figure 1.

Genealogy In this context, genealogy comprises the phylogenetic tree topology and the branch lengths along that tree.

Robinson–Foulds metric(RF) A distance measure between two phylogenetic tree topologies, considering how many branches are found in one topology and not the other. This distance metric is also twice the Nearest-Neighbor Interchange distance.

Nearest-Neighbor Interchange(NNI) A method of generating a new tree topology proposal by swapping two subtrees across a branch. NNI can be used as a tree topology distance metric between two topologies, indicating how many NNI swaps would be required to convert one topology into the other.

VI.5 Materials and Methods

VI.5.1 Mitochondrial Sequences

The 629-taxon mitochondrial gene alignment from Goldstein et al. was used for this study [48]. Three data sets were produced from these sequences by arbitrarily selecting 10, 20, and 50 sequences from this alignment and the selected taxa names are shown in

Supplemental Tables VI.1, VI.2, VI.3, and VI.4, respectively. For the 10 taxa alignment, only the 495 amino acid Cytochrome C Oxidase subunit 1 (COI) sequences were used out of the entire 13 gene alignment, in order to reduce the information used to make the tree inference. When all 13 genes in the alignment are used, the posterior of a single tree is over 99%, but we wanted to test this method on a data set where the correct tree is not easy to identify. For the 20 and 50 taxa alignments, the first 1000 columns of the entire mitochondrial alignment were selected. Mitochondrial sequences were used in order to avoid lineage sorting complications, for the large number of taxa sequenced, and for the ease of identifying orthologs.

VI.5.2 The Markov Katana Algorithm

The Markov Katana algorithm begins by selecting a random sample of sites from the entire multiple sequence alignment (MSA, A, see algorithm 1). The number of sites sampled is the sample size, f (usually expressed as a percentage), times the number of sites in the MSA. This sampling with replacement, similar to the jackknife, is why we chose to call the method a "katana". An initial neighbor-joining genealogy G_0 is generated from this sample. The neighbor-joining algorithm is sped up by precalculating the distances among the extant sequences for each site independently. Then when the algorithm calculates the total distance between taxa, it simply sums the distances from the sites in the sample.

Then the algorithm enters the main body of the MCMC loop. For each generation of the MCMC, the sample of sites is resampled. Resampling involves randomly selecting a site from the sample to be replaced by a random site from the MSA. This process is repeated a uniformly distributed number of times up to the jump size j. A neighbor-joining tree is proposed given the new sample of sites, and a likelihood is calculated for that new tree given the entire alignment. The likelihoods of the old tree and the new tree are compared, and the new tree and sample are kept according to the Metropolis-Hastings algorithm [174].

Given alignment A, sample size f, sequence sample S, genealogy G, likelihood $\mathcal{L}$, K generations, jump size j
$S_0 \leftarrow \texttt{Sample}(A, f)$
$G_0 \leftarrow \texttt{NJ}(S_0)$
for $k \leftarrow 1$ **to** K **do**
  $\tilde{S} \leftarrow \texttt{Resample}(S_{k-1}, A, j)$
  $\tilde{G} \leftarrow \texttt{NJ}(\tilde{S})$
  With probability $\alpha = \min\{1, \frac{\mathcal{L}(\tilde{G},A)}{\mathcal{L}(G_{k-1},A)}\}$, set $S_k \leftarrow \tilde{S}, G_k \leftarrow \tilde{G}$
end

Algorithm 1: Markov Katana algorithm

VI.5.3 Posterior Calculations

To obtain the posterior, the uncorrected distribution of trees after the initial Markov Katana run must be corrected for the bias induced by the proposal mechanism. Since the sampling procedure generates genealogies based on a sample of the MSA, the trees produced will not be uniformly distributed. In these runs, the sample size as a percentage, f, and the jump size, j, were variable parameters and differed among runs as specified. For a given sampled generation, k, the alignment sample at that generation produced a NJ genealogy, G_k, with topology, T_i, where i indicates the index of the topology. The proportion of times that each different topology was produced by the chain out of K sampled generations in the chain is an estimator of the uncorrected posterior $\hat{U}_f(T_i)$ for a given sample size, f, or

$$\hat{U}_f(T_i) = \frac{1}{K} \sum_{k=1}^{K} \delta(T_i, G_k) \tag{VI.1}$$

where $\delta(T_i, G_k)$ is a delta function equal to 1 if G_k has topology T_i and otherwise 0. To obtain the corrected topology posterior, $C(T_i)$, we first estimate the topology sampling bias $\hat{\beta}_f(T_i)$ induced by NJ proposals with sampling fraction f, by sampling K' genealogies from a separate chain in which all proposals are accepted, to obtain

$$\hat{\beta}_f(T_i) = \frac{1}{K'} \sum_{k'=1}^{K'} \delta(T_i, G_{k'}) \tag{VI.2}$$

We note that this procedure is identical to obtaining a NJ partial bootstrap, but by running the Markov chain with a given jump size, we can obtain the connectedness among topologies, providing a natural topological distance measure.

We then recognize that

$$U_f(T_i) \propto \beta_f(T_i) \int L(G_k)P(G_k) \bullet \delta(T_i, G_k)dk = \beta_f(T_i)C(T_i) \tag{VI.3}$$

where $L(G_k)$, $\beta_f(G_k)$, and $P(G_k)$, are the likelihood, the genealogy sampling bias induced by NJ proposals with sampling fraction f, and the prior, respectively. Here we assume a flat prior across all tree topologies, and we integrate over all proposed branch lengths within a topology. The next step is to divide the uncorrected topology posterior by the sampling distribution induced by the proposals to obtain

$$\hat{C}(T_i) \propto \frac{\hat{U}_f(T_i)}{\hat{\beta}_f(T_i)} \tag{VI.4}$$

We normalized the corrected posteriors by dividing by the sum of all corrected posteriors over all topologies sampled. We can do this division because the sampling distribution is very broad relative to the uncorrected posterior distribution. In practice, the division improves the posterior estimate slightly, for our set of sequences.

When estimating the topology posteriors in this way, it is possible for the sampling distribution prior probabilities to bias the uncorrected posteriors in such a way that the division does not result in a good estimate of the posteriors. One change to the method which would result in an improved posterior calculation rather than estimates would be to include the sampling distribution priors into the Metropolis-Hastings algorithm step in the Markov chain Monte Carlo. The typical acceptance probability α for the genealogy $\tilde{G}$ given the previous genealogy G_{k-1} using the Metropolis-Hastings algorithm is:

$$\alpha = min\{1, \frac{\mathcal{L}(\tilde{G}) * prior(\tilde{G})}{\mathcal{L}(G_{k-1}) * prior(G)}\} \tag{VI.5}$$

For the Markov Katana algorithm proposed in this paper, the priors for each tree topology are considered equal during the MCMC run and the posterior is corrected for differing sampling distribution priors after the MCMC has been run. One could, however, estimate the sampling distributions for each topology before of the posterior MCMC run, and then use those proper sampling distribution priors in the posterior estimation.

There are a number of problems that could arise from this correction method. First if during the course of the posterior estimation, a tree topology is proposed for which a prior has not been sampled, then one would have to provide an estimate of that topology's prior somehow. One could have a fallback prior for all topologies that were not sampled during the prior estimation step, however this would bias the posterior calculations. Second, retrieving the correct precalculated sampling distribution prior of a tree topology requires calculating the topology ID for every step in the MCMC. Calculating the topology ID is a non-trivial operation requiring recursion through the entire tree structure, and so it would add to the computation required at every generation of the MCMC, potentially slowing it down considerably.

VI.5.4 Program Details

A Perl program, Markov Katana, was written to implement the Markov chain bootstrapping algorithm. Markov Katana takes multiple sequence alignments in the FASTA format and outputs phylogenetic trees in the Newick format, along with likelihood values. Another program Forest was written to analyze the trees generated by Markov Katana to calculate tree and branch frequencies. Markov Katana and Forest were tested on and are compatible with current Unix-based operating systems as well as Windows. The program PAML was used to calculate the likelihoods for the trees using the entire alignment of 495 amino acids [172]. The data and code can be found at https://github.com/PollockLaboratory/MarkovKatana.

VI.5.5 Modifying Implementation of NJ in Markov Katana to Improve Branch Length Estimation

In initial runs, the NJ algorithm often generated unrealistically short branches, so to counteract this we lengthened the shortest branches by adding a random number from 0 to 2 substitutions (a branch length increase of 0 to 2/495). This limited the effect of these implausibly short branches in the proposal mechanism. Short branches were still possible, but extremely short branches were not as likely to be proposed.

VI.5.6 Branch Calculations

It may sometimes be useful and possibly more accurate to calculate branch (a.k.a. a species bi-partition, or edge) posteriors $\tilde{U}_f(B_l)$ directly over the sample of trees,

$$\tilde{U}_f(B_l) = \frac{1}{K} \sum_k \delta(B_l, G_k) \tag{VI.6}$$

where $\delta(B_l, G_k)$ is a delta function equal to 1 if G_k has branch B_l, and otherwise 0. The $\tilde{}$ symbol indicates that the branch uncorrected posteriors were calculated directly. In this case, it is necessary to appropriately adjust for the sampling distribution on the branch induced by topological constraints [224], which is contained in both $\tilde{U}_f(B_l)$ and a similarly obtained

$$\tilde{\beta}_f(B_l) = \frac{1}{K'} \sum_{k'=1}^{K'} \delta(B_l, G_{k'}) \tag{VI.7}$$

This prior is put back into the posterior calculation as,

$$\hat{C}(B_l) \propto \frac{\tilde{U}_f(B_l)}{\tilde{\beta}_f(B_l)} * P(B_l|N, s_l) \tag{VI.8}$$

where $P(B_l|N, s_l)$ is the prior probability of branch B_l induced by topological structures, N is the total number of extant species in the tree, and s_l is the smaller number of species that are partitioned to one side of branch B_l. $P(B_l|N, s_l)$ can be calculated directly (see Methods).

Branch priors can be calculated as

$$P(B_l|N, s_l) = \frac{T_s^r * T_{N-s}^r}{T_N^u} \tag{VI.9}$$

where T_x^r and T_x^u are respectively the number of possible rooted and unrooted topologies with x taxa, N is the total number of taxa being evaluated, and s is the smaller number of taxa that are segregated on one side or the other of branch B_b [224].

VI.5.7 Branch Length Approximations

When comparing the probabilities of topologies with respect to a multiple sequence alignment, often the maximum likelihood for the topology is used by computing the maximum likelihood estimates for the branch lengths. Markov Katana is estimating the posteriors for each topology and so ideally the likelihood would be integrated (marginalized) over all possible branch lengths. This is infeasible for a method such as Markov Katana, since this integration of the likelihood would be very computationally expensive and would make the method so slow as to be unusable. Markov Katana is often run for hundreds of thousands or millions of generations, and so integrating over branch lengths for each proposed tree would be too computationally expensive.

Instead of full integration, we allow the algorithm to sample from the branch lengths. With enough samples, the approximation approaches the full integral of the likelihood. We consider each set of branch lengths sampled to be a point in the multidimensional branch length space over which we are approximating the integral. We assume that the likelihood at that point is representative of the area around this point. In this way, we can approximate the integral for the high likelihood parts of the multidimensional branch length space by sampling. Since the proposed trees are generated from a neighbor joining algorithm based on a subset of the alignment, it is unlikely that the branch lengths are optimal for the tree given the entire alignment. However for the trees which are high likelihood and therefore proposed often, Markov Katana will produce an effective approximation of the integral across the likely branch lengths. The highest likelihood

Leaves	Unrooted topologies
10	$2.02 * 10^6$
20	$2.22 * 10^{20}$
50	$2.75 * 10^{76}$

Table VI.1: The numbers of distinct binary unrooted topologies for different numbers of leaves.

trees will have the best approximations to the integral, while little time will be spent approximating the integral for lower likelihood trees.

VI.5.8 Sampled tree space

Ideally when sampling using a MCMC, the transition method employed would be ergodic, or could jump from any state to any other state in a single step. Every state would theoretically be reached eventually. In terms of tree topology sampling, this means that the proposal mechanism would reach all possible tree topologies if given enough time. The number of possible topologies of unrooted binary trees with n leaves is [143]:

$$T(n) = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{VI.10}$$

The numbers of possible topologies for 10, 20, and 50 taxa are listed in table VI.1. For all alignment sizes except 10, the number of trees actually tested during a typical Bayesian phylogenetic tree search is far smaller than the number of possible topologies to be tested.

The size of the tree topology space possibly sampled by the Markov Katana method depends on the number of combinations of sites in the sampled multiple sequence alignment, resulting in the maximum number of topologies in the sample space being $T_{MK} = (f * M)^M$ where f is the sample size and M is the number of sites in the multiple sequence alignment. The actual sample space is probably much smaller since it is unlikely that every combination of sites will map to a unique topology.

VI.6 Results

VI.6.1 Estimating Topology Posteriors

We began by analyzing a 10-taxon Cytochrome C Oxidase subunit 1 (COI) amino acid alignment (495 residues) that was chosen so that there would be a moderate level of topological uncertainty in the posterior (Fig. VI.1). Preliminary evaluations indicated that NJ trees on bootstrapped data have a distribution of topologies that are relatively similar among distance types (Supplemental Figure VI.1). Although there is considerable noise to the estimates for very small frequencies, and there is a slight shift towards higher frequencies with the Markov Katana difference NJ, overall the two measures have a nearly linear relationship. This gave us confidence that NJ trees based on differences rather than corrected distances might be sufficiently accurate for our purposes, so to keep the NJ calculations as simple and fast as possible for initial testing, distances were generated using the simple difference matrix. The likelihoods of the proposed tree topologies were then evaluated using the mtMam substitution rate matrix model on the entire sequences [141]. Continuing to keep things simple for initial testing, we used a flat prior on branch lengths, although we imagine that most future implementations will want to incorporate other priors here, such as the commonly used exponential priors on branch lengths [225].

To understand the differences in topology sampling bias estimates obtained using different sample sizes, f, Markov Katana was run with sample fractions ranging from 100% (495 sites) down to 20% (99 sites). We chose 100% (495 sites) sample size bootstraps for the NJ proposals along with a jump size of 10% (which is 50 sites) as standard reference conditions, which had acceptance probabilities of about 30%. The topology sampling biases for smaller f become somewhat more even, with the least frequent topologies about 10x more frequent for $f = 20\%$ than for $f = 100\%$ (Fig. VI.2). At the same time, the number of topologies with sampling probabilities greater than 10^{-6} increased from 5,975 for $f = 100\%$ to 21,198 for $f = 20\%$. Predictably, comparisons of replicate sampling distribution runs indicated an increasing variance in estimated biases with decreasing

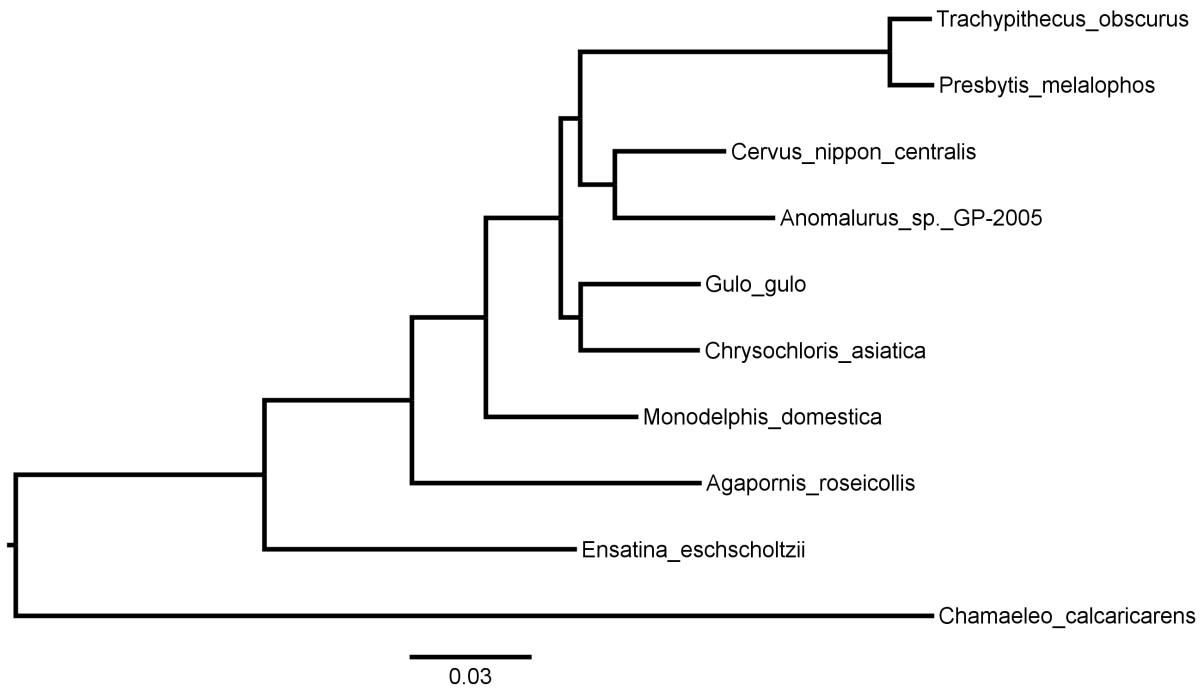


Figure VI.1: An example phylogenetic tree for the 10-taxon dataset. A tree for the 495 amino acid Cytochrome C Oxidase subunit 1 (COI) alignment used in initial analyses. The tree was generated by the neighbor-joining algorithm in Markov Katana. Given the intentionally limited data used here for the purpose of testing the Markov Katana method, this tree should not be interpreted as being a "true" or species tree.

probabilities in the runs (Sup. Fig. VI.2).

The posterior correction (Equation VI.4) appears to work well across a broad range of sample sizes (Figure VI.3). The corrected topology posteriors for sample fractions from 20% to 95% were all highly correlated with the topology posteriors for sample size 100%. It should be noted that the uncorrected posteriors are only slightly less correlated with each other than are the corrected posteriors (Supplemental Figure VI.3), meaning that the answer would have been similar without the correction. It is probably best to use the correction anyway, because in more complicated situations it may make more of a difference, and it is not too much trouble to obtain and is correct.

The corrected posterior estimates appear to be most noisy when the sampling distribution estimate is small and therefore poorly estimated. This is not entirely surprising given that the sampling distribution is in the denominator. Because the sampling distribution calculations are computationally inexpensive (they do not require a likelihood calculation), it is possible to obtain a couple orders of magnitude more data for them than for the uncorrected posterior estimates. While estimating the sampling distribution more precisely is important for the correction, many of the trees examined have topologies that are not found in the posterior. A potential means to increase accuracy of relevant topologies in the sampling distribution is to limit the sampling distribution chains to those topologies seen in the uncorrected posterior.

In order to verify that the method is estimating the posteriors well, we compared the corrected posteriors from Markov Katana with the posterior estimates from MrBayes (Fig. VI.4) [151, 152]. MrBayes was run using the mtMam model for 100,000 generations sampled every 10 generations, and a summary of the tree probabilities was generated using a 25% burnin. The MrBayes run block is given in Supplementary Figure VI.4. The estimates generally agree, and Markov Katana does not show significant bias either high or low. Many of the trees sampled by Markov Katana were not sampled by MrBayes, and only trees sampled by both are shown in Figure VI.4. The topology posteriors were also

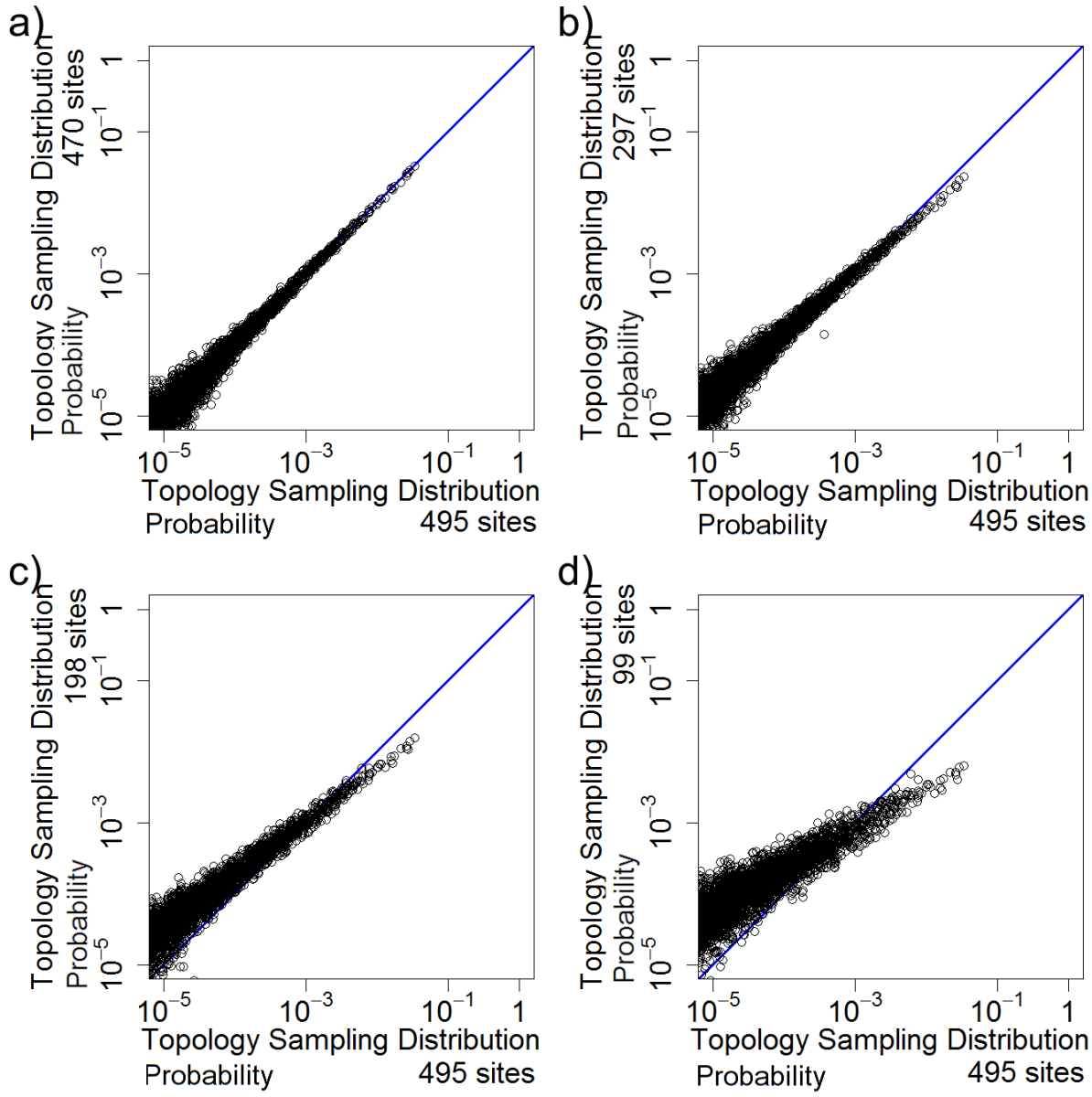


Figure VI.2: The topology sampling distribution becomes more evenly distributed as the sample size decreases. Markov Katana was run using different sample sizes and the estimates of the topology probabilities in the sampling distributions for different sample sizes and the same jump sizes are shown. Each point is a tree topology found in all the sampling distributions. The sampling distributions of sample size a) 95%, b) 60%, c) 40%, d) 20% are compared against the sampling distribution of sample size 100% on the x axis. The blue line indicates where x and y values are equal. As the sample size decreases, the sampling distribution compared to 100% becomes more flat and evenly distributed across all topologies. The sampling distributions were averaged over triplicate runs of 1,000,000 generations and used a jump size of 10%.

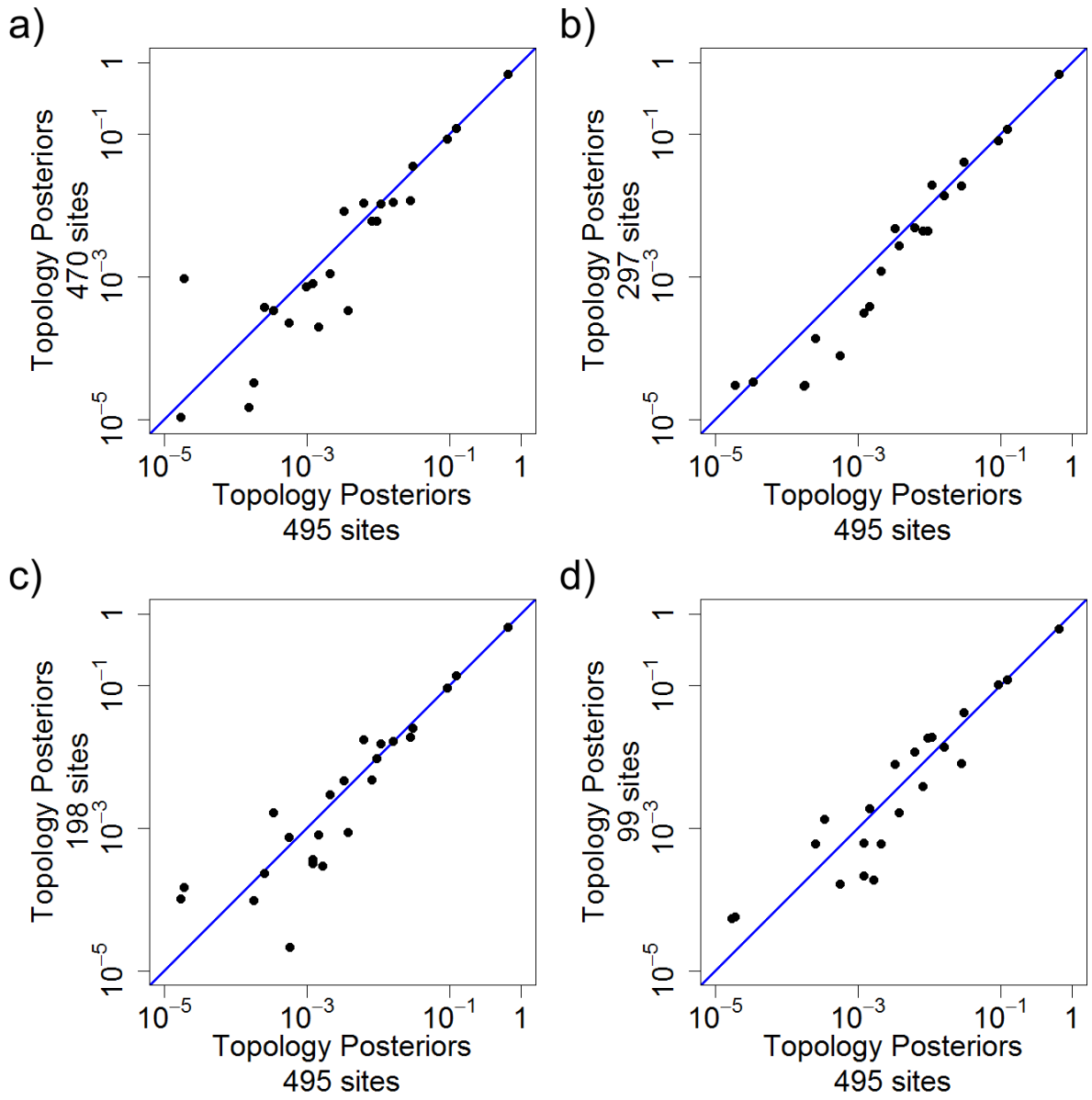


Figure VI.3: The corrected tree topology posteriors calculated using different sample sizes are linear over a wide range of different sample sizes. The corrected topology posterior distributions of sample fractions a) 95%, b) 60%, c) 40%, d) 20% are compared to the corrected topology posterior distribution of sample fraction 100% on the x axis. The blue line indicates where x and y values are equal. The topology posteriors are highly correlated and very linear among all the different sample sizes tested, despite having to correct for the differing sampling distributions of the different sample sizes. The differences among sampling distributions is shown in Figure VI.2. The uncorrected posteriors were averaged over triplicate runs of 50,000 generations and used a jump size of 10%.

estimated using BEAST, however the results differed so much from MrBayes that they are not shown here [226]. The differences in posteriors are apparently due to differences in implementing the model and not allowing for the the same options, especially in the tree searching methods employed.

VI.6.2 Effect of Sample Fraction and Jump Size on the Markov Chain

Although the posterior estimates were comparable for all sample sizes, it is still worthwhile to consider the effect of both sample size and jump size on the mixing efficiency of the Markov chain. One important factor in considering the mixing efficiency of the Markov chain is the acceptance probability. The acceptance probability in Markov Katana is the likelihood that a proposed tree topology will be accepted by the chain. The acceptance probability is defined by α in Algorithm VI.1. If the acceptance probability is too high, then the MCMC will explore the tree topology space very slowly and may never reach the high likelihood trees. One possible fix for a slowly moving MCMC is to propose moving to tree topologies which are more different than the current topology. If the acceptance probability is too low, then many low likelihood trees will be proposed and the high likelihood trees again may be undersampled. Reducing the distance between the current tree and the proposed tree helps increase the acceptance probability.

Acceptance probabilities varied widely depending on the jump size. We ran Markov Katana with many different jump sizes. In general, a jump size of 5-10 sites appears to be a minimum in order to ensure movement across topologies, and 50 sites is probably a maximum for this data set. The jump size directly determines the maximum number of sites resampled in any given generation of the MCMC, and the acceptance probability varies strongly with the number of sites resampled. In order to visualize how the acceptance probabilities Markov Katana method change with the number of sites resampled, Markov Katana was run on the same 495-amino acid and 10-taxon data set with different sample fractions. The effect of the number of sites resampled on acceptance probability is shown for two different sample sizes in Figure VI.5. Although in this example the jump size is

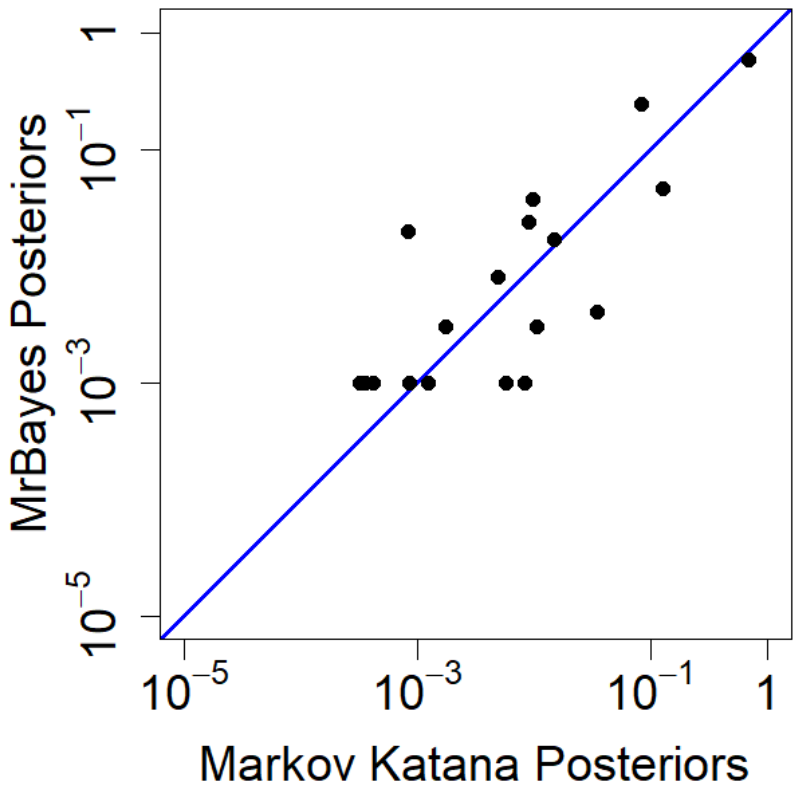


Figure VI.4: In order to test the accuracy of the Markov Katana method, the Markov Katana corrected tree topology posterior estimates are compared with posterior estimates from MrBayes for the same 10-taxon data set. MrBayes is a commonly used tool for phylogenetic tree search and uses a Bayesian approach like Markov Katana. The posteriors show a strong correlation and linearity, despite MrBayes sampling many fewer trees than Markov Katana, indicating that Markov Katana can estimate the correct posteriors of tree topologies for this data set. The blue line indicates where x and y values are equal.

held fixed at 10%, since the sample sizes varied from 100% to 80%, the number of sites resampled varied. Overall the acceptance probability decreased with a higher number of sites resampled. The acceptance probability also dropped faster for the 80% sample size run than the 100% sample size run. With smaller sample sizes (e.g., 80% shown here), the jump size is a larger proportion of the sample and reduces the acceptance probability more rapidly.

We also considered the effect of jump size on both the sampling distribution and the uncorrected posterior Markov chain estimates. For the initial sampling distribution estimation procedure, the most well-mixed chain is of course the one with independent bootstraps (j=100%), but the chain also mixes well with lower jump sizes. It is necessary to have smaller jump sizes because a high proportion (99.9%) of the random samples are not in the uncorrected posterior topology set. For this analysis, the optimal jump size was j=85%. This result did not differ much for a range of sample sizes. Although differing in detail, the jump size analysis for the uncorrected posterior had similar results to the biased sampling distribution analysis.

In order to show how the jump size alters the tree topology proposals, we analyzed the distances between current topology and the proposed topology for every generation in two Markov Katana runs with different jump sizes (10% and 70%, see Figure VI.6). The distance between tree topologies can be measured in a number of ways, but for this figure, we chose to use the Robinson–Foulds (RF) metric, which considers how many branches are found in one topology and not the other. In the Markov Katana runs, we counted every time a new topology was proposed and when it was accepted, then split this count by the RF distance between the current and proposed topologies. Rejected proposals and proposals to the same topology (e.g. the current topology and the proposed topology are the same, but may have different branch lengths) are considered to have a RF distance 0. Increasing the jump size in general increases the average Robinson–Foulds distance between jumps and proposals, as shown by parts c) and d) of Figure VI.6. In

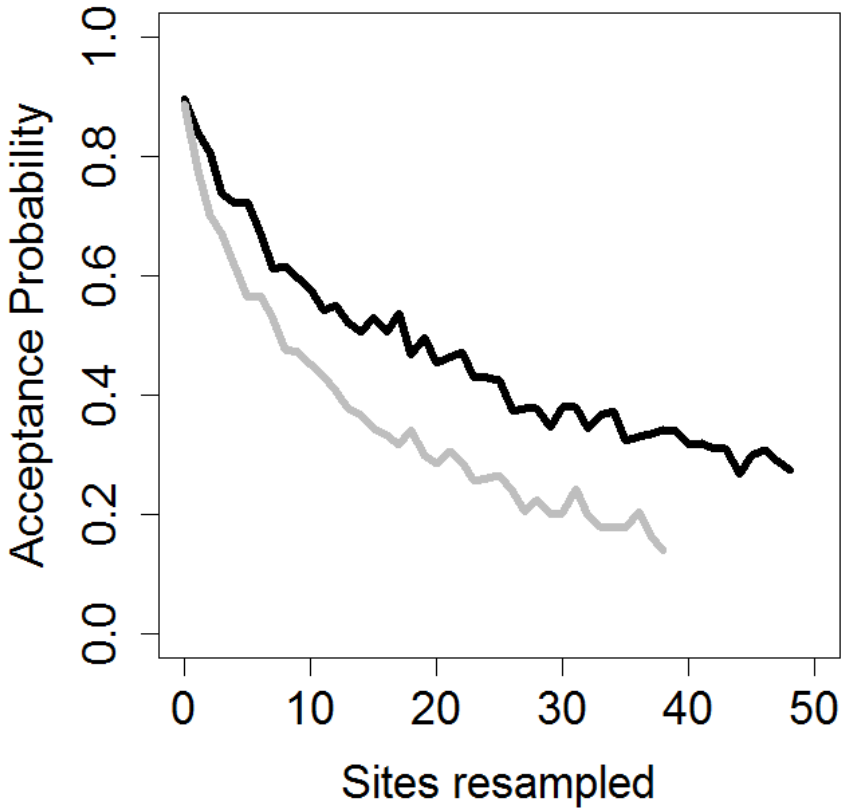


Figure VI.5: In order to visualize how the acceptance probabilities Markov Katana method change with the number of sites resampled, Markov Katana was run on the same 495-amino acid and 10-taxon data set with different sample fractions. The average acceptance probabilities are shown for proposals that resampled different numbers of sites. The bootstrapped sample fraction for neighbor-joining (NJ) proposals was 100% for the black line and 80% for the gray line. The jump size was 10% for both runs. The gray line decreases faster because each additional resampled site is a larger fraction of the total sample size. Acceptance probabilities were determined by the average of 3 independent 50,000-generation runs of Markov Katana.

the 10% jump size run, the largest distance between the current topology and proposed topology was 8, while in the 70% jump size run, the largest distance was 12. The whole proposal distribution is shifted toward higher distance proposals in the 70% jump size run, indicating that a larger jump size does propose more different tree topologies. The number of acceptances for proposals of each RF distance were remarkably similar for the 10% and 70% jump size runs, with only a slight increase in the number of proposals with a distance of 8 in the 70% run.

VI.6.3 The Structure of Tree Space

In order to visualize the tree topologies and the distances between them, a network representation of the 12 tree topologies that had a posterior of 0.001 or higher was constructed for Figure VI.7. Every topology is shown as a node in the network. The size of each node represents the relative posterior of the topology, and the edges of the graph indicate Nearest-Neighbor Interchange (NNI) distances of one (black) or two (blue) between the tree topologies. The tree topology space of this test data was clearly divided into two clusters of trees shown by the intragroup connections and the few intergroup connections. Given the connectivity of the network, other tree topology sampling procedures may have difficulty jumping between groups. A table of the posteriors of the labeled tree topologies is shown in Table VI.2.

VI.6.4 Estimating Larger Trees

In order to test Markov Katana's ability to estimate trees with more than 10 taxa, Markov Katana was run on two larger data sets with 20 taxa and 50 taxa and 1000 sites from the mitochondrial alignment. Multiple runs of Markov Katana on the 20 taxon alignment converged to similar uncorrected posteriors as shown in Figure VI.8.

The corrected posteriors averaged over three Markov Katana runs with sample size of 100% and jump size of 10% are compared to the topology posteriors from MrBayes in Figure VI.9. The agreement between the Markov Katana and MrBayes is not good. The two tools use different branch length approximation methods, which might account for

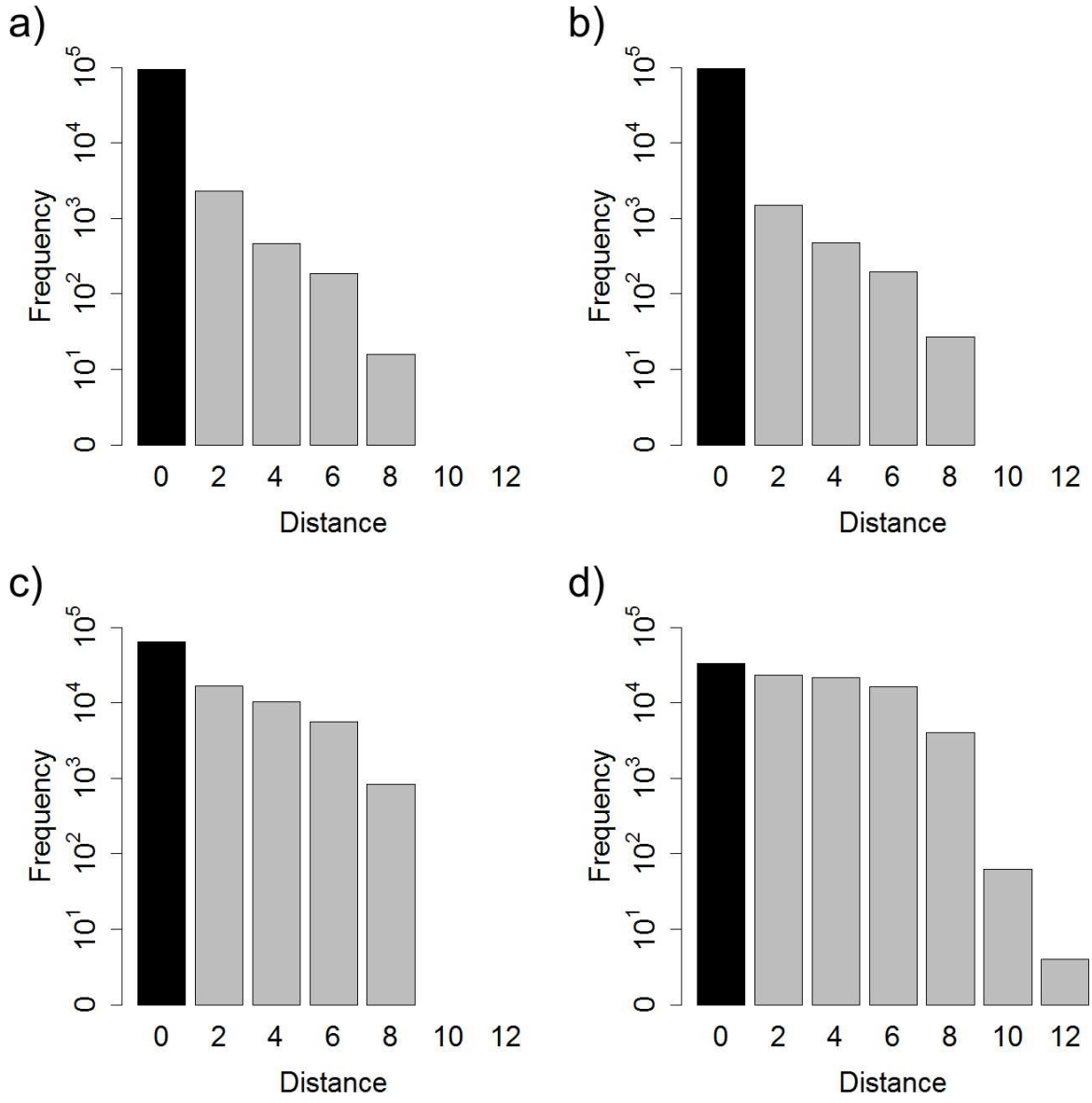


Figure VI.6: In order to show how the jump size alters the tree topology proposals, we analyzed the distances between current topology and the proposed topology for every generation in two Markov Katana runs with different jump sizes. Increasing the jump size in general increases the average Robinson–Foulds (RF) distance between jumps and proposals, which considers how many branches are found in one topology and not the other. Parts a) and b) show the distance between trees in jumps (*accepted* proposals) for jump size 10% and 70% respectively. Parts c) and d) show the distance between trees in proposals for jump size 10% and 70% respectively. Rejected jumps are considered jumps of distance 0. The whole proposal distribution for the 70% jump size run is shifted toward higher distance proposals compared to the 10% jump size run, indicating that a larger jump size does propose more different tree topologies. The number of acceptances for proposals of each RF distance were remarkably similar for the 10% and 70% jump size runs, with only a slight increase in the number of proposals with a distance of 8 in the 70% run.

Label	ID	Posterior
A	16405	0.6924
B	78835	0.0670
C	36915	0.0659
D	92575	0.0480
E	26545	0.0385
F	57985	0.0370
G	80055	0.0285
H	39655	0.0104
I	2955	0.0028
J	82665	0.0026
K	8085	0.0024
L	88595	0.0009

Table VI.2: Posterior probability for topologies with substantial representation (greater than 0.001) in the uncorrected posterior for the 10-taxon dataset. The topologies are labeled for reference in Figure VI.7.

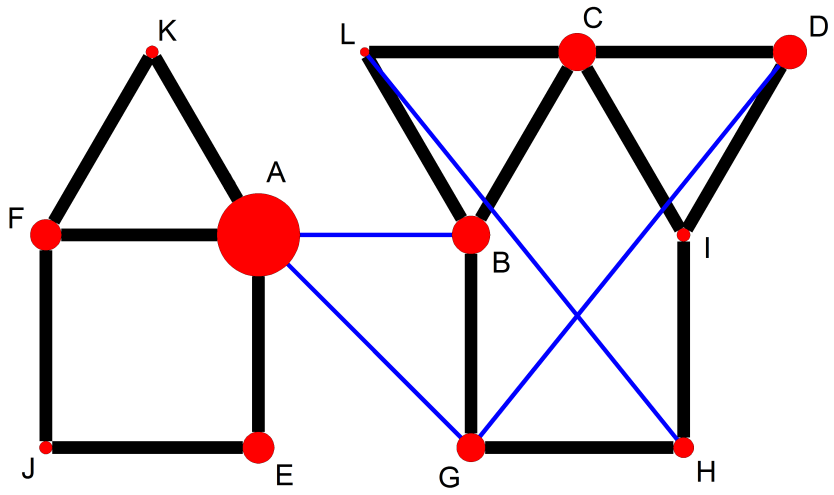


Figure VI.7: The high posterior tree topologies are shown as a network with connections indicating distance between topologies. Topologies shown have > 0.001 posterior probabilities. The topologies are labeled A through L as in Table VI.2. The size of the circle shows the relative posterior probability. Black and blue lines indicate distances of 1 and 2 nearest-neighbor interchange (NNI) (RF distance of 2 and 4), respectively. There are two groups of tree topologies with high connectivity within then and few connections between them. One group is tree topologies A, E, F, J, and K (shown on the left) and the other group is topologies B, C, D, G, H, I, and L (shown on the right).

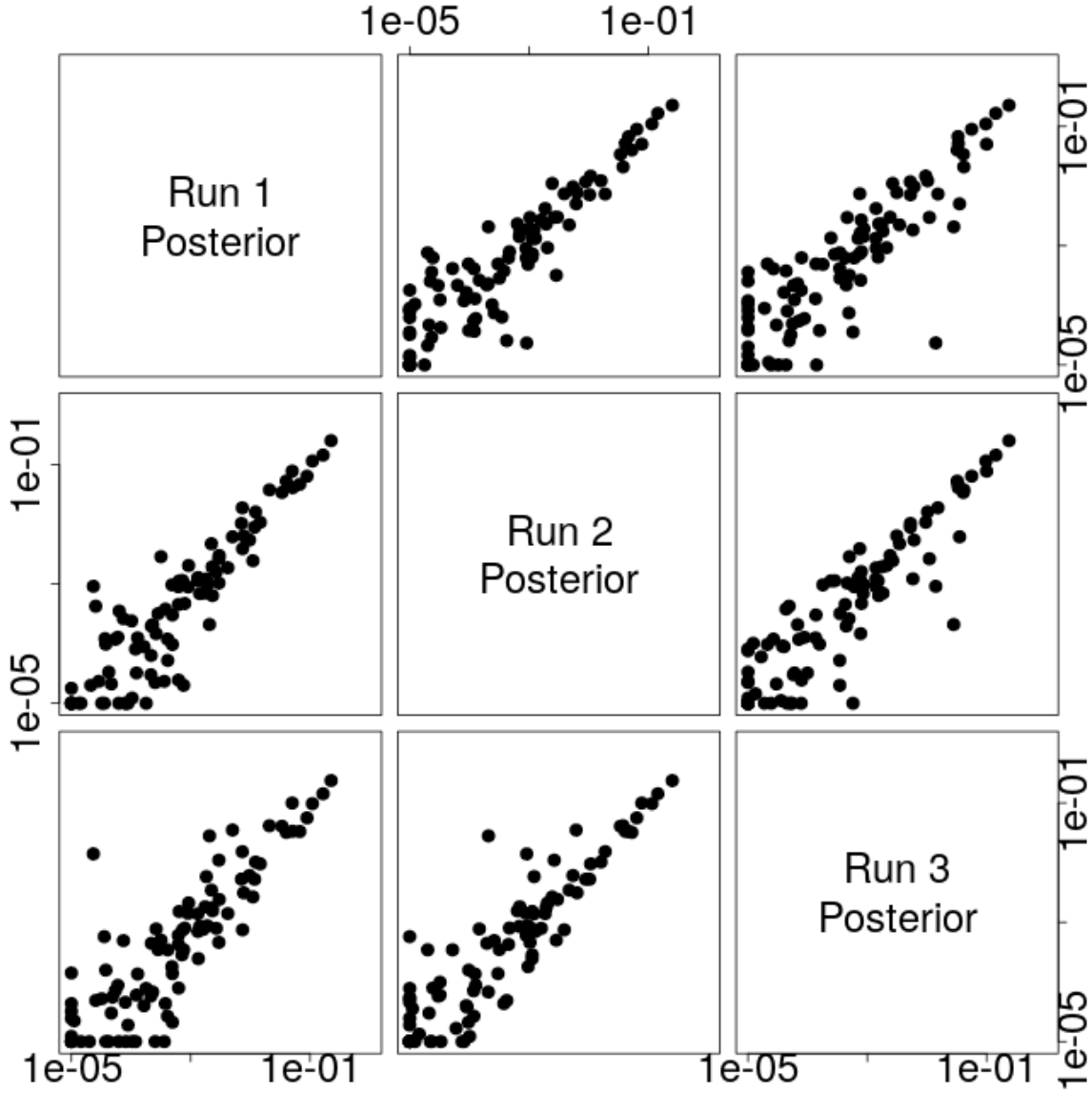


Figure VI.8: Uncorrected posterior estimates are similar across multiple runs of Markov Katana on a 20 taxon alignment. Markov Katana was run using a sample fraction of 100% and a jump size of 10%. The uncorrected posteriors were averaged over triplicate runs of 1,000,000 generations. Tree topology posteriors below 10^{-5} were set to a minimum value of 10^{-5}.

Generations	Sample Size	Jump Size	Effective sample size	Time (h:m:s)
50,000	60%	10%	370	24:16:14
100,000	100%	10%	497	40:37:27
100,000	100%	70%	400	39:36:26

Table VI.3: The computational performance of Markov Katana is shown here with a few example run times for the 10 taxon data set. The run times are robust against increasing the sample size and the jump size for these runs.

Generations	Sample Size	Jump Size	Run Time
1,000,000	100%	10%	5 days 13:01:38
1,000,000	50%	10%	4 days 03:25:05
1,000,000	100%	1%	5 days 12:45:08

Table VI.4: Example run times for the 20 taxon tree

some of the discrepancy. The correlation could possibly be improved if MrBayes produced topology posteriors down to 10^{-5}.

The run time required for convergence on 20 taxa was substantially higher than for the 10 taxon alignment. The neighbor joining algorithm used scales as N^2 with the number of taxa, and so the run times for fitting 20 taxa were around five and a half days, rather than the one day required to estimate the 10 taxon tree. Example run times for estimating the 20 taxon tree are provided in Figure VI.4.

Due to the quickly expanding run times with the number of taxa, the run time to fit the 50 taxon alignment would have required far too much time to converge to the correct posterior distribution. Running Markov Katana for 1,000,000 generations on the 50 taxon tree would have required around 115 days and 18 hours, since each generation took around 10 seconds to complete (data not shown). The current neighbor joining algorithm is implemented in Perl 5. A faster implementation of the neighbor joining algorithm would speed up the generation time of Markov Katana. If a neighbor joining heuristic which scales better with the number of taxa were used such as NINJA [227], then the upper limit on the size of the tree that can be estimated would increase dramatically.

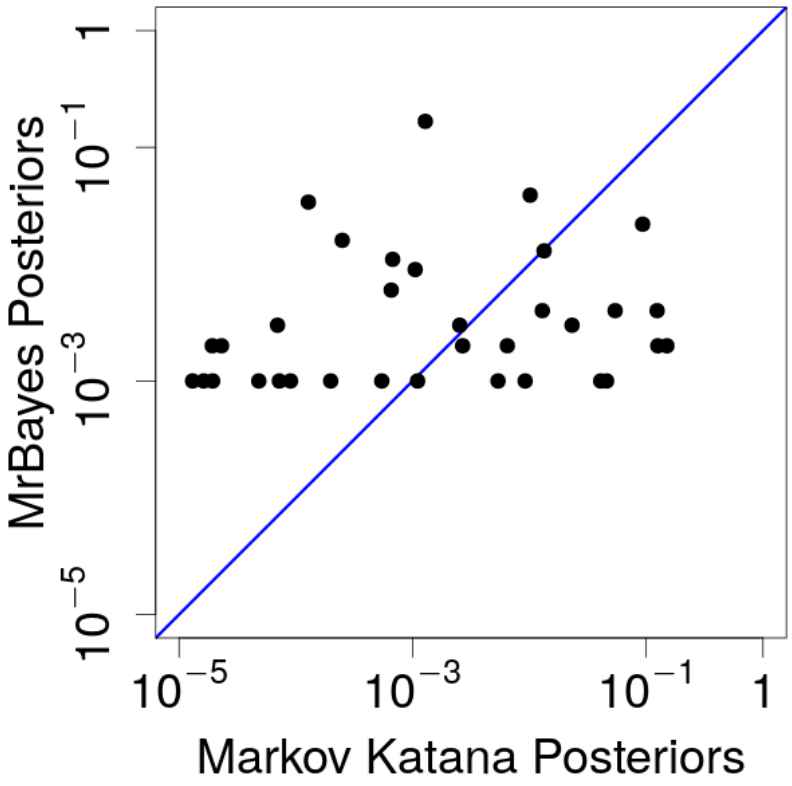


Figure VI.9: Markov Katana corrected tree topology posterior estimates are compared with MrBayes estimates for the same 20-taxon data set. The blue line indicates where x and y values are equal.

VI.7 Discussion

We have demonstrated here that the Markov Katana bootstrapping approach to phylogenetic tree searching can be a highly effective means for finding Bayesian posterior topologies and branches. It is able to take advantage of the speed of approximate distance-based methods to propose new trees, but retains the reliability of Bayesian methods. Many previous phylogenetic tree-search methods use the provided sequences for only the likelihood calculations, but Markov Katana introduces a new way to explore tree space informed by the sequences. Including the sequence data in the tree search improves the fraction of high likelihood trees proposed and allows efficient jump proposals between even distant topologies.

For the 10-taxon dataset, the NJ algorithm is extremely fast, and the overall speed of the Markov Katana computation was limited by the likelihood calculations. A table with some examples of runs with the number of generations, likelihood effective sample sizes, and times is presented in Table VI.3. As the number of taxa grows beyond 200, the NJ algorithm slows dramatically and dominates computation times (data not shown). This could be alleviated using fast heuristic NJ algorithms or external programs such as RapidNJ that are optimized for large alignments [228]. Our current implementation calculates the distance contribution of each site only once and so is not hindered by the complexity of the distance measure. We did not see a great difference in the proposal bias for the two distance measures we compared, but further exploration of the performance of alternative distances may in some cases be warranted.

This method has been applied to topology posteriors, but it could easily be applied to branch posteriors also. Calculating branch posteriors and branch length distributions using a similar method could expand the usefulness of the Markov Katana method for searching tree space. Then the posterior of a tree could be decomposed into the probability of each of its branches, assuming independence of the branch probabilities.

We used PAML for the likelihood calculations, but any program that computes

likelihoods could potentially be used. The simplicity and adjustability of the approach means that it could be easily incorporated into existing sequence analysis packages (e.g., MrBayes, PAUP*, HyPhy, and PAML [152, 229, 230, 172]). We used a Perl script to implement the Markov Katana algorithm and demonstrate the method as simply as possible, but we expect that Markov Katana can be easily integrated directly into existing programs, which would then undoubtedly be much faster. We did not see the benefit in constructing a new likelihood program from scratch, although we believe the methodology would interact well with our existing context-dependent Bayesian analysis program, PLEX [87].

VI.8 Declarations

VI.8.1 Ethics approval and consent to participate

Not applicable

VI.8.2 Consent for publication

Not applicable

VI.8.3 Availability of data and material

The datasets generated and/or analyzed during the current study are available in the Markov Katana repository, https://github.com/PollockLaboratory/MarkovKatana.

VI.8.4 Competing interests

The authors declare that they have no competing interests.

VI.8.5 Funding

We acknowledge the support of the National Institutes of Health (NIH; GM083127 and GM097251) to DDP.

VI.8.6 Authors' contributions

SP developed the scripts, performed the research, and wrote the manuscript. KF and DP contributed to the manuscript and early testing of the scripts. ZW and TC contributed to early versions of the manuscript. All authors read and approved the final manuscript.

VI.8.7 Acknowledgements

Thanks to Seena D. Shah, who contributed to early versions of coding on Markov Katana.

Tree ID	MrBayes Posterior	Markov Katana Posterior
79499	99.9%	96.6
3419	<1%	1.2%
77299	<1%	1.0%

Table VI.5: Posterior estimations from MrBayes and Markov Katana for the 10 taxon simulated tree and alignment.

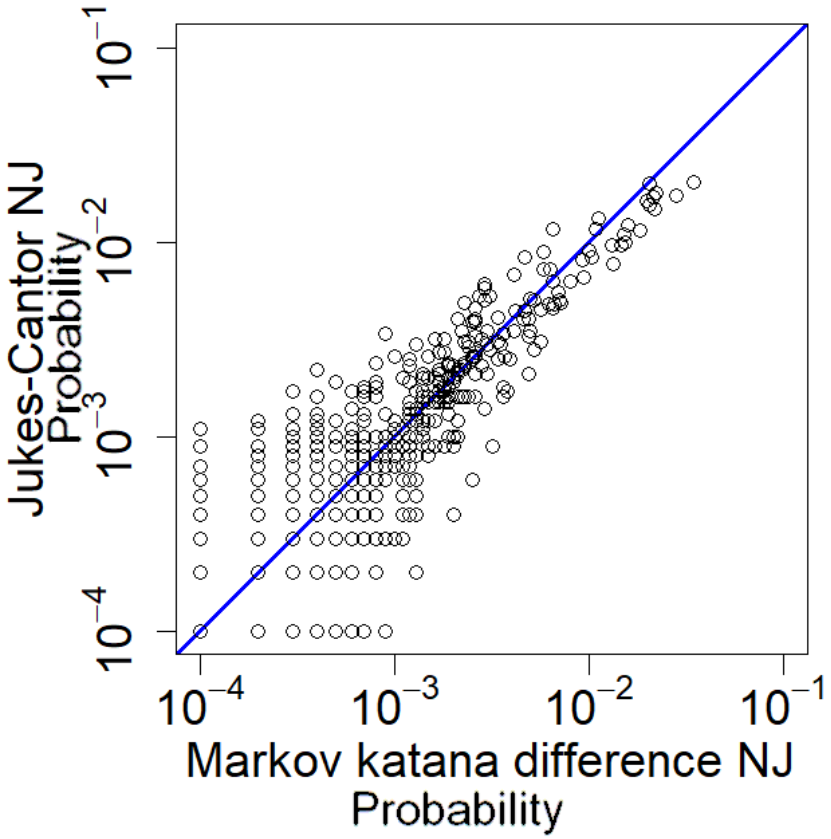
VI.9 Supplementary Material

VI.9.1 Simulation testing

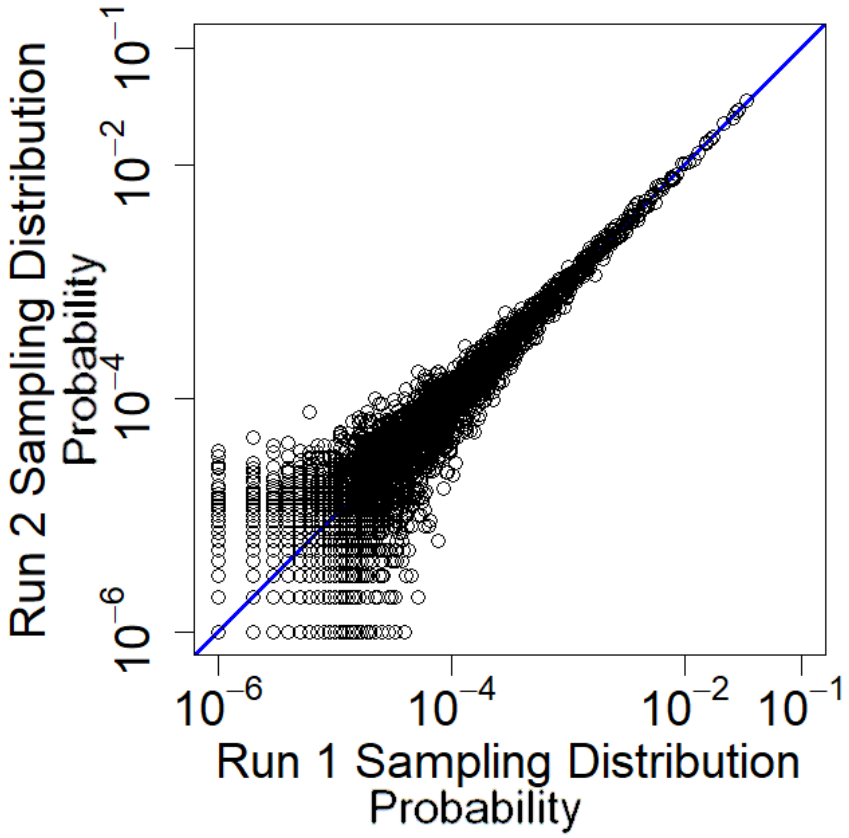
In order to determine if Markov Katana would detect the correct tree given a simulated data set, a tree was simulated using T-REX then sequences were generated for the tips using Seq-Gen and the WAG model [182, 231]. The tree had an average branch length of 0.1, 10 leaf taxa, and 500 sites were simulated for the alignment. The number of sites was selected to correspond to size of the Cytochrome C Oxidase subunit 1 (COI) mitochondrial alignment. MrBayes was used to estimate the trees from the generated sequences, fitting under the WAG model and was run for 100,000 generations. Markov Katana estimated the tree given the simulated multiple sequence alignments using a sample size of 1.0 and jump size of 0.1 for 100,000 generations and burnin of 10,000 generations, and calculating the tree likelihoods using the WAG model.

With this simulated data set, both MrBayes and Markov Katana provide very high posterior probabilities for the correct topology, with the ID 79499. A table of topology posteriors is provided in table VI.5, with the Markov Katana posteriors being uncorrected. In fact, since the sampling distribution favored the correct tree so often (due to the neighbor joining method employed), the corrected posterior gave less weight to the right tree than the uncorrected posterior. Both programs sampled other topologies at very low rates and converged to the correct answer. We may then conclude that when the answer is obvious and not obscured by the complexities of real mitochondrial evolution, Markov Katana can detect the correct tree.

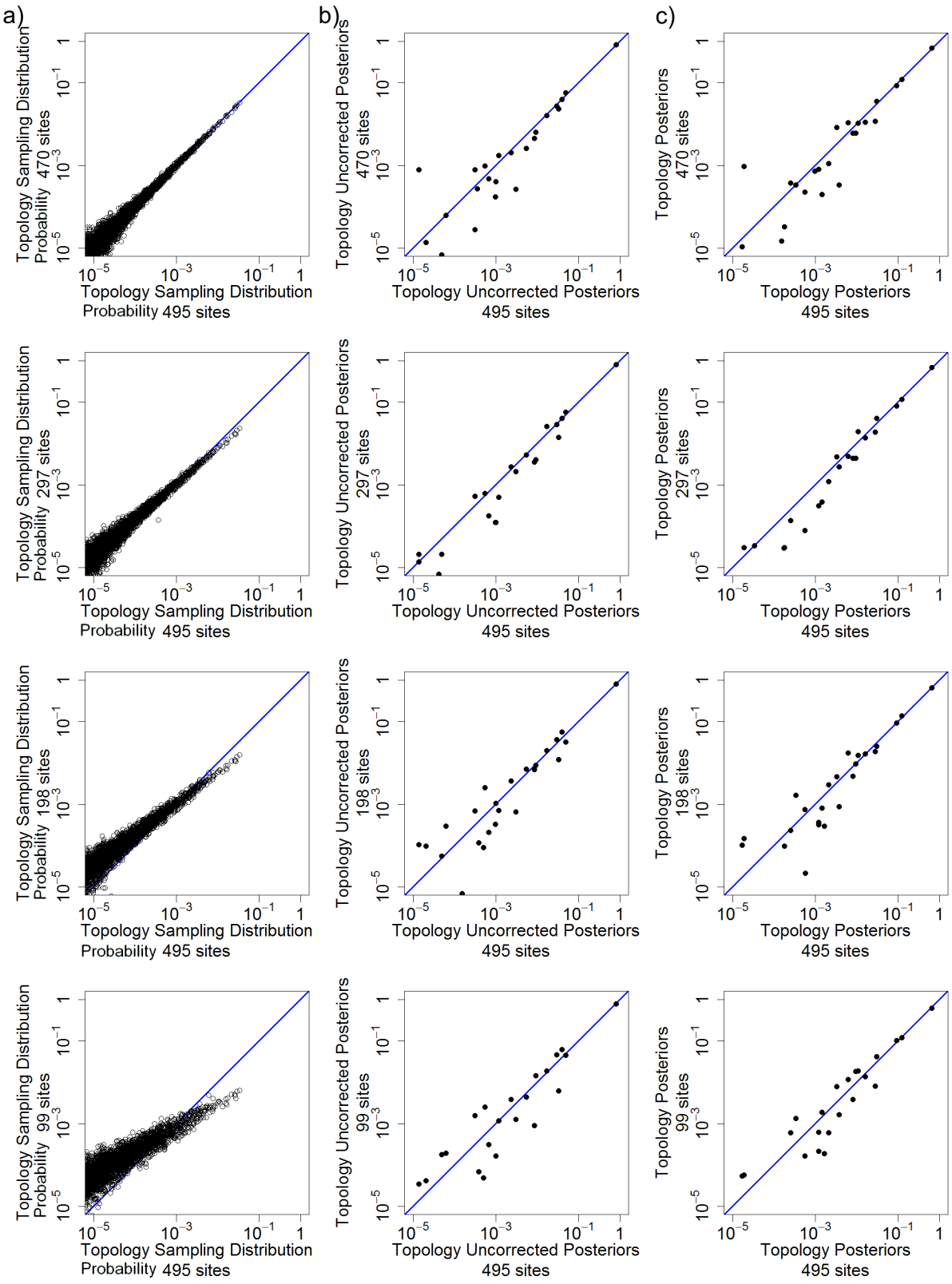
VI.9.2 Figures



Supplementary Figure VI.1: Comparing NJ implementations. Markov Katana bootstrapped topology probabilities were compared with those from a different neighbor-joining program, RapidNJ. 10,000 bootstraps were run.



Supplementary Figure VI.2: Variance of the sampling distribution among runs. The sampling distributions of two different runs with sample size 100% and jump size 10% and 1 million bootstraps are shown.



Supplementary Figure VI.3: Comparing sampling distribution and posteriors across sample size. The sampling distribution (column a), uncorrected posteriors (column b) and corrected posteriors (column c) for 470 sites, 297 sites, 198 sites, and 99 sites, all versus 495 sites.

Agapornis roseicollis
Anomalurus sp. GP-2005
Cervus nippon centralis
Chamaeleo calcaricarens
Chrysochloris asiatica
Ensatina eschscholtzii
Gulo gulo
Monodelphis domestica
Presbytis melalophos
Trachypithecus obscurus

Supplementary Table VI.1: Species in the Cytochrome C Oxidase Subunit 1 Alignment

```
begin mrbayes;
    set autoclose=yes nowarn=yes;
    execute COI.10taxa.20160106.aa.NoGap.nexus;
    prset aamodelpr = fixed(mtmam);
    mcmc ngen=100000 samplefreq=10 file=mbout.nex2;
    sumt relburnin=yes burninfrac=0.25;
end;
```

Supplementary Figure VI.4: The MrBayes block run for method validation. The file named "COI.10taxa.20160106.aa.NoGap.nexus" contains the multiple sequence alignment.

```
begin mrbayes;
    set autoclose=yes nowarn=yes;
    execute 20taxa.1000aa.NoGap.nexus;
    prset aamodelpr = fixed(mtmam);
    mcmc ngen=100000 samplefreq=10 file=20taxa_mbout.nex;
    sumt relburnin=yes burninfrac=0.25;
end;
```

Supplementary Figure VI.5: The MrBayes block run for method validation for the 20 taxon alignment. The file named "20taxa.1000aa.NoGap.nexus" contains the multiple sequence alignment.

VI.9.3 Tables

Agkistrodon piscivorus
Apus apus
Canis lupus lupus
Chamaeleo africanus
Echymipera rufescens australis
Episoriculus fumidus
Eulemur macaco macaco
Galago senegalensis
Gekko gecko
Hemiechinus auritus
Hemiphaga novaeseelandiae
Martes zibellina
Megaptera novaeangliae
Otis tarda
Phoenicopterus ruber roseus
Procolobus badius
Scolecomorphus vittatus
Todiramphus sanctus vagans
Tomistoma schlegelii
Vulpes vulpes

Supplementary Table VI.2: Species in the 20 taxon mitochondrial alignment

Achalinus meiguensis
Aegotheles cristatus
Agkistrodon piscivorus
Alectura lathami
Apus apus
Bambusicola thoracica
Batrachoseps attenuatus
Batrachuperus gorganensis
Bombina variegata
Canis lupus lupus
Chamaeleo africanus
Chamaeleo dilepis
Echymipera rufescens australis
Eothenomys chinensis
Episoriculus fumidus
Equus caballus
Eulemur fulvus mayottensis
Eulemur macaco macaco
Eumetopias jubatus
Galago senegalensis
Gekko gecko
Hemiechinus auritus
Hemiphaga novaeseelandiae
Hydromantes brunus
Hylobates lar

Supplementary Table VI.3: The first 25 species in the 50 taxon mitochondrial alignment

Hynobius amjiensis
Hynobius chinensis
Iguana iguana
Macrotis lagotis
Martes zibellina
Megaptera novaeangliae
Muntiacus reevesi micrurus
Mus musculus molossinus
Ornithorhynchus anatinus
Otis tarda
Pachyhynobius shangchengensis
Paleosuchus trigonatus
Phaethon rubricauda
Phoenicopterus ruber roseus
Platysternon megacephalum
Polychrus marmoratus
Procolobus badius
Rattus exulans
Scolecomorphus vittatus
Sminthopsis douglasi
Smithornis sharpei
Todiramphus sanctus vagans
Tomistoma schlegelii
Ursus thibetanus formosanus
Vulpes vulpes

Supplementary Table VI.4: The last 25 species in the 50 taxon mitochondrial alignment

CHAPTER VII

CONCLUSION*

The main goal of this thesis has been to improve the methods of studying evolution and propose a novel and general method for inferring amino acid propensities that change over sites and time. I have motivated the creation of new models using inferred molecular convergence. I have tested a method for calculating changing amino acid propensities. I have proposed a new way of testing models by integrating across likely trees using a Neighbor joining method named "Markov Katana". In this research, I bring together the ideas of phylogenetics, molecular convergence, and changing amino acid propensities in order to attempt to achieve an integrated understanding of protein evolution.

We have described the role of mechanisms and phenomenological descriptions as components of statistical empirical models, and described recent developments in mechanistic descriptions of the evolution of functional molecules, such as proteins. The role of fast thermodynamic evolutionary simulations is pivotal in discerning how proteins, as thermodynamic entities, should evolve, and what sorts of effects thermodynamics have on evolutionary outcomes. These thermodynamic models provide a potential explanation for patterns of epistasis, coevolution, average substitution rate differences over long periods of time, molecular convergence changes over time, and the evolutionary Stokes shift, which are fundamental problems for current statistical empirical models. We believe that a statistical mechanic-like treatment of protein sequence evolution points to a mechanistic explanation for many, if not all, of these phenomena, with the added benefit that it may greatly reduce the number of phenomenological parameters needed for future statistical empirical models of evolution.

*Portions of this chapter were previously published in *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts* 2017, *Encyclopedia of Evolutionary Biology* 2016, and Molecular Biology and Evolution 2015 volume 32 issue 6 and are included with the permission of the copyright holder.

VII.1 The Importance of Convergence

Current treatment of convergent events generally assumes that nonadaptive convergence at the molecular level is well predicted by simple time-averaged and site-averaged models. However, our analysis of real proteins and model-based simulations demonstrates that the rate of convergence changes over time and can be extremely high for recently diverged proteins. The convergence data presented here provide additional evidence that our understanding of how proteins evolve needs to be fundamentally revised. The patterns of convergent evolution observed may cause difficulties for phylogenetic reconstructions, but they can also provide important information about adaptation and adaptive bursts, as well as allowing us to investigate the underlying topology of the fitness landscape and the nature of the substitution process.

Convergence probability is closely related to the number of amino acids that are acceptable at a given site at a given time. If a small hydrophobic amino acid is required, the probability that two acceptable substitutions in different lineages will result in the same small hydrophobic amino acid can be quite high. Constraints at another site requiring large flexible amino acids will result in a similarly high probability of convergence. If the substitution model is inferred by averaging over different sites, or the same site at different times, including instances where only small hydrophobic, or large flexible, or aromatic, or charged amino acids are required, the result is a model with few constraints that allows a wide variety of different amino acids. These simple models will overestimate the number of acceptable amino acid substitutions and underestimate the probability of convergence.

The high rate of convergence and the strong dependence of the convergence rate on evolutionary distance strongly suggest the importance of variation in the substitution rate across sites and over time. The idea of fluctuating amino acid substitution rates over time is an important feature of evolutionary Stokes-shift theory [35]. According to this theory, the fitness of an amino acid for any site, and therefore the propensities for the

amino acid at that site, is dependent on how well suited it is to the environment formed by the amino acids at neighboring and interacting sites. As substitutions at neighboring sites alter the environment of a site, the amino acid propensities of that site will also be altered, resulting in fluctuating substitution rates at that site. Homologous but divergent proteins in other species will likely have fluctuated differently, meaning that the sets of acceptable amino acids at each position will diverge with evolutionary distance, causing a falloff in the convergence probability. In Stokes-shift theory, divergence in substitution models at a site is strongly coupled to substitutions at that site, so the convergence rate will also be significantly lower following a substitution.

The Stokes-Fisher (SF) model makes three additional predictions. First as the selection at different sites in the protein will be of different and fluctuating magnitude, there should be large differences and fluctuations in the convergence probability. Second we would expect more buried locations to be under more stringent constraints, resulting in a higher convergence probability than exposed locations. Third we expect the selective constraints at buried locations to diverge slowly because the residues around such locations are also buried and evolve slowly, resulting in a slower decline in the convergence probability with increasing evolutionary distance. All these predictions are matched by the observations of mitochondrial proteins.

Both heterogeneity of selection at different sites in the protein and fluctuations in selection over evolutionary time can cause models that neglect these effects to underestimate convergence rates. In particular, the CAT model [39], which includes spatial variation and excludes temporal variation, generates initially high C/D ratios that decline over evolutionary distance in a similar manner as the SF model. Similar drops in C/D ratios can also be seen in other highly parameterized site-specific models of spatial variation (data not shown). However, the effect of spatial versus temporal variation can be distinguished by considering the evolutionary distance dependence of C/D ratios from the same ancestral states. This ratio increases with evolutionary distance when a model is used

(CAT) that includes only spatial variation. Sites with fewer constraints are more likely to undergo changes, and therefore less likely to have the same ancestral states at longer divergence times. As a result the sites with the same ancestral states become increasingly the highly constrained sites with lower sequence entropy. As more constrained sites have higher C/D ratios, this means that C/D for these sites will increase with evolutionary distance. In contrast, when there are temporal changes in selection, diverging sequences will increasingly be under different selective constraints. This can result in a decreasing C/D ratio with increasing evolutionary distance. A fluctuating temporal component is not surprising, as no plausible biophysical model would allow site-specific constraints to remain fixed in the face of divergence in the rest of the protein, and there is other strong evidence for coevolution (or epistasis) among residue positions [42, 35, 36].

The effects of fluctuating and poorly estimated neutral convergence may have substantial effects on phylogenetic inference. Although truly neutral convergence is expected to be unbiased to any particular phylogenetic solution, it may well add considerable noise that would mask true phylogenetic signal. The distance dependence of the convergence probability may also interact in complex ways with the well known phylogenetic problem of long-branch attraction [143], and we expect that extensive analyses will be necessary to sort out such interactions. Furthermore, it is clear that our new understanding of fluctuating substitution processes suggests a multitude of new questions about how protein evolution operates and the role of convergence analysis in understanding protein evolution. Can we use convergence to better estimate instantaneous constraints? Can we understand the role of interactions between different amino acid substitutions at different distances in a protein structure, and how substitutions at those positions affect the probability of convergence? Can we use convergence estimates over different lengths of time to better understand the rates of fluctuation in constraints both with and without substitution at a target site? The inclusion of variation in the substitution process across sites and over time - details that standard models currently lack - should be included in future

evolutionary models to obtain more accurate descriptions of protein evolution.

VII.2 Detecting propensity shifts

In estimating how amino acid propensities change over time, I have shown how substitutions at adjacent sites correlate with large shifts in amino acid propensities. A novel method for determining the amino acid propensities of sites at specific times on the tree was proposed and tested via simulation. This propensity estimation method was then applied to the mitochondrial genome of 629 vertebrates. Substitutions were found to increase the average propensity shifts of adjacent sites at the same time as the substitution occurred. Many examples of large propensity shifts which correlated with adjacent substitutions were found and a few examples are shown.

The Acceptability model has been shown to be effective at estimating the propensities at sites even as they change over time, as demonstrated by the simulation results. When comparing to the model proposed by Usmanova et al., the Acceptability model is simpler and easier to fit to data [57]. The proposed model can be applied to real, full sized trees and sequences, as opposed to only quadruplets. It also does not make assumptions about what amino acids are forbidden from a site altogether, allowing any amino acid to be acceptable at any site, and is general enough to allow any amino acid substitution probability matrix to be used.

The Acceptability model is designed to be simple and yet allow the evolution process to vary over sites and time. Possible extensions to this model could relax some assumptions inherent to this model. First there is a single parameter which governs both the switch on and switch off rates ν, which could be split into two different rates: switch on rate ν_{on} and switch off rate ν_{off}. This would eliminate the assumption that the on and off rates are equal. It could also reduce or remove the need for the prior on the acceptable set size ρ, since the ratio of $\frac{\nu_{on}}{\nu_{off}}$ would determine the steady state equilibrium probability that any amino acid is on at any point in time.

It is also possible that the resident amino acid might increase the switch on rate for

other amino acids with similar physicochemical properties. The switch on rates would then be described by a matrix with each row corresponding to the resident amino acid and each cell is the switch on rate for each amino acid in the column. A model like this might be sufficient to explain the substitution rates seen in current matrix substitution rate models, however the explanation for these rates would be different. The resident amino acid would not randomly substitute to different amino acids at different rates, rather the resident amino acid would change the probability of similar amino acids becoming acceptable. An explanation like this brings us closer to a mechanistic model instead of a descriptive model.

One could also use more fitness levels than the binary fitness used here, if the data support having more levels. One complication is that the complexity of the Gibbs sampling increases linearly with the number of levels added. Adding more levels may add to the computational cost without learning more about the evolution process. The substitution process among high fitness amino acids could also be expanded above the simple one parameter model used here. If the nucleotide information is available, one could incorporate a model which allows for transition/transversion rate differences, such as Kimura's model [21]. One could use an entire matrix to describe the substitution process such as WAG or mtMam [142, 141].

Now that we can estimate the propensities at all the sites in an alignment and at the ancestral species, we can ask what other factors impact propensity shifts. There are many questions about how the structure of a protein influences its evolution that we might be able to answer. Do substitutions at one site cause large shifts in the propensities of sites which interact in the 3D structure with that site? We can probe whether there are sites which influence each other's propensities strongly but are in fact far away in the 3D structure of a protein. If we observe a substitution from a neutral amino acid to a charged amino acid, do the propensities for the oppositely charged amino acids increase at sites nearby in 3D space? Once we can calculate how the propensities have changed

over time, we can start answering questions like these.

VII.3 Markov Katana

We have demonstrated here that the Markov Katana bootstrapping approach to phylogenetic tree searching can be a highly effective means for finding Bayesian posterior topologies and branches. It is able to take advantage of the speed of approximate distance-based methods to propose new trees, but retains the reliability of Bayesian methods. Many previous phylogenetic tree-search methods use the provided sequences for only the likelihood calculations, but Markov Katana introduces a new way to explore tree space informed by the sequences. Including the sequence data in the tree search improves the fraction of high likelihood trees proposed and allows efficient jump proposals between even distant topologies.

For the 10-taxon dataset, the NJ algorithm is extremely fast, and the overall speed of the Markov Katana computation was limited by the likelihood calculations. As the number of taxa grows beyond 200, the NJ algorithm slows dramatically and dominates computation times (data not shown). This could be alleviated using fast heuristic NJ algorithms or external programs such as RapidNJ that are optimized for large alignments [228]. Our current implementation calculates the distance contribution of each site only once and so is not hindered by the complexity of the distance measure. We did not see a great difference in the proposal bias for the two distance measures we compared, but further exploration of the performance of alternative distances may in some cases be warranted.

This method has been applied to topology posteriors, but it could easily be applied to branch posteriors also. Calculating branch posteriors and branch length distributions using a similar method could expand the usefulness of the Markov Katana method for searching tree space. Then the posterior of a tree could be decomposed into the probability of each of its branches, assuming independence of the branch probabilities.

We used PAML for the likelihood calculations, but any program that computes

likelihoods could potentially be used. The simplicity and adjustability of the approach means that it could be easily incorporated into existing sequence analysis packages (e.g., MrBayes, PAUP*, HyPhy, and PAML [152, 229, 230, 172]). We used a Perl script to implement the Markov Katana algorithm and demonstrate the method as simply as possible, but we expect that Markov Katana can be easily integrated directly into existing programs, which would then undoubtedly be much faster. We did not see the benefit in constructing a new likelihood program from scratch, although we believe the methodology would interact well with our existing context-dependent Bayesian analysis program, PLEX [87].

VII.4 Future Work

Work has already begun to implement the Markov Katana method of tree sampling and complex time-independent models such as the Acceptability model, into a single program called SimPLEX, which is an updated version of PLEX [87]. This program is designed to efficiently fit complex site- and time-heterogeneous models to large amounts of phylogenetic data. It will be flexible enough to handle many different kinds of heterogeneities and dependencies within the models tested.

Most of the computational work involved in fitting the Acceptability model is in calculating the likelihood and sampling the hidden states. At present the likelihood is recalculated after the hidden states are Gibbs sampled. A possible improvement in calculating the likelihood would be simply update the likelihood during the Gibbs sampling and only recalculating the full likelihood occasionally. One could test if the likelihood updating method calculates the likelihood exactly correctly by comparing the updated likelihood with the fully calculated. If the updating method introduces slight errors, then there may be a discrepancy in the calculated likelihoods. Since each site is completely independent, they can be sampled in parallel. Using parallel threads for each Gibbs sampling of the hidden states for each site should improve the sampling speed substantially, making the speedup roughly equal to the number of sites in the multiple

sequence alignment. If the two improvements above are implemented, then each site would update the likelihood independently, possibly causing race conditions. This problem could be solved by either decomposing the overall likelihood into site-specific likelihoods or using a data structure that allows for concurrent access to the likelihood value.

One difficulty with a model like this is that it usually infers that only a single amino acid is acceptable at a time. The average instantaneous constraint ends up being higher than the two to four amino acids at a time that has been suggested previously [48]. It would be difficult to improve this metric, however, because the only information about the acceptability of an amino acid at any given point in time is the resident amino acid and nearby substitutions. One could construct a model which infers how likely a new amino acid would be to become acceptable given the current amino acid. This would result in a matrix of probabilities, each row of which might have the two to four amino acid constraint.

VII.4.1 Adding Structure to Evolution

After the Acceptability model, next steps could include building models with interactions among sites by directly modeling epistasis and allowing the resident amino acid at one site to influence the acceptable amino acids at another, global protein stability requirements, and perhaps even protein-ligand binding. Now that we can consistently infer when amino acids became acceptable along a tree, we can begin to investigate what physical forces cause amino acids to be high or low fitness at different positions in a protein. We already have strong evidence of amino acid frequency differences between the surface and the core of proteins, and we have solid biophysical reasons to expect those differences. Biophysics and evolutionary history have both shown that it is beneficial for some classes of proteins to have hydrophobic cores and charged or polar amino acids near the water exposed surfaces. The structure of the protein heavily influences the fitness of amino acids at different sites. How many other contributing physical factors could we discover that influence the fitnesses?

Not only could we hope to explain why certain sites have specific acceptabilities, but we might also learn why those acceptabilities change over time. Perhaps a substitution occurred at a site nearby in the protein structure shifting the local charged environment from positive to hydrophobic. This increases the fitness of hydrophobic amino acids and decreases the negatively charged amino acids resulting in a substitution to a hydrophobic residue. If the protein binds to another protein, a substitution in the binding interface of the other protein could precipitate a change in the propensities of protein's binding site. Now that we can map the substitution history along with the acceptability histories, we can begin to answer these questions.

One of the main callings in science is to construct models of our world that capture the major features of reality. We can choose those models to be statistical descriptions of phenomena, or we can seek to discover the underlying causes. It is time for us to move away from descriptive models of the evolutionary process and move toward models grounded in mechanism. Including the structure of an evolving protein is just one step in this transition. There are a huge number of forces that are essential to protein evolution that we could not model before. In the near future we will.

References

[1] D. W. Thompson. "The history of animals—Aristotle". In: *London: John Bell* (1907).

[2] René Descartes. *Descartes: The world and other writings*. Cambridge University Press, 1998.

[3] Pier Luigi Luisi. "About various definitions of life". In: *Origins of Life and Evolution of the Biosphere* 28.4 (1998), pp. 613–622.

[4] Carol E. Cleland and Christopher F. Chyba. "Defining 'life'". In: *Origins of Life and Evolution of the Biosphere* 32.4 (2002), pp. 387–393.

[5] Erwin Schrödinger. *What Is Life? the physical aspect of the living cell and mind*. Cambridge University Press, Cambridge, 1944.

[6] Theodosius Dobzhansky. "Nothing in Biology Makes Sense except in the Light of Evolution". In: *The American Biology Teacher* 35.3 (Mar. 1, 1973), pp. 125–129. ISSN: 0002-7685, 1938-4211. DOI: 10.2307/4444260. URL: http://abt.ucpress.edu/content/35/3/125 (visited on 10/31/2018).

[7] Charles Darwin. "On the origin of species by means of natural selection". In: *Murray, London* (1859).

[8] Carl Woese. "The universal ancestor". In: *Proceedings of the National Academy of Sciences* 95.12 (June 9, 1998), pp. 6854–6859. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.95.12.6854. URL: http://www.pnas.org/content/95/12/6854 (visited on 10/31/2018).

[9] W. Ford Doolittle. "Phylogenetic Classification and the Universal Tree". In: *Science* 284.5423 (June 25, 1999), pp. 2124–2128. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.284.5423.2124. URL: http://science.sciencemag.org/content/284/5423/2124 (visited on 10/31/2018).

[10] J. David Archibald. "Fossil evidence for a Late Cretaceous origin of "hoofed" mammals". In: *Science* 272.5265 (1996), pp. 1150–1153.

[11] J. David Archibald, Alexander O. Averianov, and Eric G. Ekdale. "Late Cretaceous relatives of rabbits, rodents, and other extant eutherian mammals". In: *Nature* 414.6859 (Nov. 2001), pp. 62–65. ISSN: 1476-4687. DOI: 10.1038/35102048. URL: https://www.nature.com/articles/35102048 (visited on 11/01/2018).

[12] Qiang Ji et al. "The earliest known eutherian mammal". In: *Nature* 416.6883 (Apr. 2002), pp. 816–822. ISSN: 1476-4687. DOI: 10.1038/416816a. URL: https://www.nature.com/articles/416816a (visited on 11/01/2018).

[13] Zhe-Xi Luo and John R. Wible. "A Late Jurassic digging mammal and early mammalian diversification". In: *Science* 308.5718 (2005), pp. 103–107.

[14] Vincent M. Sarich and Allan C. Wilson. "Rates of albumin evolution in primates". In: *Proceedings of the National Academy of Sciences* 58.1 (1967), pp. 142–148.

[15] S. V. Muse and B. S. Weir. "Testing for equality of evolutionary rates." In: *Genetics* 132.1 (Sept. 1, 1992), pp. 269–276. ISSN: 0016-6731, 1943-2631. URL: http://www.genetics.org/content/132/1/269 (visited on 10/31/2018).

[16] Jerzy Neyman. "Molecular studies of evolution: a source of novel statistical problems". In: *Statistical decision theory and related topics*. Elsevier, 1971, pp. 1–27.

[17] Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002. URL: http://books.google.com/books?hl=en&lr=&id=fT1Iu-h6E-oC&oi=fnd&pg=PR7&dq=burnham+anderson+model+selection+&ots=teto33_Hn0&sig=KJIpVYGDIFpMGDTVIBPlHIPU1Kc (visited on 02/04/2015).

[18] Anthony WF Edwards. "The reconstruction of evolution: estimation by maximum likelihood". In: *Genetical Society programme* (1965).

[19] Luigi L. Cavalli-Sforza and Anthony WF Edwards. "Phylogenetic analysis: models and estimation procedures". In: *Evolution* 21.3 (1967), pp. 550–570.

[20] Thomas H. Jukes and Charles R. Cantor. "Evolution of protein molecules". In: *Mammalian protein metabolism* 3 (1969), pp. 21–132. URL: http://books.google.com/books?hl=en&lr=&id=FDHLBAAAQBAJ&oi=fnd&pg=PA21&dq=Jukes+T,

+Cantor+C.+1969.+%22Evolution+of+protein+molecules.&ots=bkdrVGT_kA&sig=UmReLQj5b_oKpbess6UN3anuJ0c (visited on 04/06/2015).

[21] Motoo Kimura. "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". In: *Journal of molecular evolution* 16.2 (1980), pp. 111–120. URL: http://link.springer.com/article/10.1007/BF01731581 (visited on 06/18/2014).

[22] N. Goldman and Ziheng Yang. "A codon-based model of nucleotide substitution for protein-coding DNA sequences." en. In: *Molecular Biology and Evolution* 11.5 (Sept. 1994), pp. 725–736. ISSN: 0737-4038, 1537-1719. URL: http://mbe.oxfordjournals.org/content/11/5/725 (visited on 06/23/2014).

[23] Nicolas Lartillot and Hervé Philippe. "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process". In: *Molecular biology and evolution* 21.6 (2004), pp. 1095–1109. URL: http://mbe.oxfordjournals.org/content/21/6/1095.short (visited on 06/18/2014).

[24] R. P. Jorré and R. N. Curnow. "The evolution of the genetic code". In: *Biochimie* 57.10 (Jan. 1976), pp. 1147–1154. ISSN: 0300-9084. DOI: 10.1016/S0300-9084(76)80576-7. URL: http://www.sciencedirect.com/science/article/pii/S0300908476805767 (visited on 04/10/2015).

[25] R. P. Jorré and R. N. Curnow. "A model for the evolution of the proteins: Cytochrome c : mammals, reptiles, insects". In: *Biochimie* 57.10 (Jan. 1976), pp. 1141–1146. ISSN: 0300-9084. DOI: 10.1016/S0300-9084(76)80575-5. URL: http://www.sciencedirect.com/science/article/pii/S0300908476805755 (visited on 04/10/2015).

[26] Georgii A Bazykin. "Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins". In: *Biology letters* 11.10 (2015), p. 20150315.

[27] Jeffrey L. Thorne, Nick Goldman, and David T. Jones. "Combining protein evolution and secondary structure." In: *Molecular Biology and Evolution* 13.5 (1996), pp. 666–673. URL: http://mbe.oxfordjournals.org/content/13/5/666.short (visited on 06/18/2014).

[28] Nick Goldman, Jeffrey L. Thorne, and David T. Jones. "Assessing the impact of secondary structure and solvent accessibility on protein evolution". In: *Genetics* 149.1 (1998), pp. 445–458. URL: http://www.genetics.org/content/149/1/445.short (visited on 02/04/2015).

[29] Douglas M. Robinson et al. "Protein Evolution with Dependence Among Codons Due to Tertiary Structure". In: *Molecular Biology and Evolution* 20.10 (2003), pp. 1692–1704. DOI: 10.1093/molbev/msg184. URL: http://mbe.oxfordjournals.org/content/20/10/1692.short.

[30] Nicolas Rodrigue et al. "Site interdependence attributed to tertiary structure in amino acid sequence evolution". In: *Gene* 347.2 (2005), pp. 207–217. URL: http://www.sciencedirect.com/science/article/pii/S0378111904007310 (visited on 06/18/2014).

[31] Nicolas Rodrigue, Hervé Philippe, and Nicolas Lartillot. "Assessing site-interdependent phylogenetic models of sequence evolution". In: *Molecular biology and evolution* 23.9 (2006), pp. 1762–1775. URL: http://mbe.oxfordjournals.org/content/23/9/1762.short (visited on 06/18/2014).

[32] Claudia L. Kleinman et al. "Statistical potentials for improved structurally constrained evolutionary models". In: *Molecular biology and evolution* 27.7 (2010), pp. 1546–1560. URL: http://mbe.oxfordjournals.org/content/27/7/1546.short (visited on 04/10/2015).

[33] Pietro Lio and Nick Goldman. "Using protein structural information in evolutionary inference: transmembrane proteins." In: *Molecular biology and evolution* 16.12 (1999), pp. 1696–1710. URL: http://mbe.oxfordjournals.org/content/16/12/1696.short (visited on 02/04/2015).

[34] Buyong Ma et al. "Protein–protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces". en. In: *Proceedings of the National Academy of Sciences* 100.10 (May 2003), pp. 5772–5777. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1030237100. URL: http://www.pnas.org/content/100/10/5772 (visited on 04/10/2015).

[35] David D. Pollock, Grant Thiltgen, and Richard A. Goldstein. "Amino acid coevolution induces an evolutionary Stokes shift". In: *Proceedings of the National Academy of Sciences* 109.21 (2012), E1352–E1359. URL: http://www.pnas.org/content/109/21/E1352.short (visited on 02/03/2015).

[36] D. D. Pollock and R. A. Goldstein. "Strong evidence for protein epistasis, weak evidence against it". In: *Proc Natl Acad Sci U S A* 111.15 (2014), E1450. ISSN: 1091-6490. DOI: 10.1073/pnas.1401112111. URL: http://www.ncbi.nlm.nih.gov/pubmed/24706894.

[37] Jeffrey M. Koshi and Richard A. Goldstein. "Models of natural mutations including site heterogeneity". In: (1998). URL: http://deepblue.lib.umich.edu/handle/2027.42/38528 (visited on 06/18/2014).

[38] Matthew W. Dimmic, David P. Mindell, and Richard A. Goldstein. "Modeling evolution at the protein level using an adjustable amino acid fitness model". In: *Ann Arbor* 1001 (2000), pp. 48109–1055. URL: ftp://vm-lux.embl.de/pub/pub/users/lercher/Ka/Dimmic_Goldstein2000PacSympBiocomput.pdf (visited on 04/10/2015).

[39] le S Quang, O. Gascuel, and N. Lartillot. "Empirical profile mixture models for phylogenetic reconstruction". In: *Bioinformatics* 24.20 (2008), pp. 2317–23. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn445. URL: http://www.ncbi.nlm.nih.gov/pubmed/18718941.

[40] Nicolas Rodrigue, Hervé Philippe, and Nicolas Lartillot. "Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles". In: *Proceedings of the National Academy of Sciences* 107.10 (2010), pp. 4629–4634. URL: http://www.pnas.org/content/107/10/4629.short (visited on 06/18/2014).

[41] Ulrike Göbel et al. "Correlated mutations and residue contacts in proteins". In: *Proteins: Structure, Function, and Bioinformatics* 18.4 (1994), pp. 309–317. URL: http://onlinelibrary.wiley.com/doi/10.1002/prot.340180402/full (visited on 06/18/2014).

[42] David D. Pollock, William R. Taylor, and Nick Goldman. "Coevolving protein residues: maximum likelihood identification and relationship to structure". In: *Journal of Molecular Biology* 287.1 (Mar. 1999), pp. 187–198. ISSN: 0022-2836. DOI: 10.1006/jmbi.1998.2601. URL: http://www.sciencedirect.com/science/article/pii/S0022283698926018 (visited on 06/18/2014).

[43] P. Lopez, D. Casane, and H. Philippe. "Heterotachy, an Important Process of Protein Evolution". en. In: *Molecular Biology and Evolution* 19.1 (Jan. 2002),

pp. 1–7. ISSN: 0737-4038, 1537-1719. URL: http://mbe.oxfordjournals.org/content/19/1/1 (visited on 04/06/2015).

[44] Yan Zhou et al. "Evaluation of the models handling heterotachy in phylogenetic inference". en. In: *BMC Evolutionary Biology* 7.1 (Nov. 2007), p. 206. ISSN: 1471-2148. DOI: 10.1186/1471-2148-7-206. URL: http://www.biomedcentral.com/1471-2148/7/206/abstract/ (visited on 03/17/2015).

[45] Bryan Kolaczkowski and Joseph W. Thornton. "A Mixed Branch Length Model of Heterotachy Improves Phylogenetic Accuracy". In: *Molecular Biology and Evolution* 25.6 (June 2008), pp. 1054–1066. ISSN: 0737-4038. DOI: 10.1093/molbev/msn042. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299401/ (visited on 03/17/2015).

[46] Nicolas Galtier. "Maximum-likelihood phylogenetic analysis under a covarion-like model". In: *Molecular Biology and Evolution* 18.5 (2001), pp. 866–873. URL: http://mbe.oxfordjournals.org/content/18/5/866.short (visited on 06/18/2014).

[47] David Penny et al. "Mathematical elegance with biochemical realism: the covarion model of molecular evolution". In: *Journal of Molecular Evolution* 53.6 (2001), pp. 711–723. URL: http://link.springer.com/article/10.1007/s002390010258 (visited on 06/18/2014).

[48] Richard A. Goldstein et al. "Nonadaptive Amino Acid Convergence Rates Decrease over Time". en. In: *Molecular Biology and Evolution* 32.6 (June 2015), pp. 1373–1381. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv041. URL: http://mbe.oxfordjournals.org/content/32/6/1373 (visited on 07/08/2015).

[49] A. U. Tamuri et al. "Identifying changes in selective constraints: host shifts in influenza". In: *PLoS Comput Biol* 5.11 (2009), e1000564. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000564. URL: http://www.ncbi.nlm.nih.gov/pubmed/19911053.

[50] Béatrice Roure and Hervé Philippe. "Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference". In: *BMC Evolutionary Biology* 11.1 (Jan. 14, 2011), p. 17. ISSN: 1471-2148. DOI: 10.1186/1471-2148-11-17. URL: https://doi.org/10.1186/1471-2148-11-17 (visited on 11/13/2018).

[51] Nicolas Galtier and Manolo Gouy. "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis." In: *Molecular biology and evolution* 15.7 (1998), pp. 871–879. URL: http://mbe.oxfordjournals.org/content/15/7/871.short (visited on 04/10/2015).

[52] Samuel Blanquart and Nicolas Lartillot. "A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution". en. In: *Molecular Biology and Evolution* 23.11 (Nov. 2006), pp. 2058–2071. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msl091. URL: http://mbe.oxfordjournals.org/content/23/11/2058 (visited on 04/02/2015).

[53] Julien Dutheil and Bastien Boussau. "Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs". en. In: *BMC Evolutionary Biology* 8.1 (Sept. 2008), p. 255. ISSN: 1471-2148. DOI: 10.1186/1471-2148-8-255. URL: http://www.biomedcentral.com/1471-2148/8/255/abstract (visited on 03/10/2015).

[54] Julien Y. Dutheil et al. "Efficient selection of branch-specific models of sequence evolution". In: *Molecular biology and evolution* (2012), mss059. URL: http://mbe.oxfordjournals.org/content/early/2012/02/02/molbev.mss059.short (visited on 04/10/2015).

[55] Herve Philippe et al. "Heterotachy and functional shift in protein evolution". In: *IUBMB life* 55.4-5 (2003), pp. 257–265. URL: http://onlinelibrary.wiley.com/doi/10.1080/1521654031000123330/full (visited on 02/02/2016).

[56] Hervé Philippe et al. "Heterotachy and long-branch attraction in phylogenetics". In: *BMC evolutionary biology* 5.1 (2005), p. 50. URL: http://www.biomedcentral.com/1471-2148/5/50/ (visited on 02/02/2016).

[57] Dinara R. Usmanova et al. "A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution". en. In: *Molecular Biology and Evolution* 32.2 (Feb. 2015), pp. 542–554. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msu318. URL: http://mbe.oxfordjournals.org/content/32/2/542 (visited on 03/01/2016).

[58] Anthony WF Edwards and Cavalli LL Sforza. "The reconstruction of evolution". In: *Heredity* 18 (1963).

[59] David Posada and Thomas R. Buckley. "Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests". In: *Systematic Biology* 53.5 (Oct. 1, 2004), pp. 793–808. ISSN: 1063-5157. DOI: 10.1080/10635150490522304. URL: https://academic.oup.com/sysbio/article/53/5/793/2842928 (visited on 10/31/2018).

[60] Paul D. Williams et al. "Assessing the accuracy of ancestral protein reconstruction methods". In: *PLoS computational biology* 2.6 (2006), e69. URL: http://dx.plos.org/10.1371/journal.pcbi.0020069 (visited on 06/19/2014).

[61] Hirotugu Akaike. "A new look at the statistical model identification". In: *Automatic Control, IEEE Transactions on* 19.6 (1974), pp. 716–723. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1100705 (visited on 02/02/2015).

[62] Gideon Schwarz et al. "Estimating the dimension of a model". In: *The annals of statistics* 6.2 (1978), pp. 461–464. URL: http://projecteuclid.org/euclid.aos/1176344136 (visited on 02/02/2015).

[63] Steven N. Goodman. "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy". In: *Annals of Internal Medicine* 130.12 (June 15, 1999), p. 995. ISSN: 0003-4819. DOI: 10.7326/0003-4819-130-12-199906150-00008. URL: http://annals.org/article.aspx?doi=10.7326/0003-4819-130-12-199906150-00008 (visited on 10/31/2018).

[64] Steven N. Goodman. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor". In: *Annals of Internal Medicine* 130.12 (June 15, 1999), p. 1005. ISSN: 0003-4819. DOI: 10.7326/0003-4819-130-12-199906150-00019. URL: http://annals.org/article.aspx?doi=10.7326/0003-4819-130-12-199906150-00019 (visited on 10/31/2018).

[65] AP Jason de Koning, Wanjun Gu, and David D. Pollock. "Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories". In: *Molecular biology and evolution* 27.2 (2010), pp. 249–265. URL: http://mbe.oxfordjournals.org/content/27/2/249.short (visited on 02/04/2015).

[66] Barbara E. Engelhardt et al. "Protein molecular function prediction by Bayesian phylogenomics". In: *PLoS computational biology* 1.5 (2005), e45.

[67] Charles B. Fenster, Laura F. Galloway, and Lin Chao. "Epistasis and its consequences for the evolution of natural populations". In: *Trends in Ecology & Evolution* 12.7 (1997), pp. 282–286.

[68] Motoo Kimura. "Evolutionary rate at the molecular level". In: *Nature* 217.5129 (1968), pp. 624–626.

[69] John H. Gillespie. *The causes of molecular evolution*. Vol. 2. Oxford University Press on Demand, 1994.

[70] Tomoko Ohta. "Slightly deleterious mutant substitutions in evolution". In: *Nature* 246.5428 (1973), p. 96.

[71] Frieder Mayer and A. Brunner. "Non-neutral evolution of the major histocompatibility complex class II gene DRB1 in the sac-winged bat Saccopteryx bilineata". In: *Heredity* 99.3 (2007), p. 257.

[72] Matthew L. Holding, James E. Biardi, and H. Lisle Gibbs. "Coevolution of venom function and venom resistance in a rattlesnake predator and its squirrel prey". In: *Proc. R. Soc. B* 283.1829 (2016), p. 20152841.

[73] Armita Nourmohammad, Jakub Otwinowski, and Joshua B. Plotkin. "Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections". In: *PLoS genetics* 12.7 (2016), e1006171.

[74] Stéphane Aris-Brosou and Ziheng Yang. "Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny". In: *Systematic Biology* 51.5 (2002), pp. 703–714.

[75] T. A. Castoe et al. "Adaptive evolution and functional redesign of core metabolic proteins in snakes". In: *PLoS One* 3.5 (2008), e2201. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0002201. URL: http://www.ncbi.nlm.nih.gov/pubmed/18493604.

[76] Miguel Arenas, Agustin Sánchez-Cobos, and Ugo Bastolla. "Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability". ENG. In: *Molecular Biology and Evolution* (Apr. 2015). ISSN: 1537-1719. DOI: 10.1093/molbev/msv085.

[77] Ugo Bastolla, Yves Dehouck, and Julian Echave. "What evolution tells us about protein physics, and protein physics tells us about evolution". In: *Current opinion in structural biology* 42 (2017), pp. 59–66.

[78] Grant Thiltgen and Richard A. Goldstein. "Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency". In: *PLoS ONE* 7.10 (Oct. 2012), e46084. DOI: 10.1371/journal.pone.0046084. URL: http://dx.doi.org/10.1371/journal.pone.0046084 (visited on 07/01/2015).

[79] Nick Goldman and Ziheng Yang. *Introduction. Statistical and computational challenges in molecular phylogenetics and evolution*. The Royal Society, 2008.

[80] Jeffrey L. Thorne. "Models of protein sequence evolution and their applications". In: *Current opinion in genetics & development* 10.6 (2000), pp. 602–605. URL: http://www.sciencedirect.com/science/article/pii/S0959437X00001428 (visited on 02/02/2016).

[81] Jeffrey M. Koshi and Richard A. Goldstein. "Mutation matrices and physical-chemical properties: Correlations and implications". In: *Proteins: Structure, Function, and Bioinformatics* 27.3 (1997), pp. 336–344.

[82] Jeffrey M. Koshi, David P. Mindell, and Richard A. Goldstein. "Beyond mutation matrices: physical-chemistry based evolutionary models". In: *Proceedings of the second annual international conference on Computational molecular biology*. ACM, 1998, pp. 140–145. URL: http://dl.acm.org/citation.cfm?id=279107 (visited on 02/02/2016).

[83] Richard A. Goldstein and David D. Pollock. "The tangled bank of amino acids". In: *Protein Science* 25.7 (2016), pp. 1354–1362.

[84] Steven E. Brenner, Cyrus Chothia, and Tim JP Hubbard. "Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships". In: *Proceedings of the National Academy of Sciences* 95.11 (1998), pp. 6073–6078.

[85] Steven Henikoff and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks". In: *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919.

[86] W. John Wilbur. "On the PAM matrix model of protein evolution." In: *Molecular biology and evolution* 2.5 (1985), pp. 434–447.

[87] A. P. Jason de Koning et al. "Phylogenetics, likelihood, evolution and complexity". en. In: *Bioinformatics* 28.22 (Nov. 2012), pp. 2989–2990. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts555. URL: http://bioinformatics.oxfordjournals.org/content/28/22/2989 (visited on 06/23/2014).

[88] Aaron L. Halpern and William J. Bruno. "Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies." In: *Molecular biology and evolution* 15.7 (1998), pp. 910–917. URL: http://mbe.oxfordjournals.org/content/15/7/910.short (visited on 06/18/2014).

[89] Jeffrey M. Koshi and Richard A. Goldstein. "Context-dependent optimal substitution matrices". In: *Protein Engineering* 8.7 (1995), pp. 641–645. URL: http://peds.oxfordjournals.org/content/8/7/641.short (visited on 06/18/2014).

[90] Jeffrey M. Koshi, David P. Mindell, and Richard A. Goldstein. "Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes." In: *Molecular biology and evolution* 16.2 (1999), pp. 173–179. URL: http://mbe.oxfordjournals.org/content/16/2/173.short (visited on 02/02/2016).

[91] Asif U. Tamuri, Mario dos Reis, and Richard A. Goldstein. "Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models". In: *Genetics* 190.3 (2012), pp. 1101–1115. URL: http://www.genetics.org/content/190/3/1101.short (visited on 02/02/2016).

[92] Wolfgang Stephan. "The rate of compensatory evolution". In: *Genetics* 144.1 (1996), pp. 419–426.

[93] Enoch Baldwin et al. "Thermodynamic and structural compensation in" size-switch" core repacking variants of bacteriophage T4 lysozyme." In: (1996).

[94] Michael S. Breen et al. "Epistasis as the primary factor in molecular evolution". en. In: *Nature* 490.7421 (Oct. 2012), pp. 535–538. ISSN: 0028-0836. DOI: 10.1038/nature11510. URL: http://www.nature.com/nature/journal/v490/n7421/full/nature11510.html (visited on 05/05/2015).

[95] DD Pollock and ST Pollard. "Parallel and Convergent Molecular Evolution". In: (2016).

[96] O. Ashenberg, L. I. Gong, and J. D. Bloom. "Mutational effects on stability are largely conserved during protein evolution". In: *Proc Natl Acad Sci U S A* 110.52 (2013), pp. 21071–6. ISSN: 1091-6490. DOI: 10.1073/pnas.1314781111. URL: http://www.ncbi.nlm.nih.gov/pubmed/24324165.

[97] Bryan Lunt et al. "Inference of direct residue contacts in two-component signaling". In: *Methods in enzymology*. Vol. 471. Elsevier, 2010, pp. 17–41.

[98] Faruck Morcos et al. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". en. In: *Proceedings of the National Academy of Sciences* 108.49 (Dec. 2011), E1293–E1301. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1111471108. URL: http://www.pnas.org/content/108/49/E1293 (visited on 06/30/2015).

[99] Martin Weigt et al. "Identification of direct residue contacts in protein–protein interaction by message passing". In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

[100] David M. McCandlish, Premal Shah, and Joshua B. Plotkin. "Epistasis and the dynamics of reversion in molecular evolution". In: *Genetics* (2016), genetics–116.

[101] Georgii A. Bazykin et al. "Extensive parallelism in protein evolution". In: *Biology direct* 2.1 (2007), p. 20.

[102] Paul H. Harvey and Mark D. Pagel. *The comparative method in evolutionary biology*. Vol. 239. Oxford university press Oxford, 1991.

[103] Ernst Mayr. *Animal species and evolution*. OCLC: 551391. Cambridge: Belknap Press of Harvard University Press, 1963.

[104] Antonis Rokas and Sean B. Carroll. "Frequent and Widespread Parallel Evolution of Protein Sequences". en. In: *Molecular Biology and Evolution* 25.9 (Sept. 2008), pp. 1943–1953. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msn143. URL: http://mbe.oxfordjournals.org/content/25/9/1943 (visited on 06/23/2014).

[105] Joe Parker et al. "Genome-wide signatures of convergent evolution in echolocating mammals". In: *Nature* 502.7470 (2013), pp. 228–231. URL: http://www.nature.com/nature/journal/v502/n7470/abs/nature12511.html (visited on 06/23/2014).

[106] Todd A. Castoe et al. "Evidence for an ancient adaptive episode of convergent molecular evolution". en. In: *Proceedings of the National Academy of Sciences* (Apr. 2009), pnas.0900233106. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0900233106. URL: http://www.pnas.org/content/early/2009/04/28/0900233106 (visited on 06/18/2014).

[107] Gregg W.C. Thomas and Matthew W. Hahn. "Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals". In: *Molecular biology and evolution* (2015), pp. 1–49.

[108] Zhengting Zou and Jianzhi Zhang. "No Genome-Wide Protein Sequence Convergence for Echolocation". en. In: *Molecular Biology and Evolution* 32.5 (May 2015), pp. 1237–1241. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv014. URL: http://mbe.oxfordjournals.org/content/32/5/1237 (visited on 06/25/2015).

[109] David T. Jones, William R. Taylor, and Janet M. Thornton. "The rapid generation of mutation data matrices from protein sequences". In: *Computer applications in the biosciences: CABIOS* 8.3 (1992), pp. 275–282. URL: http://bioinformatics.oxfordjournals.org/content/8/3/275.short (visited on 06/18/2014).

[110] Simon Whelan, Pietro Liò, and Nick Goldman. "Molecular phylogenetics: state-of-the-art methods for looking into the past". In: *Trends in Genetics* 17.5 (May 2001), pp. 262–272. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(01)02272-7. URL: http://www.sciencedirect.com/science/article/pii/S0168952501022727 (visited on 05/11/2017).

[111] Zhengting Zou and Jianzhi Zhang. "Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations?" en. In: *Molecular Biology and Evolution* (Apr. 2015), msv091. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv091. URL: http://mbe.oxfordjournals.org/content/early/2015/04/09/molbev.msv091 (visited on 04/15/2015).

[112] Zhengting Zou and Jianzhi Zhang. "Gene tree discordance does not explain away the temporal decline of convergence in mammalian protein sequence evolution". In: *Molecular biology and evolution* 34.7 (2017), pp. 1682–1688.

[113] Fábio K. Mendes, Yoonsoo Hahn, and Matthew W. Hahn. "Gene tree discordance can generate patterns of diminishing convergence over time". In: *Molecular biology and evolution* 33.12 (2016), pp. 3299–3307.

[114] Premal Shah, David M. McCandlish, and Joshua B. Plotkin. "Contingency and entrenchment in protein evolution under purifying selection". In: *Proceedings of the National Academy of Sciences* (2015), p. 201412933.

[115] Michael B. Doud, Orr Ashenberg, and Jesse D. Bloom. "Site-specific amino acid preferences are mostly conserved in two closely related protein homologs". In: *Molecular biology and evolution* 32.11 (2015), pp. 2944–2960.

[116] Richard A. Goldstein and David D. Pollock. "Sequence entropy of folding and the absolute rate of amino acid substitutions". In: *Nature Ecology & Evolution* 1.12 (Dec. 2017), pp. 1923–1930. ISSN: 2397-334X. DOI: 10.1038/s41559-017-0338-9. URL: https://www.nature.com/articles/s41559-017-0338-9 (visited on 07/12/2018).

[117] Richard A. Goldstein. "The evolution and evolutionary consequences of marginal thermostability in proteins". In: *Proteins: Structure, Function, and Bioinformatics* 79.5 (2011), pp. 1396–1407. URL: http://onlinelibrary.wiley.com/doi/10.1002/prot.22964/full (visited on 06/18/2014).

[118] Darin M. Taverna and Richard A. Goldstein. "Why are proteins marginally stable?" In: *Proteins: Structure, Function, and Bioinformatics* 46.1 (2002), pp. 105–109.

[119] Andrew D. Foote et al. "Convergent evolution of the genomes of marine mammals". en. In: *Nature Genetics* 47.3 (Mar. 2015), pp. 272–275. ISSN: 1061-4036. DOI: 10.1038/ng.3198. URL: http://www.nature.com/ng/journal/v47/n3/abs/ng.3198.html (visited on 06/25/2015).

[120] Yang Liu et al. "Convergent sequence evolution between echolocating bats and dolphins". In: *Current Biology* 20.2 (2010), R53–R54. URL: http://www.sciencedirect.com/science/article/pii/S0960982209020739 (visited on 02/05/2015).

[121] Guojie Zhang et al. "Comparative genomics reveals insights into avian genome evolution and adaptation". en. In: *Science* 346.6215 (Dec. 2014), pp. 1311–1320. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1251385. URL: http://www.sciencemag.org/content/346/6215/1311 (visited on 06/25/2015).

[122] Caro-Beth Stewart, James W. Schilling, and Allan C. Wilson. "Adaptive evolution in the stomach lysozymes of foregut fermenters". In: (1987). URL: http://www.nature.com/nature/journal/v330/n6146/abs/330401a0.html (visited on 02/02/2015).

[123] John E. Schienman et al. "Duplication and Divergence of 2 Distinct Pancreatic Ribonuclease Genes in Leaf-Eating African and Asian Colobine Monkeys". en. In: *Molecular Biology and Evolution* 23.8 (Aug. 2006), pp. 1465–1479. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msl025. URL: http://mbe.oxfordjournals.org/content/23/8/1465 (visited on 10/26/2015).

[124] Jianzhi Zhang. "Parallel Functional Changes in the Digestive RNases of Ruminants and Colobines by Divergent Amino Acid Substitutions". In: *Molecular Biology and Evolution* 20.8 (Aug. 1, 2003), pp. 1310–1317. ISSN: 0737-4038. DOI: 10.1093/molbev/msg143. URL: https://academic.oup.com/mbe/article/20/8/1310/1081568 (visited on 11/01/2018).

[125] Zhen Liu et al. "Parallel Sites Implicate Functional Convergence of the Hearing Gene Prestin among Echolocating Mammals". en. In: *Molecular Biology and Evolution* 31.9 (Sept. 2014), pp. 2415–2424. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msu194. URL: http://mbe.oxfordjournals.org/content/31/9/2415 (visited on 02/26/2015).

[126] K T J Davies et al. "Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence". In: *Heredity* 108.5 (May 2012), pp. 480–489. ISSN: 0018-067X. DOI: 10.1038/hdy.2011.119. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3330687/ (visited on 10/23/2015).

[127] H. A. Wichman et al. "Experimental evolution recapitulates natural evolution". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 355.1403 (Nov. 29, 2000), pp. 1677–1684. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2000.0731. URL: http://rstb.royalsocietypublishing.org/content/355/1403/1677 (visited on 11/01/2018).

[128] Olivier Tenaillon et al. "The Molecular Diversity of Adaptive Convergence". en. In: *Science* 335.6067 (Jan. 2012), pp. 457–461. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1212986. URL: http://www.sciencemag.org/content/335/6067/457 (visited on 10/26/2015).

[129] Jungeui Hong and David Gresham. "Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments". In: *PLOS Genetics* 10.1 (Jan. 9, 2014), e1004041. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004041. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004041 (visited on 11/01/2018).

[130] K. D. Yokoyama and D. D. Pollock. "SP transcription factor paralogs and DNA-binding sites coevolve and adaptively converge in mammals and birds". In: *Genome Biol Evol* 4.11 (2012), pp. 1102–17. ISSN: 1759-6653. DOI: 10.1093/gbe/evs085. URL: http://www.ncbi.nlm.nih.gov/pubmed/23019068.

[131] Jeff Arendt and David Reznick. "Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation?" In: *Trends in Ecology & Evolution* 23.1 (2008), pp. 26–32. URL: http://www.sciencedirect.com/science/article/pii/S016953470700287X (visited on 02/25/2015).

[132] Pascal-Antoine Christin, Daniel M. Weinreich, and Guillaume Besnard. "Causes and evolutionary significance of genetic convergence". In: *Trends in Genetics* 26.9 (Sept. 2010), pp. 400–405. ISSN: 0168-9525. DOI: 10.1016/j.tig.2010.06.005. URL: http://www.sciencedirect.com/science/article/pii/S0168952510001289 (visited on 12/04/2014).

[133] J. Zhang and S. Kumar. "Detection of convergent and parallel evolution at the amino acid sequence level." en. In: *Molecular Biology and Evolution* 14.5 (May 1997), pp. 527–536. ISSN: 0737-4038, 1537-1719. URL: http://mbe.oxfordjournals.org/content/14/5/527 (visited on 06/23/2014).

[134] Erica Bree Rosenblum, Christine E. Parent, and Erin E. Brandt. "The Molecular Basis of Phenotypic Convergence". In: *Annual Review of Ecology, Evolution, and Systematics* 45.1 (Nov. 23, 2014), pp. 203–226. ISSN: 1543-592X. DOI: 10.1146/annurev-ecolsys-120213-091851. URL: https://www.annualreviews.org/doi/10.1146/annurev-ecolsys-120213-091851 (visited on 11/01/2018).

[135] Nicolas Lartillot, Henner Brinkmann, and Hervé Philippe. "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model". en. In: *BMC Evolutionary Biology* 7.Suppl 1 (Feb. 2007), S4. ISSN: 1471-2148. DOI: 10.1186/1471-2148-7-S1-S4. URL: http://www.biomedcentral.com/1471-2148/7/S1/S4/ (visited on 02/01/2015).

[136] Neeraja M. Krishnan et al. "Ancestral Sequence Reconstruction in Primate Mitochondrial DNA: Compositional Bias and Effect on Functional Inference". en. In: *Molecular Biology and Evolution* 21.10 (Oct. 2004), pp. 1871–1883. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msh198. URL: http://mbe.oxfordjournals.org/content/21/10/1871 (visited on 10/23/2015).

[137] Yong-Yi Shen et al. "Parallel evolution of auditory genes for echolocation in bats and toothed whales". In: *PLoS genetics* 8.6 (2012), e1002788. URL: http://dx.plos.org/10.1371/journal.pgen.1002788 (visited on 02/05/2015).

[138] Asif U. Tamuri, Mario dos Reis, and Richard A. Goldstein. "Using Site-wise Mutation-Selection Models to Estimate the Distribution of Selection Coefficients from Phylogenetic Data". In: *Genetics* (2011), genetics–111. URL: http://www.genetics.org/content/early/2011/12/29/genetics.111.136432.short (visited on 06/18/2014).

[139] Jeffrey M. Koshi and Richard A. Goldstein. "Analyzing site heterogeneity during protein evolution". In: *Biocomputing 2001*. World Scientific, 2000, pp. 191–202. URL: https://books.google.com/books?hl=en&lr=&id=nZ3VCgAAQBAJ&oi=fnd&pg=PA191&dq=RA+goldstein+phylogenetics&ots=lwUCjSFa_3&sig=qZvfeul_EIUh9XdS9aaLal1l0zc (visited on 02/02/2016).

[140] B. P. Blackburne, A. J. Hay, and R. A. Goldstein. "Changing selective pressure during antigenic changes in human influenza H3". In: *PLoS Pathog* 4.5 (2008), e1000058. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1000058. URL: http://www.ncbi.nlm.nih.gov/pubmed/18451985.

[141] Ziheng Yang, R. Nielsen, and M. Hasegawa. "Models of amino acid substitution and applications to mitochondrial protein evolution". In: *Mol Biol Evol* 15.12 (1998), pp. 1600–11. ISSN: 0737-4038. URL: http://www.ncbi.nlm.nih.gov/pubmed/9866196.

[142] Simon Whelan and Nick Goldman. "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach". In: *Molecular biology and evolution* 18.5 (2001), pp. 691–699. URL: http://mbe.oxfordjournals.org/content/18/5/691.short (visited on 06/18/2014).

[143] Joseph Felsenstein. *Inferring phylogenies*. Vol. 2. Sunderland, Massachusetts: Sinauer Associates, 2004. URL: http://www.sinauer.com/media/wysiwyg/tocs/InferringPhylogenies.pdf (visited on 07/10/2014).

[144] Mark A. Larkin et al. "Clustal W and Clustal X version 2.0". In: *Bioinformatics* 23.21 (2007), pp. 2947–2948. URL: http://bioinformatics.oxfordjournals.org/content/23/21/2947.short (visited on 02/03/2015).

[145] Ari Löytynoja and Nick Goldman. "An algorithm for progressive multiple alignment of sequences with insertions". In: *Proceedings of the National academy of sciences of the United States of America* 102.30 (2005), pp. 10557–10562. URL: http://www.pnas.org/content/102/30/10557.short (visited on 02/03/2015).

[146] N. M. Krishnan, S. Z. Raina, and D. D. Pollock. "Analysis of among-site variation in substitution patterns". In: *Biol Proced Online* 6 (2004), pp. 180–188. ISSN: 1480-9222. DOI: 10.1251/bpo88. URL: http://www.ncbi.nlm.nih.gov/pubmed/15361931.

[147] Neeraja M. Krishnan et al. "Detecting Gradients of Asymmetry in Site-Specific Substitutions in Mitochondrial Genomes". In: *DNA and Cell Biology* 23.10 (Oct. 2004), pp. 707–714. ISSN: 1044-5498. DOI: 10.1089/dna.2004.23.707. URL: http://online.liebertpub.com/doi/abs/10.1089/dna.2004.23.707 (visited on 06/24/2015).

[148] H. Akaike, B. N. Petrov, and F. Csaki. "Second International Symposiumon Information Theory". In: (1973).

[149] Hirotugu Akaike. "Information measures and model selection". In: *Int Stat Inst* 44 (1983), pp. 277–291.

[150] J. A. A. Nylander. "MrModeltest v2". In: *Program distributed by the author* (2004).

[151] John P. Huelsenbeck, Fredrik Ronquist, et al. "MRBAYES: Bayesian inference of phylogenetic trees". In: *Bioinformatics* 17.8 (2001), pp. 754–755. URL: http://www.naturhistoriska.se/download/18.42129f1312d951207af800041595/Huelsenbeck_et_al_BioInfo_2001.pdf (visited on 02/02/2015).

[152] Fredrik Ronquist and John P. Huelsenbeck. "MrBayes 3: Bayesian phylogenetic inference under mixed models". In: *Bioinformatics* 19.12 (2003), pp. 1572–1574. URL: http://bioinformatics.oxfordjournals.org/content/19/12/1572.short (visited on 02/02/2015).

[153] Sanzo Miyazawa and Robert L. Jernigan. "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation". In: *Macromolecules* 18.3 (1985), pp. 534–552. URL: http://pubs.acs.org/doi/abs/10.1021/ma00145a039 (visited on 06/18/2014).

[154] Ylva Lindqvist et al. "Three-dimensional structure of a mammalian purple acid phosphatase at 2.2 \AA resolution with a μ-(hydr) oxo bridged di-iron center". In: *Journal of molecular biology* 291.1 (1999), pp. 135–147. URL: http://www.sciencedirect.com/science/article/pii/S0022283699929625 (visited on 06/18/2014).

[155] Motoo Kimura. "Some problems of stochastic processes in genetics". In: *The Annals of Mathematical Statistics* (1957), pp. 882–901. URL: http://www.jstor.org/stable/2237051 (visited on 06/18/2014).

[156] Motoo Kimura. "On the probability of fixation of mutant genes in a population". In: *Genetics* 47.6 (1962), p. 713. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210364/ (visited on 06/18/2014).

[157] James F. Crow, Motoo Kimura, et al. "An introduction to population genetics theory." In: *An introduction to population genetics theory.* (1970). URL: http://www.cabdirect.org/abstracts/19710105376.html (visited on 06/18/2014).

[158] Miguel Arenas. "Trends in substitution models of molecular evolution". In: *Frontiers in genetics* 6 (2015), p. 319.

[159] David D. Pollock et al. "Mechanistic Models of Protein Evolution". en. In: *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*. Springer, Cham, 2017, pp. 277–296. DOI: 10.1007/978-3-319-61569-1_15. URL: https://link.springer.com/chapter/10.1007/978-3-319-61569-1_15 (visited on 01/22/2018).

[160] Margaret O. Dayhoff and Robert M. Schwartz. "A model of evolutionary change in proteins". In: *In Atlas of protein sequence and structure*. Citeseer, 1978. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315 (visited on 06/18/2014).

[161] Si Quang Le, Cuong Cao Dang, and Olivier Gascuel. "Modeling protein evolution with several amino acid replacement matrices depending on site rates". eng. In:

Molecular Biology and Evolution 29.10 (Oct. 2012), pp. 2921–2936. ISSN: 1537-1719. DOI: 10.1093/molbev/mss112.

[162] Sarah K. Hilton and Jesse D. Bloom. "Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence". In: *Virus Evolution* 4.2 (July 1, 2018). DOI: 10.1093/ve/vey033. URL: https://academic.oup.com/ve/article/4/2/vey033/5163287 (visited on 11/13/2018).

[163] "Detecting amino acid preference shifts with codon-level mutation-selection mixture modelsmar and Rodrigue, Nicolas". In: *BMC evolutionary biology* 19.1 (2019), p. 62.

[164] David M McCandlish, Premal Shah, and Joshua B Plotkin. "Epistasis and the dynamics of reversion in molecular evolution". In: *Genetics* 203.3 (2016), pp. 1335–1351.

[165] Hugh K. Haddox et al. "Mapping mutational effects along the evolutionary landscape of HIV envelope". In: *Elife* 7 (2018), e34420.

[166] Sergey Gavrilets. "Evolution and speciation on holey adaptive landscapes". In: *Trends in Ecology & Evolution* 12.8 (Aug. 1, 1997), pp. 307–312. ISSN: 0169-5347. DOI: 10.1016/S0169-5347(97)01098-7. URL: http://www.sciencedirect.com/science/article/pii/S0169534797010987 (visited on 11/07/2018).

[167] Sergey Gavrilets and Janko Gravner. "Percolation on the fitness hypercube and the evolution of reproductive isolation". In: *Journal of theoretical biology* 184.1 (1997), pp. 51–64.

[168] Erik Van Nimwegen, James P. Crutchfield, and Martijn Huynen. "Neutral evolution of mutational robustness". In: *Proceedings of the National Academy of Sciences* 96.17 (1999), pp. 9716–9720.

[169] Janko Gravner, Damien Pitman, and Sergey Gavrilets. "Percolation on fitness landscapes: Effects of correlation, phenotype, and incompatibilities". In: *Journal of Theoretical Biology* 248.4 (Oct. 21, 2007), pp. 627–645. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2007.07.009. URL: http://www.sciencedirect.com/science/article/pii/S0022519307003335 (visited on 11/13/2018).

[170] Fabian Sievers et al. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”. In: *Molecular systems biology* 7.1 (2011). URL: http://onlinelibrary.wiley.com/doi/10.1038/msb.2011.75/full (visited on 02/01/2016).

[171] Rodrigo Gouveia-Oliveira, Peter W Sackett, and Anders G Pedersen. “MaxAlign: maximizing usable data in an alignment”. In: *BMC bioinformatics* 8.1 (2007), p. 312.

[172] Ziheng Yang. “PAML 4: phylogenetic analysis by maximum likelihood”. In: *Mol Biol Evol* 24.8 (2007). Yang, Ziheng Journal Article Research Support, Non-U.S. Gov’t United States Mol Biol Evol. 2007 Aug;24(8):1586-91. Epub 2007 May 4., pp. 1586–91. ISSN: 0737-4038 (Print)0737-4038. DOI: 10.1093/molbev/msm088. URL: http://dx.doi.org/10.1093/molbev/msm088.

[173] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092. URL: http://scitation.aip.org/content/aip/journal/jcp/21/6/10.1063/1.1699114 (visited on 02/05/2015).

[174] W. K. Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. ISSN: 0006-3444. URL: https://academic.oup.com/biomet/article-abstract/57/1/97/2721936/Monte-Carlo-sampling-methods-using-Markov-chains (visited on 05/11/2017).

[175] Stuart Geman and Donald Geman. “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[176] Walter M. Fitch and Etan Markowitz. “An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution”. In: *Biochemical genetics* 4.5 (1970), pp. 579–593. URL: http://link.springer.com/article/10.1007/BF00486096 (visited on 04/06/2015).

[177] Chris Tuffley and Mike Steel. “Modeling the covarion hypothesis of nucleotide substitution”. In: *Mathematical Biosciences* 147.1 (Jan. 1, 1998), pp. 63–91. ISSN: 0025-5564. DOI: 10.1016/S0025-5564(97)00081-3. URL: http://www.sciencedirect.com/science/article/pii/S0025556497000813 (visited on 11/15/2018).

[178] John P. Huelsenbeck. "Testing a Covariotide Model of DNA Substitution". In: *Molecular Biology and Evolution* 19.5 (May 1, 2002), pp. 698–707. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a004128. URL: https://academic.oup.com/mbe/article/19/5/698/1067820 (visited on 11/15/2018).

[179] Toshimichi Ikemura. "Codon usage and tRNA content in unicellular and multicellular organisms." In: *Molecular biology and evolution* 2.1 (1985), pp. 13–34.

[180] Hiroshi Akashi. "Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy." In: *Genetics* 136.3 (1994), pp. 927–935.

[181] Kai Zeng and Brian Charlesworth. "Estimating Selection Intensity on Synonymous Codon Usage in a Nonequilibrium Population". In: *Genetics* 183.2 (2009), pp. 651–662. ISSN: 0016-6731. DOI: 10.1534/genetics.109.101782. eprint: https://www.genetics.org/content/183/2/651.full.pdf. URL: https://www.genetics.org/content/183/2/651.

[182] Alix Boc, Alpha B. Diallo, and Vladimir Makarenkov. "T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks". In: *Nucleic Acids Research* 40.W1 (2012), W573–W579.

[183] Nicolas Rodrigue and Nicolas Lartillot. "Detecting Adaptation in Protein-Coding Genes Using a Bayesian Site-Heterogeneous Mutation-Selection Codon Substitution Model". In: *Molecular Biology and Evolution* 34.1 (Oct. 2016), pp. 204–214. ISSN: 0737-4038. DOI: 10.1093/molbev/msw220. eprint: http://oup.prod.sis.lan/mbe/article-pdf/34/1/204/24246271/msw220.pdf. URL: https://doi.org/10.1093/molbev/msw220.

[184] Freek J. Vonk et al. "The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system". en. In: *Proceedings of the National Academy of Sciences* 110.51 (Dec. 2013), pp. 20651–20656. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1314702110. URL: http://www.pnas.org/content/110/51/20651 (visited on 06/25/2015).

[185] Kenji Fukushima et al. "Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory". In: *Nature Ecology & Evolution* 1 (2017), p. 0059. URL: https://www.nature.com/articles/s41559-016-0059?dom=icopyright&src=syn (visited on 06/19/2017).

[186] Zhengyuan O. Wang and David D. Pollock. "Context dependence and coevolution among amino acid residues in proteins". In: *Methods in enzymology* 395 (2005), pp. 779–790. URL: http://www.sciencedirect.com/science/article/pii/S0076687905950404 (visited on 06/18/2014).

[187] Jeremiah D. Hackett et al. "Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates". In: *Molecular Biology and Evolution* 24.8 (2007), pp. 1702–1713. URL: http://mbe.oxfordjournals.org/content/24/8/1702.short (visited on 05/09/2017).

[188] Adrian Reyes-Prieto and Debashish Bhattacharya. "Phylogeny of nuclear-encoded plastid-targeted proteins supports an early divergence of glaucophytes within Plantae". In: *Molecular biology and evolution* 24.11 (2007), pp. 2358–2361. URL: http://mbe.oxfordjournals.org/content/24/11/2358.short (visited on 05/09/2017).

[189] Roger A. Craig and Li Liao. "Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices". In: *BMC Bioinformatics* 8 (2007), p. 6. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-6. URL: http://dx.doi.org/10.1186/1471-2105-8-6 (visited on 05/09/2017).

[190] Hans Bodlaender, Mike Fellows, and Tandy Warnow. "Two strikes against perfect phylogeny". In: *Automata, Languages and Programming* (1992), pp. 273–283. URL: http://www.springerlink.com/index/Y1L403152W1N6R22.pdf (visited on 05/09/2017).

[191] Luciano Brocchieri. "Phylogenetic Inferences from Molecular Sequences: Review and Critique". en. In: *Theoretical Population Biology* 59.1 (Feb. 2001), pp. 27–40. ISSN: 00405809. DOI: 10.1006/tpbi.2000.1485. URL: http://linkinghub.elsevier.com/retrieve/pii/S0040580900914850 (visited on 05/09/2017).

[192] N. Saitou and M. Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular Biology and Evolution* 4.4 (July 1987), pp. 406–425. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040454. URL: https://academic.oup.com/mbe/article/4/4/406/1029664/The-neighbor-joining-method-a-new-method-for (visited on 05/09/2017).

[193] Mark A. Hershkovitz and Detlef D. Leipe. "Phylogenetic Analysis". en. In: *Bioinformatics*. Ed. by Andreas D. Baxevanis and B. F. Francis Ouellette. John Wiley & Sons, Inc., 1998, pp. 189–230. ISBN: 978-0-470-11060-7. DOI: 10.1002/

9780470110607.ch9. URL: http://onlinelibrary.wiley.com/doi/10.1002/9780470110607.ch9/summary (visited on 05/09/2017).

[194] Kei Takahashi and Masatoshi Nei. "Efficiencies of Fast Algorithms of Phylogenetic Inference Under the Criteria of Maximum Parsimony, Minimum Evolution, and Maximum Likelihood When a Large Number of Sequences Are Used". In: *Molecular Biology and Evolution* 17.8 (Aug. 2000), pp. 1251–1258. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026408. URL: https://academic.oup.com/mbe/article/17/8/1251/992808/Efficiencies-of-Fast-Algorithms-of-Phylogenetic (visited on 05/11/2017).

[195] Joseph Felsenstein. "Evolutionary trees from DNA sequences: A maximum likelihood approach". en. In: *Journal of Molecular Evolution* 17.6 (Nov. 1981), pp. 368–376. ISSN: 0022-2844, 1432-1432. DOI: 10.1007/BF01734359. URL: https://link.springer.com/article/10.1007/BF01734359 (visited on 05/09/2017).

[196] John P. Huelsenbeck and Keith A. Crandall. "Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood". In: *Annual Review of Ecology and Systematics* 28.1 (1997), pp. 437–466. DOI: 10.1146/annurev.ecolsys.28.1.437. URL: http://dx.doi.org/10.1146/annurev.ecolsys.28.1.437 (visited on 05/09/2017).

[197] Jack Sullivan and Paul Joyce. "Model Selection in Phylogenetics". In: *Annual Review of Ecology, Evolution, and Systematics* 36.1 (2005), pp. 445–466. DOI: 10.1146/annurev.ecolsys.36.102003.152633. URL: http://dx.doi.org/10.1146/annurev.ecolsys.36.102003.152633 (visited on 05/10/2017).

[198] Maria Anisimova and Olivier Gascuel. "Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative". In: *Systematic Biology* 55.4 (Aug. 2006), pp. 539–552. ISSN: 1063-5157. DOI: 10.1080/10635150600755453. URL: https://academic.oup.com/sysbio/article/55/4/539/1675125/Approximate-Likelihood-Ratio-Test-for-Branches-A (visited on 05/10/2017).

[199] Elchanan Mossel and Eric Vigoda. "Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees". en. In: *Science* 309.5744 (Sept. 2005), pp. 2207–2209. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1115493. URL: http://science.sciencemag.org/content/309/5744/2207 (visited on 05/10/2017).

[200] Joseph Felsenstein. "Distance Methods for Inferring Phylogenies: A Justification". In: *Evolution* 38.1 (1984), pp. 16–24. ISSN: 0014-3820. DOI: 10.2307/2408542. URL: http://www.jstor.org/stable/2408542 (visited on 05/18/2017).

[201] John P. Huelsenbeck and Mark Kirkpatrick. "Do Phylogenetic Methods Produce Trees with Biased Shapes?" In: *Evolution* 50.4 (1996), pp. 1418–1424. ISSN: 0014-3820. DOI: 10.2307/2410879. URL: http://www.jstor.org/stable/2410879 (visited on 05/10/2017).

[202] Xuhua Xia. "Topological bias in distance-based phylogenetic methods: problems with over-and underestimated genetic distances". In: *Evolutionary Bioinformatics* 2 (2006). URL: http://search.proquest.com/openview/186314449a175c3222c8c80061e94530/1?pq-origsite=gscholar&cbl=1026404 (visited on 05/10/2017).

[203] John P. Huelsenbeck. "Performance of Phylogenetic Methods in Simulation". In: *Systematic Biology* 44.1 (Mar. 1995), pp. 17–48. ISSN: 1063-5157. DOI: 10.1093/sysbio/44.1.17. URL: https://academic.oup.com/sysbio/article-abstract/44/1/17/1667053/Performance-of-Phylogenetic-Methods-in-Simulation (visited on 05/10/2017).

[204] David D Pollock. "Increased Accuracy in Analytical Molecular Distance Estimation". In: *Theoretical Population Biology* 54.1 (Aug. 1998), pp. 78–90. ISSN: 0040-5809. DOI: 10.1006/tpbi.1998.1362. URL: http://www.sciencedirect.com/science/article/pii/S0040580998913624 (visited on 05/10/2017).

[205] David D. Pollock and William J. Bruno. "Assessing an Unknown Evolutionary Process: Effect of Increasing Site-Specific Knowledge Through Taxon Addition". en. In: *Molecular Biology and Evolution* 17.12 (Dec. 2000), pp. 1854–1858. ISSN: 0737-4038, 1537-1719. URL: http://mbe.oxfordjournals.org/content/17/12/1854 (visited on 06/18/2014).

[206] Michael J. Sanderson and Amy C. Driskell. "The challenge of constructing large phylogenetic trees". In: *Trends in Plant Science* 8.8 (Aug. 2003), pp. 374–379. ISSN: 1360-1385. DOI: 10.1016/S1360-1385(03)00165-1. URL: http://www.sciencedirect.com/science/article/pii/S1360138503001651 (visited on 05/10/2017).

[207] M. D. Hendy and David Penny. "Branch and bound algorithms to determine minimal evolutionary trees". In: *Mathematical Biosciences* 59.2 (June 1982), pp. 277–290. ISSN: 0025-5564. DOI: 10.1016/0025-5564(82)90027-X. URL: http://www.sciencedirect.com/science/article/pii/002555648290027X (visited on 07/10/2017).

[208] Marco Salemi, Philippe Lemey, and Anne-Mieke Vandamme. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. en. Cambridge University Press, Mar. 2009. ISBN: 978-0-521-87710-7.

[209] Kevin C. Nixon. "The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis". en. In: *Cladistics* 15.4 (Dec. 1999), pp. 407–414. ISSN: 1096-0031. DOI: 10.1111/j.1096-0031.1999.tb00277.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1096-0031.1999.tb00277.x/abstract (visited on 05/11/2017).

[210] R. A. Vos. "Accelerated Likelihood Surface Exploration: The Likelihood Ratchet". In: *Systematic Biology* 52.3 (June 2003), pp. 368–373. ISSN: 1063-5157. DOI: 10.1080/10635150390196993. URL: https://academic.oup.com/sysbio/article/52/3/368/1665193/Accelerated-Likelihood-Surface-Exploration-The (visited on 05/11/2017).

[211] Simon Whelan. "New Approaches to Phylogenetic Tree Search and Their Application to Large Numbers of Protein Alignments". In: *Systematic Biology* 56.5 (Oct. 2007), pp. 727–740. ISSN: 1063-5157. DOI: 10.1080/10635150701611134. URL: https://academic.oup.com/sysbio/article/56/5/727/1694763/New-Approaches-to-Phylogenetic-Tree-Search-and (visited on 05/11/2017).

[212] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. "Optimization by Simulated Annealing". In: *Science* 220.4598 (1983), pp. 671–680. ISSN: 0036-8075. URL: http://www.jstor.org/stable/1690046 (visited on 05/11/2017).

[213] Charles J. Geyer. "Markov Chain Monte Carlo Maximum Likelihood". en_US. In: Interface Foundation of North America, 1991. URL: http://conservancy.umn.edu/handle/11299/58440 (visited on 05/11/2017).

[214] Hideo Matsuda. "Construction of Phylogenetic Trees from Amino Acid Sequences using a Genetic Algorithm". In: *Genome Informatics* 6 (1995), pp. 19–28. DOI: 10.11234/gi1990.6.19.

[215] Alexei J. Drummond et al. "Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data". In: *Genetics* 161.3 (July 1, 2002), pp. 1307–1320. ISSN: 0016-6731, 1943-2631. URL: http://www.genetics.org/content/161/3/1307 (visited on 07/02/2018).

[216] Ian J. Wilson and David J. Balding. "Genealogical inference from microsatellite data". In: *Genetics* 150.1 (1998), pp. 499–510.

[217] S. Hohna, M. Defoin-Platel, and A. J. Drummond. "Clock-constrained tree proposal operators in Bayesian phylogenetic inference". In: *2008 8th IEEE International Conference on BioInformatics and BioEngineering*. 2008 8th IEEE International Conference on BioInformatics and BioEngineering. Oct. 2008, pp. 1–7. DOI: 10.1109/BIBE.2008.4696663.

[218] Clemens Lakner et al. "Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics". In: *Systematic Biology* 57.1 (Feb. 2008), pp. 86–103. ISSN: 1063-5157. DOI: 10.1080/10635150801886156. URL: https://academic.oup.com/sysbio/article/57/1/86/1704335/Efficiency-of-Markov-Chain-Monte-Carlo-Tree (visited on 05/11/2017).

[219] Joseph Felsenstein. "Confidence Limits on Phylogenies: An Approach Using the Bootstrap". In: *Evolution* 39.4 (1985), pp. 783–791. ISSN: 0014-3820. DOI: 10.2307/2408678. URL: http://www.jstor.org/stable/2408678 (visited on 05/11/2017).

[220] Andrey Zharkikh and Wen-Hsiung Li. "Estimation of Confidence in Phylogeny: The Complete-and-Partial Bootstrap Technique". In: *Molecular Phylogenetics and Evolution* 4.1 (Mar. 1995), pp. 44–63. ISSN: 1055-7903. DOI: 10.1006/mpev.1995.1005. URL: http://www.sciencedirect.com/science/article/pii/S1055790385710056 (visited on 05/11/2017).

[221] Bradley Efron, Elizabeth Halloran, and Susan Holmes. "Bootstrap confidence levels for phylogenetic trees". en. In: *Proceedings of the National Academy of Sciences* 93.23 (Nov. 1996), pp. 13429–13429. ISSN: 0027-8424, 1091-6490. URL: http://www.pnas.org/content/93/23/13429 (visited on 05/11/2017).

[222] Michael E. Alfaro, Stefan Zoller, and François Lutzoni. "Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence". In: *Molecular Biology and Evolution* 20.2 (Feb. 2003), pp. 255–266. ISSN: 0737-4038. DOI: 10.1093/molbev/msg028. URL: https://academic.oup.com/mbe/article/20/2/255/1003275/Bayes-or-Bootstrap-A-Simulation-Study-Comparing (visited on 05/11/2017).

[223] M. K. Kuhner, J. Yamato, and J. Felsenstein. "Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling."

en. In: *Genetics* 140.4 (Aug. 1995), pp. 1421–1430. ISSN: 0016-6731, 1943-2631. URL: http://www.genetics.org/content/140/4/1421 (visited on 07/10/2017).

[224] Kurt M. Pickett and Christopher P. Randle. "Strange bayes indeed: uniform topological priors imply non-uniform clade priors". In: *Molecular Phylogenetics and Evolution* 34.1 (Jan. 2005), pp. 203–211. ISSN: 1055-7903. DOI: 10.1016/j.ympev.2004.09.001. URL: http://www.sciencedirect.com/science/article/pii/S1055790304002817 (visited on 01/23/2018).

[225] Ziheng Yang and Bruce Rannala. "Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny". In: *Systematic Biology* 54.3 (June 2005), pp. 455–470. ISSN: 1063-5157. DOI: 10.1080/10635150590945313. URL: https://academic.oup.com/sysbio/article/54/3/455/1728906/Branch-Length-Prior-Influences-Bayesian-Posterior (visited on 05/11/2017).

[226] Marc A. Schard et al. "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10". In: *Virus Evolution* 40.1 (2012), W573–W579.

[227] Travis J. Wheeler. "Large-scale neighbor-joining with NINJA". In: 5724.6 (2009), pp. 375–389.

[228] Martin Simonsen, Thomas Mailund, and Christian N. S. Pedersen. "Rapid Neighbour-Joining". en. In: *Algorithms in Bioinformatics*. Ed. by Keith A. Crandall and Jens Lagergren. Vol. 5251. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 113–122. ISBN: 978-3-540-87361-7. DOI: 10.1007/978-3-540-87361-7_10. URL: http://link.springer.com/10.1007/978-3-540-87361-7_10 (visited on 05/18/2017).

[229] D. L. Swofford. *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sunderland, Massachusetts, 2003.

[230] Sergei L. Kosakovsky Pond and Spencer V. Muse. "HyPhy: hypothesis testing using phylogenies". In: *Statistical methods in molecular evolution*. Springer, 2005, pp. 125–181. URL: http://link.springer.com/content/pdf/10.1007/0-387-27733-1_6.pdf (visited on 05/18/2017).

[231] Andrew Rambaut and Nicholas C. Grass. "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees". In: *Bioinformatics* 13.3 (1997), pp. 235–238.

APPENDIX A

GENOME OF THE PITCHER PLANT CEPHALOTUS REVEALS GENETIC CHANGES ASSOCIATED WITH CARNIVORY*

A.1 Abstract

Carnivorous plants exploit animals as a nutritional source and have inspired long-standing questions about the origin and evolution of carnivory-related traits. To investigate the molecular bases of carnivory, we sequenced the genome of the heterophyllous pitcher plant Cephalotus follicularis, in which we succeeded in regulating the developmental switch between carnivorous and non-carnivorous leaves. Transcriptome comparison of the two leaf types and gene repertoire analysis identified genetic changes associated with prey attraction, capture, digestion and nutrient absorption. Analysis of digestive fluid proteins from C. follicularis and three other carnivorous plants with independent carnivorous origins revealed repeated co-options of stress-responsive protein lineages coupled with convergent amino acid substitutions to acquire digestive physiology. These results imply constraints on the available routes to evolve plant carnivory.

A.2 Article

Carnivorous plants bear extensively modified leaves capable of attracting, trapping and digesting small animals, and absorbing the released nutrients (1,2) . Plant carnivory evolved independently in several lineages of flowering plants, providing a classic model for the study of convergent evolution (3) . Cephalotus follicularis (Cephalotus), a carnivorous plant native to southwest Australia that belongs to the monospecific family Cephalotaceae in the order Oxalidales, forms both carnivorous pitcher leaves and non-

*Portions of this chapter were previously published in Nature Ecology & Evolution, 2017, volume 1, issue 3, and are included with the permission of the copyright holder. Authors include Kenji Fukushima, Xiaodong Fang, David Alvarez-Ponce, Huimin Cai, Lorenzo Carretero-Paulet, Cui Chen, Tien-Hao Chang, Kimberly M. Farr, Tomomichi Fujita, Yuji Hiwatashi, Yoshikazu Hoshi, Takamasa Imai, Masahiro Kasahara, Pablo Librado, Likai Mao, Hitoshi Mori, Tomoaki Nishiyama, Masafumi Nozawa, Gergő Pálfalvi, Stephen T. Pollard, Julio Rozas, Alejandro Sánchez-Gracia, David Sankoff, Tomoko F. Shibata, Shuji Shigenobu, Naomi Sumikawa, Taketoshi Uzawa, Meiying Xie, Chunfang Zheng, David D. Pollock, Victor A. Albert, Shuaicheng Li, and Mitsuyasu Hasebe.

carnivorous flat leaves (Fig. 1). Co-existence of the two types of leaf in a single individual plant provides a unique opportunity to understand the genetic basis of plant carnivory through comparative analysis of these serially homologous organs. To this end, we sequenced the Cephalotus genome. A total of 305 Gb of Illumina reads were generated for contig assembly and scaffolding, and 17 Gb of PacBio reads for inter-contig gap filling (Supplementary Table 1). The resulting assembly consists of 16,307 scaffolds totalling 1.61 Gb with an N50 length of 287 kb (Supplementary Table 2), corresponding to 76% of the estimated genome size (Supplementary Fig. 1a). Long-terminal repeat retrotransposons account for 76% of the genome (Supplementary Tables 3 and 4). Syntenic block comparison with the robusta coffee genome, which maintained diploidy since the ancient split from the Cephalotus lineage 4 , reveals mostly one-to-one mappings (Fig. 1c and Supplementary Table 5), indicating that the Cephalotus genome has not experienced further whole genome duplications since the hexaploidy event at the origin of core eudicots 5 (Supplementary Note 1). We annotated 36,503 protein-coding genes (Supplementary Fig. 1b–e), and 72 microRNA (miRNA) loci (Supplementary Table 6) and their potential targets (Supplementary Table 7) using RNA-sequencing (RNA-seq) data of representative tissues (Supplementary Tables 8–10). Orthologous gene groups (orthogroups) were defined using OrthoMCL 6 for the complete gene sets of Cephalotus and eight eudicot species (Supplementary Tables 11 and 12). Analysis of shared singletons indicates that core eudicot genes are conserved in the Cephalotus genome (Supplementary Note 2 and Supplementary Table 13).

Maximum-likelihood gene gain and loss analysis detected lineage-specific expansion of 492 orthogroups in Cephalotus (Supplementary Table 14). Gene ontology (GO) enrichment analysis (Supplementary Tables 15–21) highlighted Cephalotus-expanded orthogroups containing purple acid phosphatases, known as a typical component of digestive fluids 1,7 (Supplementary Table 17). RNase T2, also known as a constituent of digestive fluids 1,8,9 , is enriched among orthogroups composed only of genes from Cephalotus and another

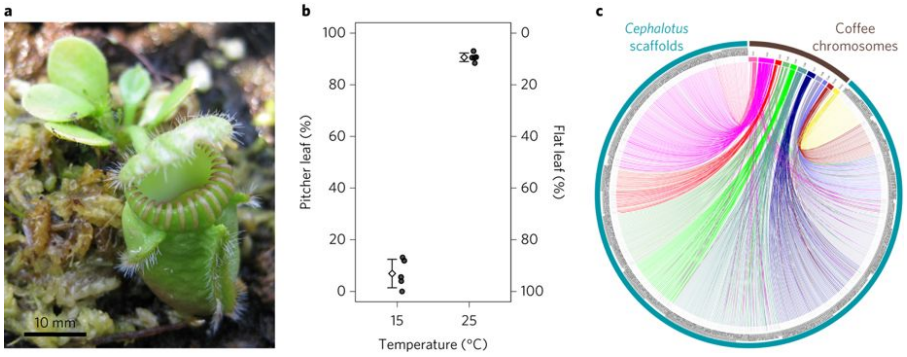


Figure A.1: a, Pitcher and flat leaves. b, Flat and pitcher leaves predominantly produced at 15 °C and 25 °C, respectively, under continuous light conditions. Diamonds and error bars indicate means and standard deviations, respectively. Each filled circle represents an independent experiment with 45 plants. c, Synteny block matching of the Cephalotus genome against the coffee genome4 revealed a one-to-one matching in most genomic loci.

carnivorous plant Utricularia gibba (Supplementary Table 18). Also, the enriched GO term 'cellular response to nitrogen levels' included ten Cephalotus-specific singleton genes encoding dihydropyrimidinases, which have the potential function of acquired nitrogen recycling (Supplementary Table 19). Nitrogen is, in turn, known to be one of the primary limiting nutrients that carnivorous plants derive from prey 1,10 .

As we succeeded in regulating the developmental switch between pitcher and flat leaves by ambient temperature (Fig. 1b and Supplementary Fig. 1f,g), their transcriptomes were compared. The pitcher transcriptome was differentially enriched with cell cycle- and morphogenesis-related GO terms (Supplementary Table 22), which may reflect the morphological complexity of pitcher leaves. Although both developmental and thermoresponsive genes may change their expression in the temperature-dependent leaf switching, certain developmental regulators related to adaxial–abaxial polarity (for example, AS2, YAB5, and WOX1 orthologues) showed higher expression levels in shoot apices bearing pitchers than those terminating in flat leaves (Supplementary Fig. 2), implying the involvement of such factors in pitcher development and evolution. In contrast, the flat leaf transcriptome was enriched with photosynthesis-related GO terms (Supplementary Table 23). These results are compatible with the distinct functional specializations of

carnivory-dominated pitcher leaves versus photosynthesis-dominated flat leaves.

Carnivorous plants attract potential prey by nectar, coloration and scent 1,11,12 . GO terms enriched in the pitcher transcriptome included 'starch metabolic process' and 'sucrose metabolic process' (Supplementary Table 22), which may be related to the production of attractive nectar. Indeed, we detected transcriptional upregulation of certain sucrose biosynthetic genes and members of sugar efflux carriers in pitcher leaves (Supplementary Fig. 3).

The epidermis of carnivorous pitfall traps often develops a slippery, waxy surface that promotes prey capture and prevents them from escaping 1,13 . A cytochrome P450 (CYP) orthogroup was expanded in the Cephalotus lineage (Supplementary Table 14). In a phylogenetic tree, these CYP genes belonged to a clade containing Arabidopsis genes involved in wax and cutin biosynthesis (CYP86 and CYP96A) 14 (Supplementary Fig. 4). These genes, as well as wax ester synthase orthologues (WSD1) 15 , showed pitcher-predominant expression and are tandemly duplicated in the genome (Supplementary Fig. 4), suggesting possible co-regulated involvement of the clusters in slippery surface formation.

Carnivorous plants secrete digestive enzymes for degradation of trapped animals 1,11,12 . Previous studies on several digestive enzymes of Nepenthes spp., Drosera spp., Dionaea muscipula and Cephalotus indicate that pathogenesis-related proteins were co-opted for digestive function as well as for preventing microbial colonization of digestive fluid (refs 16,17,18,19 and refs in Supplementary Table 24). To further investigate the origin and evolution of digestive enzymes of Cephalotus and three other distantly related carnivorous plants (Drosera adelae, N. alata and Sarracenia purpurea), we sequenced fragments of digestive fluid proteins and identified 35 corresponding genes (Fig. 2a and Supplementary Tables 25–28). As Drosera and Nepenthes trace back to a common carnivorous origin in Caryophyllales 3,20 , the four species including Cephalotus therefore cover three independent origins of plant carnivory. Together with previously identified

enzyme sequences including proteins from Dionaea (Supplementary Table 24), we inferred phylogenetic relationships among the digestive fluid proteins (Fig. 2b and Supplementary Fig. 5a–ah). Glycoside hydrolase family 19 (GH19) chitinase, -1,3-glucanase, PR-1-like protein, thaumatin-like protein, purple acid phosphatase and RNase T2 genes showed orthologous relationships among carnivores despite their multiple origins. This result suggests that orthologous genes were repeatedly co-opted for digestive functions in independent carnivorous plant lineages.

To infer putative ancestral functions of these independently arisen digestive fluid proteins, we examined the expression patterns of their phylogenetically most closely related Arabidopsis genes (Supplementary Fig. 5a–ah). Compared with other genes in the same families, these Arabidopsis genes showed a significant tendency to be upregulated on various biotic and abiotic stresses ($P < 0.02$, randomization test) (Supplementary Fig. 5ai). This result suggests that co-option from stress-responsive proteins is a general evolutionary trend in the repeated evolution of carnivorous plant enzymes. Whether they are currently bifunctional—having both carnivorous and non-carnivorous roles—is unclear, but tissue-specific basal expression is probably optimized for carnivory in Cephalotus and N. alata, as the genes are preferentially expressed in their pitcher traps (Fig. 2c,d).

In Cephalotus, three aspartic proteases were identified in the digestive fluid proteome. We found three genomic clusters of aspartic protease genes containing both pitcher-preferential and constitutively expressed genes (Supplementary Fig. 5a–c). Together with the inferred tandem duplications of CYP, this result highlights the roles of gene duplication and subsequent functional divergence in carnivorous plant evolution.

The repeated evolutionary utilization of similar genes may have been accompanied by convergent responses to carnivory-specific selective pressures at the amino acid substitution level. To test this, we developed a tree-based method for the detection of molecular convergence in multigene families, using phylogeny reconciliation between third codon position-derived gene trees and a consensus species tree (Supplementary Note 3, see

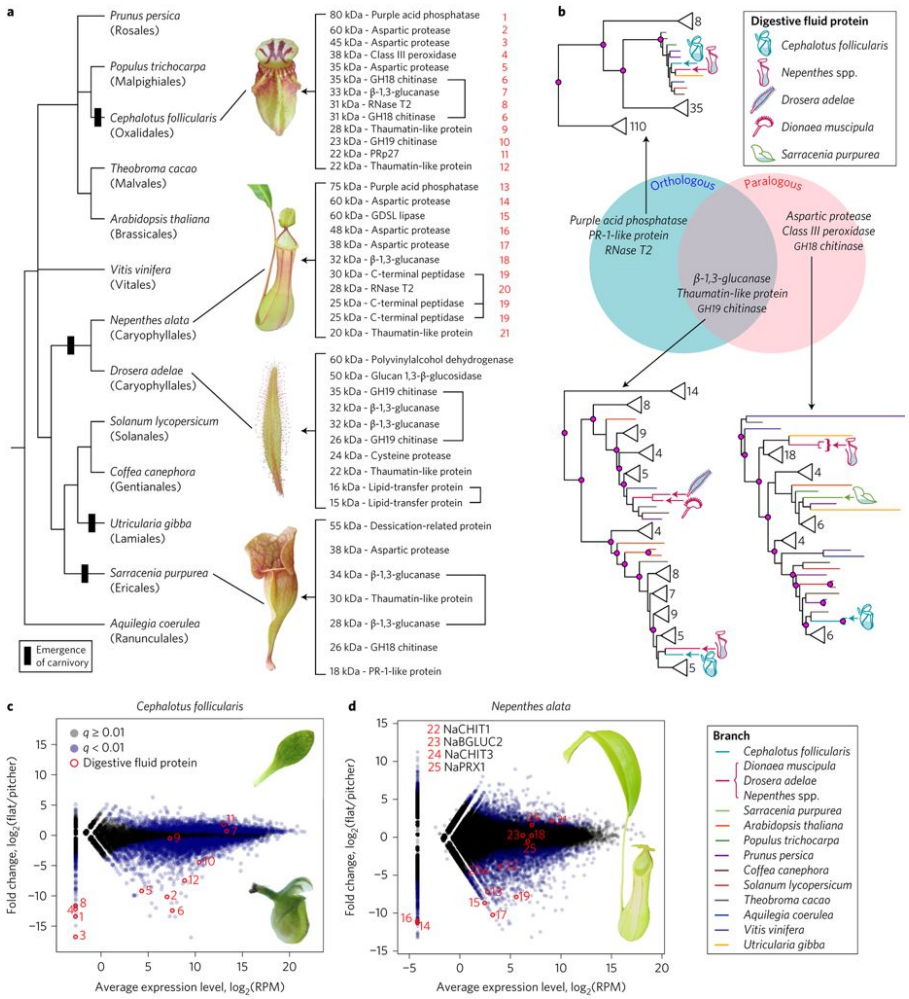


Figure A.2: a, Phylogenetic relationships of independently evolved carnivorous plants and, to their right, the digestive fluid proteins identified through proteomic analysis. Polytomy in the tree represents topological discrepancy between previously reported plastid and nuclear phylogenies (see Methods). Brackets connect protein variants likely to originate from the same gene. b, Phylogeny-based orthologue–paralogue classification. Branch colours denote species identities. Magenta on internal nodes indicates inferred duplication events. Gene numbers in collapsed clades are shown next to triangles. The collapsed clades do not contain genes encoding the digestive fluid proteins but may contain other Cephalotus genes as well as non-carnivorous plant genes. Complete trees are available in Supplementary Fig. 5. c,d, Transcriptome comparison of flat and pitcher organs in Cephalotus (c) and N. alata (d). Red numbers indicate positions of genes encoding digestive fluid proteins identified in this work (1–21, shown in a) and previous studies (22–25, Supplementary Table 24), several of which are outliers showing trap-specific expression.

Methods for the choice of a species tree). Using reconciled trees, the number of digestive enzyme-specific convergent substitutions was inferred on the basis of Bayesian ancestral sequence reconstructions (Fig. 3a,b and Supplementary Fig. 6). By comparing convergent substitution numbers and empirically calculated background-level expectations 21 , we found that GH19 chitinases (Fig. 3a,b), purple acid phosphatases (Supplementary Fig. 6i,j) and RNase T2s (Supplementary Fig. 6m,n) significantly accumulated convergent amino acid substitutions. For all three enzymes, two pitfall-type carnivorous pitcher plants, Cephalotus and N. alata, were associated as convergent branch pairs. Furthermore, for RNase T2, significant molecular convergence was also detected between Cephalotus and the common ancestor of the three Caryophyllales species, D. adelae, D. muscipula and N. alata, which produce sticky, snap and pitfall traps, respectively. Parsimonious inference of character evolution indicates that trapping strategy diversified after the establishment of carnivory in the Caryophyllales 3,20 . Therefore, molecular adaptation of RNase T2 probably occurred both during the evolution of carnivory and subsequently during the establishment of the specific capture strategy of pitfall traps. It is noteworthy that the Cephalotus RNase T2 and purple acid phosphatase genes are located adjacent to each other within a 40 kb interval of the Cephalotus genome (Supplementary Fig. 7). This placement could indicate an arrangement favoured by adaptive, positionally correlated co-expression 22 of these modified carnivorous enzymes (Supplementary Note 4). In light of similar cases of convergent evolution shown for animal digestive enzymes 23,24 , we propose that major changes in nutritional strategy impose a selective pressure strong enough to override evolutionary contingency in both plants and animals.

As protein structure imposes major constraints on amino acid substitutions 25,26,27 , we mapped amino acid residues identified as convergent onto corresponding 3D enzyme models. Convergent positions do not overlap with or cluster around catalytically essential amino acids (Supplementary Fig. 8). Instead, they tend to be located at exposed positions to an extent comparable to divergent substitutions (Fig. 3c), despite the

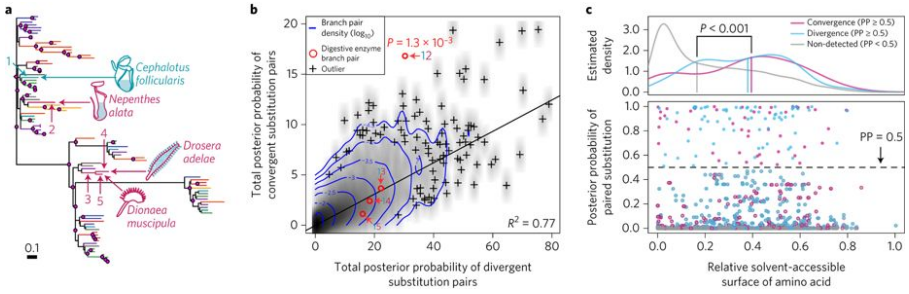


Figure A.3: a, GH19 chitinase phylogeny obtained from the phylogeny reconciliation. Identified digestive enzyme genes are indicated by trap illustrations. Magenta on internal nodes indicates inferred duplication events. The bar indicates 0.1 nucleotide substitutions per site. The complete tree is available in Supplementary Fig. 6q. b, Accumulation of convergent amino acid substitutions in GH19 chitinases. The positions of digestive enzyme branch pairs are indicated by red circles with corresponding numbers in a. Grey tones indicate branch pair density. The line shows a linear regression. c, Relationships between substitution processes and amino acid exposure in protein structures. As the convergent branch pairs in different families showed similar patterns (Supplementary Fig. 8e), data from GH19 chitinases, purple acid phosphatases and RNase T2s are pooled. The bottom panel shows posterior probabilities (PP) of convergent (pink) and divergent (light blue) substitution pairs. The top panel shows density distributions of convergent and divergent loci (PP $\geq$ 0.5, filled pink and light blue in the lower panel) as well as non-detected positions (PP $<$ 0.5, filled grey with outline colour according to the substitution types). P value indicates a statistical difference of medians (vertical lines) (randomization test).

prediction that more exposed positions result in lower convergence probability 28 . Exposed sites are structurally less constrained, and substitutions in such sites are likely to change their interactions with other molecules in solution, rather than changing protein conformation 25,26,27 . During the evolution of digestive enzymes, selective pressures may have come from the digestive fluid environment, which include the presence of insect-derived substrates, high endogenous proteolytic activity, low pH and microbial invasion or symbiosis 1,11,12 . As exposed residues constitute the protein–environment interface, the convergent amino acid substitutions may have been critical factors for the convergent establishment of carnivory across the angiosperms.

In the final phase of carnivorous plant physiology, digested molecules are absorbed into the plant body to promote growth and reproduction 1,29 . We found that various transporters were preferentially expressed in pitcher leaves (Supplementary Table 29). One pitcher-predominant transporter showed phylogenetic affinity to the AMMONIUM TRANSPORTER 1 (AMT1) subfamily (Supplementary Fig. 9), which contains the previously characterized carnivory-related D. muscipula gene DmAMT1 30 . This result, together with the repeated co-option of digestive enzymes already described, indicates utilization of common genetic programs and evolutionary pathways in independently evolved carnivorous plant lineages.

The Cephalotus genome has allowed us to discover numerous genes associated with evolutionary transition to carnivory in plants. In particular, the high degree of convergent evolution in digestive enzymes indicates that there are few available evolutionary pathways for angiosperms to become carnivorous.

A.3 Methods

A.3.1 Plant materials and culture conditions

Axenically grown plants of C. follicularis were obtained from CZ Plants Nursery (Trebovice, Czech Republic) and were maintained in polycarbonate containers (60 x 60 x 100 mm) containing half-strength Murashige and Skoog solid medium 31 supplemented

with 3% sucrose, 1x Gamborg's vitamins, 0.1% 2-(N-morpholino)ethanesulfonic acid, 0.05% Plant Preservative Mixture (Plant Cell Technology) and 0.3% Phytagel, at 25 °C in continuous light. For transcriptome sequencings, D. adelae was cultivated in a peat pot in an incubator at 25 °C in continuous light. N. alata was grown in soil in a greenhouse. S. purpurea was grown in peat-based soil and was maintained in a field. For digestive fluid sampling, C. follicularis, D. adelae, N. alata and S. purpurea were grown in a greenhouse.

A.3.2 Culture conditions for leaf fate regulation

Shoot apices with one or two expanded leaves were collected with fine forceps from plants grown at 25 °C and planted on medium. The plantlets were grown for 12 weeks under a light intensity of 20–40 mol m2 s1. Numbers of youngest pitcher and flat leaves on main shoots were counted for each plantlet (Fig. 1b). Leaves with intermediate shapes were counted as either of the two categories based on morphological similarity.

A.3.3 DNA isolation

Total genomic DNA was isolated from young flat leaves and pitcher leaves of axenically grown plants. Collected leaves were homogenized in liquid nitrogen using a mortar and pestle. The homogenate was transferred into 2x CTAB extraction buffer (2% cetyltrimethylammonium bromide (CTAB), 1.4 M NaCl, 100 mM Tris-HCl (pH 8.0), 20 mM EDTA (pH 8.0)) preheated to 80 °C and was gently agitated at 60 °C for 1 h. An equal volume of chloroform:isoamyl alcohol (25:1) was added and agitated using a rotator at 20 r.p.m. for 10 min at room temperature. After centrifugation at 9,000 x g for 30 min at room temperature, supernatants were transferred to new tubes and supplemented with 1/10 volume of 10% CTAB and an equal volume of chloroform:isoamyl alcohol (25:1). The tubes were shaken with a rotator for 10 min. After centrifugation, supernatants were again transferred to new tubes and an equal volume of isopropanol was added. The tubes were centrifuged and supernatants were discarded. The crude DNA pellet was rinsed with 5 ml of 70% EtOH and air-dried for 10 min. The pellet was dissolved in 200 l of TE (pH 8.0) containing 0.1 mg/ml RNase A, and gently agitated for 60 min at 37 °C. A 1/20

volume of 20 mg ml1 Proteinase K was added, and tubes were incubated at 56 °C for 30 min. Subsequently, the DNA solution was further purified using Qiagen Genomic-tip, following the manufacturer's instructions. DNA concentration was determined using fluorometry with Qubit 2.0 (Life Technologies).

A.3.4 Genome sequencing

Whole-genome shotgun short-read sequences were generated with an Illumina HiSeq 2000 to a depth of approximately 150-fold of the 2 Gb Cephalotus genome using paired-end and mate-pair protocols, according to the manufacturer's instructions (Supplementary Table 1). For long read sequencing, genomic DNA samples were sheared to 6kb or 10kb using g-Tube (Covaris, Massachusetts). Libraries were prepared with DNA Template Prep Kit 2.0 (Pacific Biosciences, California) (3–10kb) following the manufacturer's instructions and sequencing was performed using PacBio RS with C2 chemistry, P2 polymerase and 45-min movies. Using 158 cells, a total of ca. 17Gb were generated with a quality cut-off value of 0.75 (Supplementary Table 1).

A.3.5 Genome size estimation

The size of the Cephalotus genome was estimated by k-mer frequency analysis using JELLYFISH 32 (Supplementary Fig. 1a).

A.3.6 Genome assembly

Illumina paired-end reads with all insert sizes, and mate-pair reads with insert sizes of 2 and 5 kb, were first assembled into 43,308 scaffolds using Allpaths-LG v42381 33 . Fragment filling was applied to paired-end libraries with insert sizes of 170 bp and 250 bp. Standard deviations of insert sizes were set to 10% of insert sizes. Gap filling and further scaffolding were performed by adding mate-pair reads with longer inserts using SSPACE 34 and GapCloser 35 . PacBio reads were subjected to two rounds of error correction using Sprai v0.2.2.3 (`http://zombie.cb.k.u-tokyo.ac.jp/sprai/`) and used for four rounds of iterative gap filling with PBJelly v12.9.14 36 . The final assembly included 16,307 scaffolds with N50 of 287 kb (Supplementary Table 2).

A.3.7 Repeat identification

Repetitive elements of the Cephalotus genome were first identified and masked for gene prediction (Supplementary Tables 3 and 4). De novo prediction of transposable elements was performed using RepeatModeler (`http://www.repeatmasker.org/RepeatModeler.html`) and LTR_FINDER (`http://tlife.fudan.edu.cn/ltr\_finder/`). Known transposable elements were found using RepeatMasker and RepeatProteinMask (`http://repeatmasker.org`). Tandem repeat sequences were screened using Tandem Repeats Finder 37 .

A.3.8 RNA extraction

Plant materials were ground in liquid nitrogen using a mortar and pestle. Total RNA was extracted using the PureLink Plant RNA Reagent (Life Technologies) and subsequently purified using the RNeasy Mini Kit (QIAGEN). DNase treatment was performed during the column purification. Total RNA was qualified using a 2100 Bioanalyzer (Agilent).

A.3.9 Transcriptome sequencing

Extracted RNA was subjected to two rounds of mRNA enrichment using Dynabeads mRNA Purification Kit (Life Technologies) according to the manufacturer's instructions. RNA-seq libraries were prepared using TruSeq RNA Sample Preparation kit v.2 (Illumina). Strand-specific mRNA libraries were constructed using the dUTP second-strand marking method 38 . These libraries were sequenced on an Illumina HiSeq 2000 with three biological replications (Supplementary Table 8).

A.3.10 Gene prediction

For gene predictions, we used homology-based, ab inito and transcript-based methods. Protein data sets of Arabidopsis thaliana, Linum usitatissimum, Manihot esculenta, Populus trichocarpa and Ricinus communis (Supplementary Table 11) were aligned to the Cephalotus genome using tblastn (cut-off: 1e5) and then homology-based gene predictions were generated using GeneWise 39 . We also used Augustus (`http://augustus.gobics.de/`), GENSCAN (`http://genes.mit.edu/GENSCAN.html`),

GlimmerHMM (https://ccb.jhu.edu/software/glimmerhmm/) and SNAP (http://korflab.ucdavis.edu/software.html) for ab initio predictions, with model parameters trained using 730 Cephalotus gene models that were well supported by homology evidence. RNA-seq data generated from 16 samples (Supplementary Table 8) were used for transcript-based predictions with the Bowtie–Tophat–Cufflinks pipeline 40 . These models were merged using GLEAN (http://glean-gene.sourceforge.net/). Finally, gene models that were not in the GLEAN non-redundant gene set but supported by both homology and RNA-seq evidences, or homology-based models (frame shift mutation not allowed and aligning rate >50%), or RNA-seq models encoding proteins 120 amino acids in length, were further added.

A.3.11 Gene annotation

Gene functions were assigned using BLAST searches (E-value cut-off of 105) against the following databases: KEGG (Release 58), nr (NCBI release 20130904), Swissprot and TrEMBL (Uniprot release 201203). Conserved protein domains were assessed by InterPro 41 and InterProScan 42 with applications including HMMPfam, HMMPanther, ProfileScan, HMMSmart, FPrintScan and BlastProDom.

A.3.12 Evaluation of genome assembly and gene prediction

Gene coverage of predicted gene sets was evaluated using CEGMA 2.4 43 (Supplementary Fig. 1b). Read mapping rates of 15 RNA-seq libraries from five tissues ranged from 74.4% to 83.6% (Supplementary Table 9), indicating consistency between the assembled genome and the sequenced transcriptome.

A.3.13 Small RNA extraction and sequencing

Plant tissues were ground in liquid nitrogen using a mortar and pestle. Total RNA was extracted using PureLink Plant RNA Reagent (Life Technologies) and subsequently purified using the miRNeasy kit (QIAGEN). DNase treatment was performed during the column purification. Briefly, for each sample, RNA of the desired size range (18–30 nucleotides) was size-fractionated and ligated with the 5' adapter and, subsequently, the 3'

adapter. Ligated RNA was then subjected to PCR with reverse transcription (RT-PCR) to produce sequencing libraries. Small RNA-seq was performed on an Illumina HiSeq 2000 (Supplementary Table 10).

A.3.14 miRNA prediction and target prediction

Cephalotus miRNA loci were predicted in the genome by both transcriptome- and homology-based methods (Supplementary Table 6). Small RNA-seq reads were mapped onto genomic inverted repeats predicted by EMBOSS einverted 44 . miRNA loci were identified from the mapping results using ShortStack v1.2.3 45 . For homology-based prediction, 7,385 mature miRNA sequences of Viridiplantae species were retrieved from miRbase release 20 46 . These miRNA sequences were mapped onto the Cephalotus genome using patscan 47 , allowing one mismatch. Putative loci mapped by less than five independent miRNAs were excluded. Secondary structures were identified from flanking regions of mapped loci (350 bp) using RNAfold of Vienna RNA Package 2.0 48 , and putative miRNA loci were predicted using miRcheck with default parameters 49 . When putative miRNAs were predicted on both strands of the same loci, the minor locus was collapsed. Putative targets of annotated miRNAs were identified using psRNATarget 50 using default settings (Supplementary Table 7).

A.3.15 OrthoMCL gene classification

Orthologues were clustered by comparison of protein data sets among A. thaliana, C. follicularis, Theobroma cacao, Vitis vinifera, Prunus persica, Coffea canephora, Solanum lycopersicum, U. gibba and P. trichocarpa using BLASTP (cut-off: 105) and OrthoMCL 6 (Supplementary Tables 11 and 12). Protein data sets of the nine genomes were BLAST searched against nr (NCBI release 20140407; BLASTP, E-value cut-off of 105). Functional terms (GO and enzyme codes) were then assigned to each query sequence using Blast2GO (`https://www.blast2go.com/`).

A.3.16 Maximum-likelihood inference of orthogroup gains and losses

We estimated the divergence times of the surveyed species using RAxML version 8 51 , employing tree topologies published previously 52,53,54 . The reported placement of P. persica (Rosales) is discrepant between plastid- 54 and nuclear-based analyses 52,53 . To account for that, we analysed phylogenetic relationships using the single-copy orthologue alignment (see below). Although the bootstrap supports were low, the maximum-likelihood tree supported the nuclear-based topology (Supplementary Fig. 1h), and therefore we placed P. persica as sister to the clade containing A. thaliana, T. cacao, P. trichocarpa and C. follicularis. The placement of V. vinifera is also different among previously published phylogenies 52,53,54 . To account for that, two alternative tree topologies with different placements of V. vinifera were assumed in this analysis. For that, we leveraged the amino acid sequence data of all single-copy orthologues, as defined by OrthoMCL (1,836 1:1 orthologues), after excluding all putative TE sequences identified in BLAST searches against different TE databases (TIGR Plant Repeat Databases 55 , TransposonPSI (`http://transposonpsi.sourceforge.net`) and NCBI's non-redundant (nr) protein database). We then aligned the sequences of each orthogroup with the program M-Coffee 56 and used trimAl 57 to automatically remove poorly aligned regions. The best-fit amino acid substitution model for each multiple sequence alignment was selected using ProtTest 58 and specified in the RAxML analysis under a partitioned scheme. We finally used r8s 59 to obtain the ultrametric trees required for the BadiRate 60 analysis, by applying the nonparametric rate smoothing algorithm 59 to the maximum-likelihood trees and fixing the age of the root to 113 Myr in both cases. This date, a compromise for the two trees we tested, was derived from the average of the 2 BEAST point estimates for the earliest split within the rosid clade (with Vitaceae as one sister lineage), as calculated in ref. 54 (their Fig. 1 and Table 2). The two trees tested are detailed below.

Tree 1:

((V_vinifera: 92.251246, (((T_cacao: 72.791098, A_thaliana: 72.791098): 5.199433,

(P_trichocarpa: 72.150935, C_follicularis: 72.150935): 5.839595): 5.054431, P_persica: 83.044961): 9.206285): 20.748754, (U_gibba: 101.840606, (C_canephora: 89.804910, S_lycopersicum: 89.804910): 12.035696): 11.159394).

Tree 2:

(((U_gibba: 82.830331, (C_canephora: 74.586612, S_lycopersicum: 74.586612): 8.243718): 14.783045, (((T_cacao: 74.611384, A_thaliana: 74.611384): 5.510359, (P_trichocarpa: 73.737693, C_follicularis: 73.737693): 6.384050): 5.848981, P_persica: 85.970724): 11.642652): 15.386624, V_vinifera: 113.000000).

To identify gene families specifically expanded in the Cephalotus genome, we followed the method implemented in refs 4 and 61 , accepting a weighted Akaike information criterion (wAIC) ratio of 2.7 for the best-fit branch model to the second-best-fit model. We ran BadiRate 60 twice, once for each of the two alternative topologies shown above. Only those families strongly supported as expanded (wAIC ratio >2.7) under both of the two alternative topologies were considered for further analyses (Supplementary Table 14).

A.3.17 GO enrichment analysis

Supplementary Table 12 shows the per species summary of orthogroups and singletons in nine plant species. Before BadiRate analyses, orthogroups containing sequences with significant similarity to transposable elements (resulting in E-values <1015 in TBLASTX searches against sequences of the RepBase v19.12 database) 62 were filtered out from all nine genomes. The functional categories (generic GO terms) differentially represented among 493, 495 and 492 Cephalotus-specific expanded genes families (grouping 2,560, 2,567 and 2,557 total genes, respectively), as identified in BadiRate analyses performed using tree 1, tree 2 and the intersection of both trees, are displayed in Supplementary Tables 15, 16 and 17, respectively. Similarly, differential representation of GO generic terms among 2,716 Cephalotus-specific singletons, 237 Cephalotus-specific two-gene families (474 total genes) and Cephalotus-specific 201 multigene families (1,714 total genes) are shown in Supplementary Tables 19, 20 and 21, respectively. Finally, differential representation of

GO generic terms among five pairs of genes unique to Cephalotus and U. gibba is presented in Supplementary Table 18. We performed significance analyses of differential distribution of GO terms by comparing different subsets of genes with the entire complement of genes in the genome using Fisher's exact test (seefor example, ref. 4). To control for multiple testing, the resulting P values were corrected according to ref. 63 .

A.3.18 Selection of differentially expressed genes

Strand-specific RNA-seq reads were mapped to gene models on the genome assembly using Tophat2 64 with minimum and maximum intron lengths of 20 and 20,000 bp, respectively (Supplementary Table 9). Transcript abundances calculated by featureCounts 65 were normalized using the iterative differentially expressed gene elimination strategy (iDEGES) 66 , which consists of sequential TMM-(edgeR-TMM) n normalization 67,68 . Using the normalized reads per million mapped reads (RPM) values, differentially expressed genes were identified by an exact test for a negative binomial distribution 69 and subsequent multiple correction by adjusting the false discovery rate to $q < 0.01$ (ref. 63 ; Fig. 2c,d and Supplementary Figs 2–5 and 9). Normalized RPM values are used in Fig. 2c,d, whereas unnormalized RPM values are plotted in Supplementary Figs 2–5 and 9. The significantly differentially expressed genes were subjected to a subsequent GO-enrichment analysis (Supplementary Tables 22 and 23).

A.3.19 Protein sequencing of digestive fluids

Digestive fluids of C. follicularis, D. adelae, N. alata and S. purpurea were collected from soil-grown plants in a greenhouse. Fluids were freeze dried and stored at room temperature. Dried samples were dissolved in a protease inhibitor cocktail (cOmplete, Mini, EDTA-free, Roche), precipitated with 8% trichloroacetic acid (TCA) and then washed with 90% acetone. They were dissolved in SDS sample buffer (62.5 mM Tris-HCl, 2% SDS, 0.25% BPB, 10% glycerol, 5% 2-mercaptoethanol, pH 6.8), denatured at 95 °C for 3 min and then separated by 12% SDS-polyacrylamide gel electrophoresis. Negative staining was performed using the Gel-Negative Stain Kit (Nacalai Tesque)

according to the manufacturer's instructions. After destaining, proteins were transferred to polyvinylidene difluoride (PVDF) membranes. N-terminal sequences of each protein band were determined by the Edman degradation method using an ABI Procise 494-HT instrument (Applied Biosystems). To obtain internal protein sequences, protein bands were dissected from the gel, destained, dehydrated with 100% acetonitrile for 5 min, dried using an evaporator and then reduced by incubating in 10 mM DTT and 25 mM ammonium bicarbonate at 56 °C for 60 min. After washing with 25 mM ammonium bicarbonate, the proteins were alkylated in 55 mM iodoacetamide and 25 mM ammonium bicarbonate for 45 min at room temperature. After washing with 50% acetonitrile containing 25 mM ammonium bicarbonate, the samples were dried using an evaporator. The proteins were in-gel-digested with 10 ng l1 trypsin in 50 mM ammonium bicarbonate, 10 ng l1 lysyl endopeptidase in 25 mM Tris-HCl (pH 9.0) or 20 ng l1 V8 protease in 50 mM phosphate buffer (pH 7.8) at 37 °C overnight. The digested peptides were extracted twice by sonication in 50% acetonitrile containing 5% trifluoroacetic acid (TFA) for 10 min. The peptides were separated by high-performance liquid chromatography (HPLC) using the Pharmacia SMART System and a reverse-phase column (RPC C2/C18 PC 3.2/3, GE Healthcare Life Sciences, or XBridge C8 5 m 2.1x100 mm, Waters) under the following conditions: constant flow rate of 200 l min1; solvent A, 0.5% TFA, solvent B, acetonitrile containing 0.5% TFA; linear gradient from 10 to 40% (B over A in % (v/v)) over 30 min (1% min1). Separated peptides were then used for protein sequencing by the Edman degradation method.

A.3.20 Transcriptome assembly and identification of transcripts encoding biochemically identified proteins

RNA-seq reads of D. adelae, N. alata, and S. purpurea (Supplementary Table 8) were assembled into transcripts using Trinity (version r2013-02-25) 70 with a 200 bp minimum contig length cut-off. Partial amino acid sequences of digestive fluid proteins were subjected to TBLASTN searches 71 against the transcriptome assemblies and the

Cephalotus gene models to identify the corresponding transcripts (Supplementary Tables 25–28). Sequence variants within a Trinity's component were considered as originating from the same gene.

A.3.21 Preparation of digestive fluid protein data sets

In addition to proteins identified in this study (Supplementary Tables 25–28), we obtained for phylogenetic analyses a number of previously published sequences of digestive fluid proteins 8,9,72,73,74,75,76,77,78 (Supplementary Table 24). Although many protein and transcript sequences for possible digestive enzymes are available (for example, refs 17,79,80,81,82,83,84), we included only genes for which complete coding sequences were available and for which their presence in digestive fluid had been biochemically validated (Supplementary Table 24, last searched 20 January 2016).

A.3.22 Phylogenetic analyses of gene families

Phylogenetic relationships of digestive enzyme genes and other carnivory-related genes were analysed along with their homologues in the annotated genomes of ten angiosperm species (Supplementary Table 11). TBLASTX searches 71 were performed against the above coding sequence (CDS) data sets with an E-value cut-off of 10. After sequence retrieval, multiple alignments were prepared using MAFFT 6.956 85 , and ambiguous codons were removed using trimAl 57 implemented in Phylogears2-2.0.2013.03.15 (`http://www.fifthdimension.jp/products/phylogears/`) with the 'gappyout' option. Poorly aligned sequences were removed using MaxAlign 86 . Phylogenetic trees were reconstructed by the maximum-likelihood method using RAxML v8.0.26 51 with the general time-reversible (GTR) model of nucleotide substitution and four discrete gamma categories of rate heterogeneity ('GTRGAMMA' option). Support for nodes was estimated by rapid bootstrapping with 100 replicates. Trees were rooted at the midpoint between the two most divergent genes. Gene duplication events shown in Figs 2b and 3a were inferred on the basis of species overlap between partitions 87 using a Python package 'ETE3' 88 . The trees were visualized using iTOL 89 .

A.3.23 Detection of orthologous relationships

Orthology of Cephalotus genes and digestive enzyme genes was inferred on the basis of tree topologies reconstructed by the maximum-likelihood method using the ten plant genomes described above (Supplementary Figs 2, 4 and 5). As we cannot exclude the possibility of parallel gene losses, a clade containing genes from at least five plant genomes was designated as a putative orthologous unit.

A.3.24 Expression profiling of Arabidopsis genes

Affymetrix ATH1 (25K) microarray data sets on stress-related experiments were retrieved from ArrayExpress 90 if two or more replicates were available on wild-type Arabidopsis plants (Supplementary Table 30). Robust multi-array average normalized expression data 91 were subjected to heatmap visualization using the R package 'gplots'. Dendrograms were constructed using the furthest neighbour method with Euclidian distances. Significance of differential expression was analysed by a randomization test with 10,000 iterations in which resamplings were performed in each gene family and the sum of expression changes was compared with the original value.

A.3.25 Evaluation of detection methods for molecular convergence

To evaluate different tree reconstruction methods, simulated gene sequences were generated using the R package 'Phylosim' 92 . We used publicly available simulated data sets for 16 fungi species 93,94 . These data sets contain 1,000 simulated tree topologies of gene families, each of which was generated under observed gene duplication and loss rates. Sequences of 300 codons were simulated on the tree topologies of the fungi data set. Codon usage was sampled from the actual frequencies in Saccharomyces cerevisiae 95 . The (transition/transversion rate) was set to 1. The (nonsynonymous / synonymous nucleotide substitution rate ratio (dN/dS)) of each codon position was randomly sampled from a gamma distribution (shape = 0.5, rate = 1). To mimic molecular convergence, two genes were randomly selected to be converged. In terminal branches of selected genes, codon usage of S. cerevisiae was replaced with a biased matrix in which frequencies of

codons coding for two randomly selected amino acids were increased. Increased frequency was calculated by multiplying the original value by 100, and then total frequencies of all codons were scaled to 1.

Gene trees were inferred by the maximum-likelihood method 51 using first, second, third and all codon positions as well as 300 nucleotide random sequences. To obtain a robust tree topology, the gene trees were reconciled with the species tree using Treefix 1.1.10 94 , which incorporates duplication-loss parsimony and a test statistic for likelihood equivalence. Reconciliation was accomplished using default settings for which 1,000 iterations of topology searches were performed and rearrangements were accepted when likelihood was not significantly reduced by the Shimodaira–Hasegawa test 96 (P value threshold of 0.05). Branch lengths of reconciled trees were optimized using RAxML 51 . Finally, the numbers of convergent and divergent substitutions were estimated from the inferred tree topologies and the original simulated alignments using CodeMLancestral 21 (Supplementary Fig. 10). Substitution pairs that result in the same descendant amino acid at the same alignment position in both branches were categorized as convergent changes, whereas the remaining substitution pairs were counted as divergent changes 21,28 .

A.3.26 Detection of molecular convergence in digestive fluid proteins

Genes encoding digestive fluid proteins identified in this study (Supplementary Tables 25–28) and previous research (Supplementary Table 24) were analysed. When corresponding gene sequences for a given species clustered together in the maximum-likelihood trees (Supplementary Fig. 5), they were considered to represent the same gene, whereafter we retained our own sequences to circumvent incorrect inference of gene duplication events in phylogeny reconciliation. A maximum-likelihood tree was reconstructed using third codon position sequences of the trimAl-processed alignments, and it was subsequently reconciled with a species tree prepared from a dated large-scale plastid phylogeny of flowering plants 54 using Treefix 94 with default parameters, except with the number of

iterations increased to 1,000. Although the plastid-based topology 54 is partly different from nuclear-based topology 52,53 (Supplementary Fig. 1h), we employed it because of the necessity to include carnivorous lineages in which nuclear genome sequences are unavailable (for example, Drosera, Nepenthes and Sarracenia). Branch lengths of the reconciled trees were optimized against trimAl-processed CDS alignments using RAxML 51 . The trees were subsequently used for Bayesian ancestral state reconstruction using PhyloBayes 97 over 12,000 generations (2,000 generations of burn-in) with an infinite mixture of GTR substitution models (CAT-GTR model) of amino acid substitution and five discrete gamma categories of rate heterogeneity to calculate posterior numbers of convergent and divergent substitution pairs. Background levels (null hypothesis) of convergent substitution pairs were estimated by a linear regression in which the posterior numbers of convergent changes were predicted by divergent changes 21 . Over-accumulation of convergent changes in a tree was examined by one-sided single-sample proportion tests 98 with Yate's continuity correction 99 and subsequent Bonferroni adjustment for multiple comparisons 100 . Digestive enzyme branch pairs among independent carnivorous plant lineages were examined in the statistical test. Corrected P values are shown in Fig. 3b and Supplementary Fig. 6.

A.3.27 Homology modelling of protein structures

Protein structures of digestive enzymes were analysed using the SWISS-MODEL Workspace 101 . Template models were selected using the 'Template Identification' tool. SWISS-MODEL Template Library IDs of selected templates were 2dkv.1.A, 3zk4.1.A and 1dix.1.A for GH19 chitinases, purple acid phosphatases and RNase T2s, respectively. Predicted models were visualized using UCSF Chimera 1.10 102 . Relative exposure of amino acid surfaces was calculated by dividing solvent-accessible surface in protein structures by the theoretical maximum of corresponding amino acids in Gly-X-Gly tripeptide contexts 103 . The relative solvent-accessible surface area for a paired amino acid substitution was reported by averaging values in proteins constituting the two clades

(Supplementary Fig. 8).

A.3.28 Data availability

The Cephalotus genome assembly and gene models are available from the DNA Data Bank of Japan (DDBJ) with the accession numbers BDDD01000001 to BDDD01016307. The genomic sequences, gene models and other source data are also available at CoGe (Genome ID = 29002) and Dryad (doi:10.5061/dryad.50tq3). The DDBJ accession numbers for DNA-seq (DRR053706–DRR053720), mRNA-seq (DRR053690–DRR051749; DRR029007–DRR29010) and small RNA-seq (DRR058704–DRR058708) are shown in Supplementary Tables 1, 8 and 10, respectively. DDBJ accessions and gene IDs for coding sequences of digestive fluid proteins are provided in Supplementary Tables 25–28.

A.4 Acknowledgements

We acknowledge the following sources for funding: MEXT/JSPS KAKENHI grant numbers 12J04926 (K.F.), 22128008 (T.N.), 221S0002 (M.K.), 16370102 (M.H.), 22128001 (M.H.) and 22128002 (M.H.); MECS CGL2013-45211 (J.R.), R01 GM097251 (D.D.P.), and the US NSF 0922742 and 1442190 (V.A.A.). The sequence data are archived at DDBJ and CoGe. This work was supported by MPRF-NIBB, I. Kajikawa and Y. Matsuzaki for cultivation, J. G. Conran and M. Waycott for locality observation and FGF-NIBB, Y. Makino, S. Ooi and K. Yamaguchi for experiments. Computations were partially performed on the NIG supercomputer and DIAF-NIBB.

A.5 Author Information

A.5.1 Kenji Fukushima & Xiaodong Fang

These authors contributed equally to this work.

A.5.2 Affiliations

National Institute for Basic Biology, Okazaki 444-8585, Japan: Kenji Fukushima, Masafumi Nozawa, Gergő Pálfalvi, Tomoko F. Shibata, Shuji Shigenobu, Naomi Sumikawa & Mitsuyasu Hasebe

Department of Basic Biology, School of Life Science, SOKENDAI (Graduate University for Advanced Studies), Okazaki 444-8585, Japan: Kenji Fukushima, Gergő Pálfalvi, Shuji Shigenobu & Mitsuyasu Hasebe

Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado 80045, USA: Kenji Fukushima, Stephen T. Pollard & David D. Pollock

BGI-Shenzhen, Shenzhen 518083, China: Xiaodong Fang, Huimin Cai, Cui Chen, Likai Mao, Meiying Xie & Shuaicheng Li

Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China: Xiaodong Fang, Huimin Cai & Shuaicheng Li

Department of Biology, University of Nevada, Reno, Nevada 89557, USA: David Alvarez-Ponce

Department of Plant Systems Biology, VIB, Ghent University, Ghent 9052, Belgium: Lorenzo Carretero-Paulet

Department of Biological Sciences, University at Buffalo, Buffalo, New York 14260, USA: Lorenzo Carretero-Paulet, Tien-Hao Chang, Kimberly M. Farr & Victor A. Albert

Department of Biological Sciences, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan: Tomomichi Fujita

School of Food, Agricultural and Environmental Sciences, Miyagi University, Miyagi 982-0215, Japan: Yuji Hiwatashi

Department of Plant Science, School of Agriculture, Tokai University, Kumamoto 869-1404, Japan: Yoshikazu Hoshi

Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-8568, Japan: Takamasa Imai & Masahiro Kasahara

Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal 643, Barcelona 08028, Spain: Pablo Librado, Julio Rozas &

Alejandro Sánchez-Gracia

Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350K Copenhagen, Denmark: Pablo Librado

Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya 464-8601, Japan: Hitoshi Mori

Advanced Science Research Center, Kanazawa University, Kanazawa 920-0934, Japan: Tomoaki Nishiyama

Department of Biological Sciences, Tokyo Metropolitan University, Hachioji 192-0397, Japan: Masafumi Nozawa

Department of Mathematics and Statistics, University of Ottawa, K1N 6N5 Ottawa, Canada: David Sankoff & Chunfang Zheng

Tohoku Medical Megabank Organization, Tohoku University, Sendai 980-8573, Japan: Tomoko F. Shibata

Department of Natural Science, Osaka Kyoiku University, Osaka 582-8582, Japan: Taketoshi Uzawa

A.5.3 Contributions

K.F., N.S. and M.H. maintained and collected samples. K.F. and M.H. established culture conditions to control the leaf dimorphism. K.F., X.F., T.N., S.L., and M.H. coordinated genome assembly and annotation. K.F., X.F., C.C., M.N., S.S. and M.X. prepared and sequenced Illumina libraries. M.N. coordinated PacBio sequencing. T.F.S. performed PacBio sequencing. T.I. and M.K. performed error correction of PacBio reads and inter-contig gap filling. K.F., X.F., H.C. and L.M. performed gene prediction and annotation. K.F. performed miRNA prediction and annotation. D.A.-P. performed singleton gene analysis. C.Z. and D.S. performed synteny analysis. P.L., A.S.-G and J.R. performed BadiRate analysis. L.C.-P. and V.A.A. performed GO-enrichment analysis. K.F., K.M.F., T.-H.C. and G.P. performed gene family analysis. T.F., Y. Hiwatashi, Y. Hoshi, H.M., N.S., T.U. and M.H. performed digestive fluid sampling and protein

sequencing. K.F., S.T.P. and D.D.P. performed convergence analysis. K.F., V.A.A. and M.H. wrote the paper with input from all authors. M.H. initiated and directed the project. K.F., V.A.A., S.L. and M.H. are representatives of each group. K.F. and X.F. should be considered joint first authors.

A.5.4 Competing interests

The authors declare no competing financial interests.

A.5.5 Corresponding authors

Correspondence to Kenji Fukushima or Victor A. Albert or Shuaicheng Li or Mitsuyasu Hasebe.

APPENDIX B

HOW TO WRITE A MARKOV CHAIN MONTE CARLO SIMULATION

B.1 Tutorial

First I came up with a simple problem to solve: fitting a straight line model to noisy linear data. I fully defined the data to be a vector of x values and a vector of y values. I used a hash called $data and two arrays with keys x and y. Then I fully defined the model used to fit the data, namely $y = mx + b$ where b and m are the parameters.I used a hash named $model with a key called parameter to hold an array of the values of the parameters.

I next had to decide what my likelihood function would be. The goal of the likelihood function is to answer the question "how likely is this model given the data?". It is equal to the probability of the data given your model. I chose not to include a prior distribution in my likelihood calculation which results in a de facto uniform prior.

The likelihood function should be at its highest when the model fits the data as well as possible and should decrease as the model fits the data worse. At first, I chose my likelihood function to be the negative of the sum of the squares of the differences between the model and the data.

One problem with this likelihood function is that all the likelihoods are negative but by definition, likelihoods must be positive. So I chose to take e and raise it to the power of the negative of the sum of the squares of the differences between the model and the data. This produces only positive values for the likelihood.

Implementing this, I wrote a squareDiff, sumOfSquareDiffs, lineFromParametersAndXs (to produce the y values predicted by the model at the given x values), and likelihoodFrom-ModelAndData (to implement the likelihood function).

This likelihood has a problem in that it is not a proper probability. One could make this likelihood a proper probability by modelling that the data has gaussian error with a variance of σ.

Now the function on the right is a proper, normalized probability that can be used in a real likelihood calculation. Notice how it introduces one more parameter, sigma, that would need to either be assumed (fixed) or sampled.

Next I needed a way to generate new models. I wrote generateRandomModel and generateNewModelFromCurrent using a step size to choose how far away from the old model the new model should be. This function takes the current model and changes the parameters slightly then returns the new model. This new model will be used when I make a proposal to jump to the new model for every generation of the MCMC.

Then I needed to implement how to decide if the new model is accepted or rejected. I used the Metropolis-Hastings (`http://en.wikipedia.org/wiki/Metropolis\%E2\%80\%93Hastings_algorithm`) method in a function called isProposedModelAccepted(). This function calculates the likelihoods of the current and proposed models then passes them to a function isProposedLikelihoodAccepted() that compares the current and proposed likelihoods and decides if the model is accepted or rejected.

Finally I wrote a function to record the likelihood and parameters of a model to calculate the posterior distribution.

Now we have all the components and functions of the MCMC defined and we can write how the MCMC should actually run. The steps in every generation to running the MCMC are:

1. generateNewModelFromCurrent(current model)
2. If isProposedModelAccepted(), copy the new model to the current model
3. Record the parameters from the current model

B.2 Analyze the output

`http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect008.htm`

B.3 Tips

- Start off by guessing a jump size. If the MCMC freezes after burnin, decrease the jump size. If it never burns in, increase jump size.
- If your acceptance rate is too low, decrease your jump size.
- MCMCs that burn in quickly might freeze later.
- Thinning is useful for reducing the amount of data to handle after running. It is not useful for estimating the posterior distributions of parameters. Correlation in samples is not a terrible thing if you integrate over enough time.

B.4 Perl source code

```
# LinearFitMCMC.pl
# Generate random almost linear data then run a Markov Chain Monte Carlo
# to determine the parameters of the line using a likelihood function of the
# sum of the squares of the differences between the model and the data.

use strict;
use warnings;

local $\ = "\n";



#srand(0);
my $step_size = 1/10; # for updating the parameters
my $generations = 10000;
my $burnin = 1000;
```

```
#my $burnin = 0;

my $x = [map {$_ * .1} (0 .. 10)];
#my $y = $x;
my $y = [map {$_ + (rand(2)-1)/10} @$x];

saveXY("data", $x, $y);

my $data = {
    x => $x,
    y => $y,
};

my $initial_model = {
    parameters => [rand, rand]
};
print "Initial parameters: $initial_model->{parameters}->[0]
    $initial_model->{parameters}->[1]";

#saveXY("model", $x, lineFromParametersAndXs($model->{parameters}, $x));

my $mcmc = {
    previous_likelihoods => [],
    previous_parameters => [],
};
```

```
my $current_model = $initial_model;
push(@{$mcmc->{previous_likelihoods}},
     likelihoodFromModelAndData($current_model, $data));
push(@{$mcmc->{previous_parameters}}, $current_model->{parameters});

for my $g (1 .. $generations) {
    my $new_model = generateNewModelFromCurrent($current_model);

    my $is_accepted = isProposedModelAccepted($data,
                                        $current_model, $new_model);
    if ($is_accepted == 1) {
#       print "model accepted: @{$new_model->{parameters}}";
        my $new_likelihood = likelihoodFromModelAndData($new_model, $data);

#       print "likelihood: $new_likelihood";
        if ($g > $burnin){
            push @{$mcmc->{previous_parameters}}, $new_model->{parameters};
            push @{$mcmc->{previous_likelihoods}}, $new_likelihood;
        }

        $current_model = $new_model;
    }
}

saveMcmc("linear_fit_mcmc", $mcmc);

sub saveMcmc {
```

```
    my ($filename, $mcmc) = @_;
    open my $fh, '>', $filename;

    print $fh "Likelihood\tB\tM";
    foreach my $i (0 .. $#{$mcmc->{previous_parameters}}) {
        print $fh
        "$mcmc->{previous_likelihoods}->[$i]\t"
        ."@{$mcmc->{previous_parameters}->[$i]}";
    }
}

sub isProposedModelAccepted {
    my ($data, $current_model, $proposed_model) = @_;

    my $current_likelihood = likelihoodFromModelAndData(
                                $current_model, $data);
    my $proposed_likelihood = likelihoodFromModelAndData(
                                $proposed_model, $data);

#   print "Current likelihood: $current_likelihood; proposed likelihood: " .
#       "$proposed_likelihood";

    my $is_accepted = isProposedLikelihoodAccepted(
           $current_likelihood, $proposed_likelihood);

    return $is_accepted;
}
```

```
# The likelihood should be e to the negative of
# the sum of squares of the difference
# between the model and the data. Must raise e
# to the power or else likelihoods
# are negative
sub likelihoodFromModelAndData {
    my ($model, $data) = @_;

    my $model_ys = lineFromParametersAndXs($model->{parameters}, $data->{x});
    return exp(-sumOfSquareDiffs($model_ys, $data->{y}))
}

# Use a simple Metropolis-Hastings
# prob of accepting proposal = min (1,
          proposed_likelihood / current_likelhood)
sub isProposedLikelihoodAccepted {
    my ($current_likelihood, $proposed_likelihood) = @_;

#   print "Acceptance ratio: " .
#         $proposed_likelihood / $current_likelihood;
    return rand() < $proposed_likelihood / $current_likelihood;
}

sub mean { my (@array) = @_;
    return sum(@array) / ($#array + 1)
}
```

```
sub sum { my (@array) = @_;
    my $sum = 0;
    foreach my $element (@array) { $sum += $element }
    return $sum
}

sub squareDiff {
    my ($x1, $x2) = @_;
    return ($x1 - $x2) * ($x1 - $x2)
}

sub sumOfSquareDiffs {
    my ($xs1, $xs2) = @_;

    if ($#$xs1 != $#$xs2) {die "xs1 and xs2 are different lengths"}
    my $sum = 0;
    foreach my $i (0 .. $#$xs1) {
         $sum += squareDiff($xs1->[$i], $xs2->[$i]);
    }
    return $sum
}

sub lineFromParametersAndXs {
    my ($parameters, $xs) = @_;

    return [map {$parameters->[0] + $_ * $parameters->[1]} @$xs];
```

```
}

sub generateNewModelFromCurrent {
    my ($current_model) = @_;

    my $new_model = {
        parameters => [
            $current_model->{parameters}->[0] + $step_size * (rand(2) - 1),
            $current_model->{parameters}->[1] + $step_size * (rand(2) - 1)
        ]
    };

    return $new_model
}

sub generateRandomModel {
    return {parameters => [rand(), rand()]};
}

sub saveXY {
    my ($filename, $x, $y) = @_;

    if ($#$x != $#$y) {die "x and y are different lengths"}

    open my $fh, ">", $filename;

    print $fh "x\ty";
```

```
    foreach my $i (0 .. $#$x) {
        print $fh "$x->[$i]\t$y->[$i]";
    }
}
```

APPENDIX C

C++ AND OBJECT ORIENTED PROGRAMMING*

C.1 Converting Perl to C++

Perl is a very convenient prototyping tool because of its flexibility, but when it comes to writing programs that are designed to do a very large amount of computation, Perl is inefficient and thus insufficient. This is why we recommend converting programs to C/C++. While there are some similarities in syntax, such as the structure of for and while loops, C/C++ is a strongly-typed language, and thus it can be non-trivial to convert Perl code into C, especially if one isn't terribly familiar with the language. This guide will discuss three main areas of difference between C/C++ and Perl: differences in the basic program structure, how to deal with C's types, and how to do object-oriented programming (OOP). For ease of reading, this document will refer to "C/C++" as "C" in all future text, even though much of what we discuss is C++ functionality. One of the main differences between Perl and C is the extra step of compiling the code. Whereas Perl code can be run from the command-line with no further steps, C code must first be converted into something machine-specific. This is typically done through the use of a Makefile.

The contents of a Makefile usually look something like this:

```
CC = g++
all : preprocessor postprocessor


preprocessor: preprocessor.o
    \$(CC) -o ../../bin/preprocessor preprocessor.o -I.
postprocessor: postprocessor.o
```

*Authors include Stephen Pollard, Aaron Wacholder, Jaime Merlano, Kathryn Hall, and Seena Shah.

```
    \$(CC) -o ../../bin/postprocessor postprocessor.o -I.
clean:
    rm *.o *.d ../../bin/preprocessor ../../bin/postprocessor
```

Here, CC is the compiler to be used. Note that there are four different targets listed: "preprocessor", "postprocessor", "all", and "clean". The "all" target simply triggers the compilation of both the "preprocessor" and "postprocessor" targets. The "clean" target removes all intermediate files, as well as the final executables. The "preprocessor" target generates the intermediate file (preprocessor.o), and then uses it to build the executable file (../../bin/preprocessor). The "postprocessor" target does the same thing for a different executable.

Whatever C program you write will need a Makefile to compile it. You will always want an "all" target that builds your executable(s) and a "clean" target that gets rid of the executable and any intermediate files built by the compiler.

**** If you make a change to your code, you will need to recompile your code before the executable will reflect the change. ****

To recompile, type "make" at the command-line. This will use the default "all" target and will recompile any .c or .cpp files which have changed. Sometimes, you will make changes to a header (.h) file. Because this file may be shared between one or more .c/.cpp files, you will need to force the compiler to recompile all of those files. To do this, you first clean the workspace with "make clean", then compile using "make".

[Cheat sheet] If you changed: .c - make .cpp - make .h - make clean, make

When a program is written in C, the program is typically broken up into files based on functionality, or based on class (see the section on 'Classes'). Each source file is generally a black box to any part of the program that is contained within another file. The corresponding header file typically contains just enough information to access the functionality contained within. Note that anything which is NOT designed to be accessed

from outside the class should be defined only within the source file and not included in the header.

For example, you might create a file seqop.cpp that contains functions relating to sequence operations that contains the following:

```
#include <string>
#include "seqop.h"
using std::string;
String getReverseComplement(String sequence) {
.
}
char getAminoAcid(String codon) {
.
}
```

and so on.

The header (.h) file would then contain just the function headers and what ever other information is necessary to understand what data will be passed in and returned back out. In this case, we need to know what a 'String' data type is, since we're passing it in and returning it from a function.

Example seqop.h:

```
#include <string>
String getReverseComplement(String);
char getAminoAcid(String);
```

Note that in seqopp.cpp, we use ¡¿'s around the standard library file, and quotes around the file we wrote ourselves. The quotes specify for the compiler to look for a user- designed header file in the same directory as the source file. When using a standard

library, you also need to tell the compiler which part(s) of the library you want to make available. This can be done by specifying the exact member, or by allowing the whole namespace to be used.

[Cheat sheet]

```
#include <standardlibrary>
#include "usergeneratedheader.h"
using <namespace>::<member>;
using namespace <namespace>;
```

The last bit of basic file structure you'll need to begin working with C code is the main() function. Perl does not have a main() function; it begins executing at the first line of code within the file. C requires you to be explicit about your entry point. You must have one and only one main() function, and your basic program should go there. It is not necessary to put a return statement inside the main() function, but it can be used to exit the program early or to indicate some sort of status about the conditions the program was terminated under.

```
int main() {
doStuff();
doMoreStuff();
if(stuffWorked()) {
return 0;
}
else {
return 1; //failed to work
}
}
```

Comments can be indicated either with a // instead of perl's , or they can be bracketed with /* */ for multi-line comments.

[Cheat sheet]

```
//This is a comment
/* This is also one big comment.
It spans multiple lines. */
```

C is a strongly-typed language

Perl is a weakly-typed language. For all intents and purposes, you have scalars, arrays, hashes, and pointers as your datatypes, and the size of data stored in your variables can change dynamically. C is far more strict and requires you be explicit about the size and type of each piece of data, and changing the size of a datastructure often requires more work. For each function, you must explicitly state what data type it returns, and what data types are passed to it.

The basic C datatypes are listed below, as well as vector and map, which are similar to Perl's array and hash data types. Note C data types have a clearly defined size – for really big numbers, you may need to use the size modifier key words to avoid overflow.

[Cheat sheet]

```
//Basic:
int x; // integer number (e.g. 1, 2, -33, 5551)
double y; // floating point number (e.g. 3.14159, 1.99999.
-0.66666667)
char c; //character (e.g. a, D, 9, ?, ~)
void temp; // no type, used when functions return nothing
bool flag; // true or false
//Size-related:
//unsigned = positive numbers only, valid options: unsigned
```

```
long, unsigned int, unsigned short, unsigned char
//Other modifiers:
//const = cannot be modified after declaration
//Pointers and simple arrays
//<type> <name>[<#ofelements>] = simple array, size is
static
//<type> *<name> = pointer
//char *<name> = C-style string if it has an '\0' on the
end
//STL types (usage to be discussed later)
#include <string>; // friendlier strings
#include <vector>; // perl-esque arrays
#include <map>; // hashes
```

If you want to save the value of a variable of one data type in a variable that has a different data type, you will need to "cast" it to the right type first. This can have some unintended effects, so do this with caution.

Examples:

```
int x = 1;
float y = 2.1;
char z = 0;
int main() {
x = x + (int)y; // x will equal 3, not 3.1
y = y + (float)x; // y will now equal 5.1 (2.1 + 3)
//CAUTION!
x = x + (int) z; // x will now equal 51 (3 + 48, because
characters are stored in ascii format, and the decimal
equivalent for 0 is 48.)
```

```
}
The correct way to convert a character to a number is as
follows:
#include <cstdlib> // includes the atoi() function
int x = 1;
float y = 2.1;
char z = 0;
int main() {
x = x + (int)y; // x will equal 3, not 3.1
x = x + atoi(z); // x will now equal 3 (3+0).
}
```

The bool type is specific to C++, and its possible values are true and false. (This is different than Perl, which uses a null value to represent false and anything else is true.)

Void is a special type and is generally used as a return type for a function that doesn't return anything, and as a generic pointer type which, via casting, allows you to pass in or return variables of different types.

C.1.1 Pointers and simple arrays

Because C is a strongly typed language, arrays and pointers must be of a specific type. Array sizes must be declared at the time they are created. If you know how big your array needs to be, and you will not need to shrink or grow it, then this is what you will want to use, but if your array changes size, then you will want to use vectors, covered later in this chapter. Arrays can be treated as special-case pointers, too, as all the entries are located adjacent in memory. Finally, these arrays do not keep track of their size – you will need to do this manually.

```
int main() {
const char nucleotides[5] = {'A','C','G','T','N'};
const char aminoacid[4][4][4] = { // A=0,C=1,G=2,T=3
```

```
{{'K'},{'N'},{'K'},{'N'}}, //AAN
{{'T'},{'T'},{'T'},{'T'}}, //ACN
{{'R'},{'S'},{'R'},{'S'}}, // AGN
{{'I'},{'I'},{'M'},{'I'}}, //ATN
{{'Y'},{'H'},{'Y'},{'H'}}, //CAN
{{'P'},{'P'},{'P'},{'P'}}, //CCN
{{'R'},{'R'},{'R'},{'R'}}, //CGN
{{'L'},{'L'},{'L'},{'L'}}.
//CTN
{{'E'},{'D'},{'E'},{'D'}}, //GAN
{{'A'},{'A'},{'A'},{'A'}}, //GCN
{{'G'},{'G'},{'G'},{'G'}}, //GGN
{{'V'},{'V'},{'V'},{'V'}}.
//GTN
{{'X'},{'Y'},{'X'},{'Y'}}, //TAN
{{'S'},{'S'},{'S'},{'S'}}, //TCN
{{'X'},{'C'},{'Y'},{'C'}}, //TGN
{{'L'},{'F'},{'L'},{'F'}}. //TTN
};
char seqarray[10] = "ATGGGCTAA"; // Note that since this
is a string, seqarray[9] = '\0';
char *seqptr = seqarray;
char temp1, temp2, temp3;
temp1 = seqarray[0]; // temp1 now equals 'A'
temp2 = *seqptr; // temp2 now equals 'A' as well
temp3 = *(seqptr+1); //temp3 now equals 'T', as *(seqptr+1)
is equivalent to doing seq[1];
```

```
seqptr++; // this moves the pointer 1 <datatype> down the
line.
temp1 = *(seqptr); // temp1 now equals 'T'.
}
```

It is possible to create an anonymous variable through the use of new and delete. (Older code might use malloc/calloc() and free() instead.) Make sure you delete any memory you allocate, or you will have a memory leak!

[Cheat sheet]

```
//Pointers:
int i, j=0;
int *ptr; // pointer to an int
ptr = &i; // equivalent of perl: \$ptr=\$i;
j=*ptr; // equivalent of perl: \$j=\$\$ptr;
//Example of new and delete:
char *oligo = new char[16];
delete [] oligo;
```

C.1.2 C++-style strings

Unless you need the speed, it's generally easier to work with C++-style strings.

```
#include <string>
using std::string;
int main() {
string sequence;
string microsat("AT");
string oligo4mer;
oligo4mer = microsat; // oligo4mer now equals "AT"
oligo4mer.append(microsat); // oligo4mer now equals "ATAT"
```

```
oligo4mer.at(2)=G; // oligo4mer now equals "ATGT"
oligo4mer[3]=C; // oligo4mer now equals "ATGC"
oligo4mer.replace(0,2,"GC"); // oligo4mer now equals "GCGC"
oligo4mer.replace(0,3,"GC"); // oligo4mer now equals "GCC"
oligo4mer.replace(0,2,"ATAT",1,3); // oligo4mer now equals
"TATC"
oligo4mer.replace(1,3,3,"N"); // oligo4mer now equals
"TNNN"
int len = oligo4mer.length();
string answer="42";
int numerical_answer = stoi(answer);
}
```

[Cheat sheet]

```
#include <string>
using std::string;
string <name>;
<name>="string_of_chars";
string <name>("string_of_chars");
<name>[<idx>]=<char>;
<name>.at(<idx>)=<char>;
int len = <name>.length();
//convert text to an integer
int numerical_value = stoi("string_of_chars");
// convert text to a floating-point number
float float_value = stof("string_of_chars");
// stod() for double, stol() for long, etc.
// the equivalent to \$string = \$string . \$another_string;
```

```
<name>.append("additional_text");
// Some examples of string replacement
<name>.replace(<start>,<len_to_be_replaced>,<newstr>);
<name>.replace(<start>,<len_to_be_replaced>,<newstr>,<start
_in_newstr>,<len_ofnewstr_to_copy>);
<name>.replace(<start>,<len_to_be_replaced>,<num_of_fill_ch
ars_to_add>,<fill_char>);
```

Using maps and vectors: the C equivalent of %hash and @array To use a map or a vector, you must include the appropriate header file. You must also define the type of the vector or map. For a vector, the capacity is how many elements it can hold before the program must spend time and effort on resizing the structure. It can be useful to pre-set the capacity to avoid spending a lot of time on resizing. Iterating through a vector is a little bit different than accessing an array. Example:

```
#include <vector>
using std::vector;
int main() {
vector <chr> seq;
seq.push_back('a');
seq.push_back('t');
seq.push_back('g');
for(vector<int>::const_iterator i= <name>.begin();i!=
<name>.end;i++) {} # forwards
for(vector<int>::reverse_iterator i= <name>.rbegin();i!=
<name>.rend;i++) {} # reverse
}
[Cheat sheet]
#include <vector>
```

```
using std::vector;
vector <<type>> <name>;
<name>.size() // The current number of elements,
equivalent to \$#array+1
<name>.capacity() // The current capacity before resizing
needs to be done, may be used to pre-allocate space so that
later resizing can be avoided.
<name>.begin() //The first element
<name>.end() // the last element
<name>.rbegin() //Beginning in reverse
<name>.rend() //ending in reverse
vector <<type>>::const_iterator <name>; //for looping
through forwards
vector <<type>>::reverse_iterator <name>; //for looping
through backwards
<name>.at(<num); // return the entry at that position,
equivalent to \$array[\$idx].
<name>.insert(<name>.begin + <num>, <value>); /* Insert a
new value into the middle of the vector at position
<name>.begin + <num>.
Note that any data already in that position or later will
no longer be at the same index it was before. It will be
shifted, and because each entry downstream of that position
will be moved, this operation will be slow. */
std::copy(<name>.begin(),<name>.end(),<destination>); //
copy the contents of a vector
<name>.erase(<index>); // erase the value at this location
```

```
<name>.erase(<indexstart>,<indexend>); // erase these and
everything in between
<name>.clear(); // empty the vector of all entries.
```

CAUTION: erase() and clear() are not a substitute for calling 'delete'. Erase and clear will empty the vector's contents, but delete must be used to destroy the vector. Maps are the C equivalent of hashes, providing a one-to-one mapping of key- value pairs. This example also shows how to use a namespace directly rather than using the "using" statement. You can also iterate through a map using an iterator, just as you would with a vector. Just make sure your iterator is of type map¡¡keytype¿,¡valuetype¿¿::iterator.

Example:

```
#include <map>;
int main() {
std::map <char, int> map1;
map1['A']=37;
map1['C']=22;
}
```

There's just a few more useful things to have in your toolkit for dealing with C types. First off, typedef is a way to define a short name for a specific type, and can make your life much easier. Second, you can define collections of data in something called a struct for anything in Perl where you'd make an anonymous hash and use specific keys to reference the values. And finally, there's a keyword const that can be used to define or declare variables whose values will not be changed. Examples:

```
#include <map>;
#include <string>;
using namespace std;
```

```
struct Chromosome {
string name;
string sequence;
bool haploid;
};
const long long HUMAN_GENOME_SIZE = 3137144693; // By
convention, constants are in all caps
int main() {
typedef map <char, double> seqmap;
seqmap nucleotideFreq;
seqmap proteinFreq;
Chromosome chrX, chrY, chr1;
chrX.name = "X";
chrY.name = "Y";
chr1.name = "1";
chrX.sequence = "ATCGAGAGAGGGTTTAA";
chrX.haploid = true;
chr1.haploid = false;
}
```

[Cheat sheet]

```
typedef <long_definition> <new_name_for_type>;
struct <name> { <type> <defined variable1>; ... <type>
<defined variableN>;}; // the closing ; is required
const <type> <VARIABLE_NAME> = <NEVER_CHANGING_VALUE>;
```

C.1.3 Writing functions in C

Besides needing to declare the types of any variables passed in or out, there are a few other details you need to know about writing functions in C. First of all, you must declare

the function before you use it. You can either write the whole function out, or you can use function prototyping, which is simply including the function header only. If these functions are meant to be called from outside the current file, the function prototype must go in the .h file, while the actual code for the function goes in the .c or .cpp file.

Example:

```
void prototypedFunction();
int main() {
prototypedFunction();
}
void prototypedFunction() {
//code goes here
}
```

Variables declared within a function live and die within that function, and start over when the function is called again. You can also pass variables by reference by declaring it in the function prototype. If you declare a variable as const in the function prototype, then the function cannot change its value. Example:

```
int countNumTimesFuncCalledByVal(int);
int countNumTimesFuncCalledByValConst(const int);
void countNumTimesFuncCalledByRef(int &);
int main() {
int i=0;
int count=0;
int countByRef=0;
count = countNumTimesFuncCalledByVal(count); // count
= 1
count = countNumTimesFuncCalledByValConst(5); // count
```

```
= 6
countNumTimesFuncCalledByRef(countByRef); //
countByRef = 1
countNumTimesFuncCalledByRef(countByRef); //
countByRef = 2
}
int countNumTimesFuncCalledByVal(int count) {
return count+1;
}
void countNumTimesFuncCalledByRef(int &count2) {
count2++;
}
int countNumTimesFuncCalledByValConst(const int prevCount)
{
int newcount = prevCount + 1;
return newcount;
}
```

C.1.4 Loops, if-else, and switch

Loops are pretty straightforward to convert to C from Perl. For loops and while loops are identical to Perl. (Incidentally, if your loop or if-statement consists of one line of code, it is not necessary to include it in 's.) There is for_each functionality for strings, vectors and maps, but to use it, you must create a function for for_each to call that is equivalent to the contents of the loop. C also adds the do-while loop, which is similar to a while loop, except it checks the conditions at the end – therefore, the loop will always execute at least once.

There is an if-else in C, but no elsif keyword, and C adds an additional decision-making structure called switch. Switch is useful when there is a number of possible values

and you wish to do different operations depending on what they are. (Technically, there is a switch-like structure in Perl, but it's a module that you must import.) Note that when using switch statements, if you do not have a break or a return statement, the code will "fall through" and continue to execute further options. C for_each loop example:

```
#include <algorithm> // required for for_each
#include <iostream>
#include <string>
using std::cout;
using std::endl;
void printNucleotide (char nt);
int main() {
String sequence = "ATTTACGGGTAATA";
std::for_each(sequence.begin(),
sequence.end(),printNucleotide);
cout << endl;
}
void printNucleotide (char nt) { cout << nt; }
Do-while example:
do {
getSequences();
}
while (fileEndIsNotReached());
If-Else vs. Switch example:
char rc_elseif(char);
char rc_switch(char);
int main() {
String sequence = "ATTTACGGGTAATA";
```

```
String rc_of_sequence = "";
String rc_of_sequence2 = "";
for(int i=sequence.length()-1;i>=0;i--) {
rc_of_sequence.append(rc_elseif(sequence[i]));
rc_of_sequence2.append(rc_switch(sequence[i]));
}
}
char rc_elseif(char nt) {
if(nt =='A') {
return 'T';
}else if (nt=='T') {
return 'A';
}else if (nt=='C') {
return 'G';
}else if (nt=='G') {
return 'C';
}else if(nt =='a') {
return 'T';
}else if (nt=='t') {
return 'A';
}else if (nt=='c') {
return 'G';
}else if (nt=='g') {
return 'C';
} else {
return nt;
}
```

```
}
char rc_switch(char nt) {
switch(nt) {
case 'A':
case 'a':
return 'T';
case 'T':
case 't':
return 'A';
case 'G':
case 'g':
return 'C';
case 'C':
case 'c':
return 'G';
default:
return nt;
}
}
char broken_switch(char nt) {
char x=nt;
switch(nt) {
case 'A':
case 'a':
x='T';
case 'T':
case 't':
```

```
x= 'A';
case 'G':
case 'g':
x= 'C';
case 'C':
case 'c':
x= 'G';
default:
}
return x; // Because there are no break or return
statements, this function will always return 'G'.
}
char switch_using_break_correctly(char nt) {
char x=nt;
switch(nt) {
case 'A':
case 'a':
x='T';
break;
case 'T':
case 't':
x= 'A';
break;
case 'G':
case 'g':
x= 'C';
break;
```

```
case 'C':
case 'c':
x= 'G';
break;
default:
}
return x; // Because there are no break or return
//statements, this function will always return 'G'.
}
```

C.1.5 Command-line I/O

C++ has the cin and cout operators which make I/O fairly easy. Unlike Perl, you cannot have a variable in the middle of a string, so each variable or expression for which you wish to print the value must be printed separately. For C++, you will use << or >> to separate your entries, rather than a , or <FILE>. To insert a \n, you can either include \n in quotes, or you can use endl; Note that the << and >> seem counterintuitive if you're an avid Linux user...

Example of command-line prints and reads:

```
#include <iostream>
using std::cout;
using std::cin;
using std::endl;
using std::flush;
int main() {
String sequence = "ATTTTAAAAGGGCCGG";
cout << "The sequence is: " << sequence << endl;
int x = 42;
cout << "The answer to life, the universe, and
```

```
everything is " << x << endl;
cout << "Twice the length of our sequence is " <<
(sequence.length() * 2) << endl;
cout << "Enter a new sequence:\n";
cin >> sequence;
cout << "Your new sequence is " << sequence << endl;
char a,b;
cout << "Enter two nucleotides.\n";
cin >> a >> b;
cout << a << " does " << ((a==b)?"":"not ") << "equal
" << b << ".\n";
}
```

File I/O is very similar.

File Example:

```
#include <iostream>
#include <fstream>
#include <string>
using namespace std;
int main(){
String filename;
cout << "Enter filename to open:\n";
cin >> filename;
ifstream inFile(filename,ios::in);
if(!inFile) { cerr << "Error opening " << filename <<
endl;}
ofstream outFile(filename.append(".out"),ios::out);
if(!outFile) { cerr << "Error opening " << filename <<
```

```
endl;}
String sequenceHeader;
inFile >> sequenceHeader;
outFile << sequenceHeader;
```

C.2 Object Oriented Programming

By Stephen Pollard

Document goal: Give a basic overview of Object Oriented Programming (OOP) that provides details that the lab needs.

What is object oriented programming?

OOP is a paradigm or method of programming in which the class (type of object) is the main focus. Inheritance of one class from another is a central concept. These classes can have data members (properties/attributes) and function members (sometimes called methods).

What are the benefits of OOP?

- Divides code base into logical pieces
- Objects hold state

What are the drawbacks?

- There is some overhead to defining objects – esp in development time.
- Takes time to learn new paradigm.

When should one use OOP?

- When developing a program with many interacting parts

When should one NOT use OOP?

- For simple programs. OOP is not worth the overhead in this case.

- When using a more functional style. Functional programming is another paradigm or method of programming that focuses more on the flow of data through the program and less about objects that hold state.

C.3 Object Oriented Example

By Stephen Pollard

Summary

The goal of this section is to teach how to take a Perl script from Perl to C++ to object oriented C++. While doing this, important object oriented and more advanced C++ concepts will be required. The example is a simple cloud building and analyzing script.

How did I convert the Linear Fit MCMC from Perl to OO C++?

How did I just convert my perl script into C++?

- comments -
- replace all the comments with //
- beware that is also used in $ to find the length of an array.

Variables -

- in perl, all variables in the main part of the script (not part of a function) are in the global scope. Although it is not good practice in C++, I left a few variables in the global scope.
- All variables need type information – int, double, bool, string, ofstream, vector

Functions -

- Had to forward declare functions in front of main so main can see them
- Define functions with type information

Includes – had to add includes for other files

Random function – since c does not have a built in way of getting a random number between 0 and 1, I had to make my own function to do that.

How did I convert my C++ program to object oriented C++?

- I used my generic MCMC program and copied and pasted from the non- OOP program.
- Memory management and construction (not work/initialization) in constructo