

Bayesian Structural Phylogenetics

by

Christopher Challis

Department of Statistical Science
Duke University

Date: _____

Approved:

Scott Schmidler, Supervisor

Robert Wolpert

Sayan Mukherjee

Jonathan Mattingly

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2013

ABSTRACT

Bayesian Structural Phylogenetics

by

Christopher Challis

Department of Statistical Science
Duke University

Date: _____

Approved:

Scott Schmidler, Supervisor

Robert Wolpert

Sayan Mukherjee

Jonathan Mattingly

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2013

Copyright © 2013 by Christopher Challis
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

This thesis concerns the use of protein structure to improve phylogenetic inference. There has been growing interest in phylogenetics as the number of available DNA and protein sequences continues to grow rapidly and demand from other scientific fields increases. It is now well understood that phylogenies should be inferred jointly with alignment through use of stochastic evolutionary models. It has not been possible, however, to incorporate protein structure in this framework. Protein structure is more strongly conserved than sequence over long distances, so an important source of information, particularly for alignment, has been left out of analyses.

I present a stochastic process model for the joint evolution of protein primary and tertiary structure, suitable for use in alignment and estimation of phylogeny. Indels arise from a classic Links model and mutations follow a standard substitution matrix, while backbone atoms diffuse in three-dimensional space according to an Ornstein-Uhlenbeck process. The model allows for simultaneous estimation of evolutionary distances, indel rates, structural drift rates, and alignments, while fully accounting for uncertainty. The inclusion of structural information enables pairwise evolutionary distance estimation on time scales not previously attainable with sequence evolution models. Ideally inference should not be performed in a pairwise fashion between proteins, but in a fully Bayesian setting simultaneously estimating the phylogenetic tree, alignment, and model parameters. I extend the initial pairwise model to this framework and explore model variants which improve agreement

between sequence and structure information. The model also allows for estimation of heterogeneous rates of structural evolution throughout the tree, identifying groups of proteins structurally evolving at different speeds. In order to explore the posterior over topologies by Markov chain Monte Carlo sampling, I also introduce novel topology + alignment proposals which greatly improve mixing of the underlying Markov chain. I show that the inclusion of structural information reduces both alignment and topology uncertainty. The software is available as plugin to the package StatAlign.

Finally, I also examine limits on statistical inference of phylogeny through sequence information models. These limits arise due to the ‘cutoff phenomenon,’ a term from probability which describes processes which remain far from their equilibrium distribution for some period of time before swiftly transitioning to stationarity. Evolutionary sequence models all exhibit a cutoff; I show how to find the cutoff for specific models and sequences and relate the cutoff explicitly to increased uncertainty in inference of evolutionary distances. I give theoretical results for symmetric models, and demonstrate with simulations that these results apply to more realistic and widespread models as well. This analysis also highlights several drawbacks to common default priors for phylogenetic analysis, I and suggest a more useful class of priors.

Dedicated to Melissa Challis

Contents

Abstract	iv
List of Tables	xi
List of Figures	xiii
Acknowledgements	xxi
1 Introduction	1
2 Stochastic Evolutionary Model of Protein Structure	6
2.1 Introduction	6
2.2 Materials and Methods	9
2.2.1 Evolutionary Model	9
2.2.2 Parameter Estimation and Computation	14
2.3 Results	16
2.3.1 Inference for Distantly Related Proteins	16
2.3.2 Phylogeny Estimation	20
2.4 Discussion	26
3 Joint Inference of Alignment and Phylogeny with Structure	34
3.1 Introduction	35
3.1.1 Including structural information	36
3.2 Probabilistic evolutionary models	37
3.2.1 Sequence and structure data	37

3.2.2	Representation of a multiple alignment	38
3.2.3	Joint model for sequence and structure	38
3.2.4	Marginal posterior	40
3.2.5	Indel model	41
3.2.6	Substitution model	42
3.3	Structural drift model	42
3.3.1	Model specification	43
3.3.2	Structural diffusion on a tree	44
3.3.3	Branch-specific structural drift rates	45
3.3.4	Non-evolutionary sources of structural variability	46
3.4	Rotations and translations	48
3.4.1	Shrinkage prior for branch-specific diffusivity	51
3.5	Priors	52
3.5.1	Alignment and tree parameters	52
3.5.2	Substitution parameters and indel model parameters	52
3.5.3	Priors for structural parameters	53
3.6	MCMC inference	53
3.6.1	Monitoring convergence	54
3.7	Results and model comparison	54
3.7.1	Structural information improves alignments	55
3.7.2	Structure reduces topological uncertainty	55
3.7.3	Structural information reduces tree errors	58
3.7.4	Structure helps select between alternative topologies	58
3.7.5	Structural drift model improves fit	63
3.8	Heterogeneity in structural drift rates	64

3.8.1	Branch-specific drift rates result in better fit	64
3.8.2	Specific examples of heterogeneity	65
3.8.3	Independence of drift rates and branch lengths	66
3.8.4	Patterns of structural divergence	67
3.8.5	Structural determinants of evolutionary drift rates	68
3.8.6	Parameter inference	69
3.9	Discussion	71
3.9.1	Key conclusions	71
3.9.2	Future work	72
3.10	Availability	74
4	The Cutoff Phenomenon and Piecewise Priors in Models of Biological Sequence Evolution	84
4.1	Introduction	84
4.2	Preliminaries	87
4.2.1	Sequence Evolution Models	87
4.2.2	Definitions	89
4.3	Symmetric Evolution	90
4.3.1	Connection to Distance Estimation	94
4.4	The Cutoff in General Sequence Models	96
4.4.1	Asymmetric Evolution	97
4.4.2	Locating the Cutoff for Finite Sequence Lengths	98
4.5	Simulations	105
4.6	Priors	108
4.6.1	Tree Priors	111
4.6.2	Simulations	117
4.7	Discussion	119

4.7.1	Insertions and Deletions	121
4.7.2	Structural Cutoff	122
5	Summary	124
A	TKF91 Transition Matrix	126
B	Data	127
B.1	Globins from Chapter 2	127
B.2	Simulated data	127
B.3	5-globin dataset	129
B.4	8- and 12-globin datasets	130
B.5	Cysteine proteinase and human protein kinases	131
B.6	Additional figures	133
B.7	Supplementary methods	134
B.7.1	Alignment accuracy and uncertainty	134
B.7.2	Linear relationship between structure and branch length in global- σ model	136
C	Proof of Cutoff	138
C.0.3	Discrete Time	144
	Bibliography	145
	Biography	158

List of Tables

3.1	Effective number of parameters, P_V , and model fit as measured by DIC for structural models with and without a drift component. Results averaged over four independent repetitions for each dataset.	63
3.2	Comparison of inference for global structural parameters on three datasets with and without drift, averaged over four repetitions from independent starting points. In the cysteine proteinase case, most of the variability is explained by baseline variance rather than evolutionary drift, although drift coefficients are significantly higher in certain regions of the tree (not shown).	70
3.3	Posterior quantiles for alignment lengths (L), and TKF92 indel model parameters for globin datasets, aggregated from four independent MCMC chains in each case. All runs used a burn-in of $10m$ iterations, followed by a sampling period of $20m$ (sequence-only) and $40m$ (structural variants).	71
B.1	PDB entries and corresponding species from Figures 2.1, 2.3, 2.4, and 2.5.	128
B.2	PDB entries and corresponding species from Figure 2.7.	129
B.3	The 5-globin dataset.	130
B.4	The 8- and 12-globin dataset, grouped according to observed clades. Sequences marked with a ‡ are present in both datasets. NsGb = non-symbiotic plant globin; Lhb = leghaemoglobin; Ngb = neuroglobin; HGbI = bacterial Hell's gate globin I; Cygb = cytoglobin; CycHb = cyclostome haemoglobin; Hb = haemoglobin; Mb = myoglobin. * - length shown for the portion present in the PDB file.	131
B.5	The cysteine proteinase dataset. Average pairwise identity using the HOMSTRAD alignment is 42%. * - length shown for the portion present in the PDB file.	132

B.6 The human protein kinase dataset. 132

List of Figures

2.1	Posterior distributions for evolutionary distance between human hemoglobin α and a series of increasingly distant globins, obtained by (a) sequence-only model, and (b) combined sequence-structure model. Distributions obtained from both models are nearly identical for the closest three orthologs (horse, turtle, stingray), but begin to diverge beyond this point. The sequence-structure model stochastically orders the proteins according to generally accepted taxonomy, while the sequence model begins to underestimate distances with the lamprey and sea cucumber, and yields completely flat, uninformative posteriors for the fruit fly, ribbon worm, nematode and tuberculosis.	17
2.2	Average 95% credible intervals and medians from 100 simulated descendants of human hemoglobin α . The sequence model with unknown alignment (a) has a sharp transition at $t = 1.5$. Removal of alignment uncertainty (b) delays the transition to 3 expected substitutions. For our combined sequence-structure model we witness this transition still later, at times > 3.5 (see Figure 1).	19
2.3	Posterior distributions of birth rate (λ) between globins of human and (a) lamprey, (b) sea cucumber, and (c) clam obtained under sequence-only (light) and sequence-structure (dark) models. Increasingly diffuse indel rate posteriors lead to underestimated evolutionary distance estimates; $\lambda = .03718$ estimated previously by Hein et al. (2000) is given as a reference (vertical line).	21

2.4	Phylogenies for a group of 24 globins (Table B.1, pairwise sequence identity 12-87%) obtained by different methods. Branch lengths in (b), (c), and (d) have been normalized for topology comparison. (a) Neighbor-joining tree using pairwise posterior mean evolutionary distances under sequence-structure model. (b) Accepted taxonomy (NCBI Taxonomy Database). (c) Topology of (a). Estimated topology closely matches NCBI taxonomy (b), with small differences. (d) Topology of neighbor-joining tree using pairwise posterior mean evolutionary distances under sequence-only model. Some groups are incorrectly separated and several species appear as zero-branch-length intermediate points. Figures created with TreeView (Page, 1996).	30
2.5	Estimated phylogenies for a subset of eight mutually distant globins (pairwise sequence identity 12-43%) . The sequence-structure model still closely matches the established NCBI taxonomy, while MAFFT begins to exhibit significant differences. Additionally, the MAFFT phylogeny has become more sensitive to parameter choice, while the sequence-structure model estimates appropriate parameters from the data.	31
2.6	Phylogenies estimated from simulated highly divergent data sets (average pairwise sequence identity 13-17%). Top: True tree used to simulate data. Left: sequence-structure model estimates. Right: MAFFT estimates. Central green circle indicates correct topology, while red, blue and yellow identify correct pairs where mismatches are made. The sequence-structure model estimates the correct topology in 6 of 10 simulations, and preserves correct pairings of the proteins in all but one. MAFFT produces the correct topology in only one data set, and in all other cases matches pairs incorrectly.	32
2.7	Phylogenetic tree estimated under the sequence-structure model on a highly divergent set of proteins (pairwise sequence identity 9-32%), from which MAFFT is unable to reconstruct a phylogeny.	33
2.8	Posterior distributions of t (light) and $\sigma^2 t$ (dark) between phycocyanin β chain of red alga and human hemoglobin α obtained under sequence-structure model. At such large distances (7% sequence identity), sequence provides no information about t and only the product $\sigma^2 t$ may still be reliably estimated through structural information.	33

3.1	Ten samples from the structural drift model on a tree, with $\sigma^2 = 0.7\text{\AA}^2$ /substitution per site, and $\tau = 70\text{\AA}^2$. With σ^2 set to zero we would see equal variability at each leaf, whereas the structural drift model proposes that structural divergence will be larger over greater evolutionary distances, in accordance with empirical observations. . . .	45
3.2	Alignment accuracy on simulated data (left two panels) for short branches (multiplier = 1) and long branches (multiplier = 2), and on the 5-globin and cysteine proteinase datasets (right panels). Shown are posterior distributions of distance to true alignment (simulated data) or HOMSTRAD alignment (globins and cysteine proteinases) obtained under the sequence-only model (red), and the structural model without (green) and with (blue) drift. In all cases structural alignments are more accurate than sequence-only, with a much lower spread of accuracy values. In many cases the drift model also offers an additional improvement in alignment accuracy. Simulated data results shown for ten realisations on an 8-taxon tree with $\sigma_k^2 = 0.7$ and $\epsilon = 0.5$, with branch lengths multiplied by the multiplier indicated. Similar results were seen with the sum-of-pairs alignment accuracy metric (not shown).	56
3.3	The two most frequently sampled tree topologies for the 5-globin data set under the sequence-only model, with posterior probabilities shown under sequence-only and structural models. Posterior probabilities were computed using the program <code>trees-consensus</code> , written by Benjamin Redelings.	58
3.4	For the cysteine proteinases the consensus topology was the same under all model variants. The labelled edges correspond to splits with significant uncertainty under the sequence-only model (the other three splits had posterior probability 1.00 in all cases). The table below the figure shows the posterior probability of each of these labelled splits under the different model variants.	59
3.5	Posterior distribution of topology errors relative to the true tree for simulated data, analysed under the structural (black) and sequence-only (grey) models, as branch lengths are doubled (left to right). The inclusion of structural information allows the tree to be accurately inferred even for large evolutionary distances, whereas the trees inferred by the sequence-only model become much less accurate. Frequencies shown for the trees on the left, with 6 (top), 8 (middle), and 10 (bottom) leaves, aggregated from 10 independent samples from the model; the maximal half Robinson-Foulds distance for a tree with n leaves is $2(n - 3)$, i.e. 3, 5 and 7 for the three trees above.	75

3.6	Consensus trees for globin datasets, taken from Hoffmann et al. (2010) and Hoffmann et al. (2012a) (top left and bottom left respectively), and inferred using the sequence-only evolutionary model of Miklós et al. (2008) (top right and bottom right). The bottom row features an augmented dataset containing plant globins, as well as a bacterial globin in our analysis. In both cases we obtain the same consensus tree as Hoffmann <i>et al.</i> , including the four-way polytomy in the 12-globin case.	76
3.7	The structurally derived trees have very low uncertainty, and the order of the splits of interest is unchanged by the inclusion of additional sequences. Consensus trees derived under the ϵ -only model (top left and bottom left), and the full structural drift model (top right and bottom right).	77
3.8	Consensus tree with branches scaled by local σ_k^2 parameters for the 12-globin dataset,	78
3.9	Distributions for σ_k^2 for leaf branches in the 12-globin dataset, estimated with low (top right) and high (bottom right) shrinkage to the global σ_k^2 , using the shrinkage mixture prior described in Supplementary Section 3.5.	79
3.10	The consensus tree for the cysteine proteinase and serine proteinase datasets (top, and bottom respectively), with branches scaled according to mean branch length (left), and mean σ_k^2 (right), showing heterogeneity in structural diffusivity coefficients across the tree.	80
3.11	The consensus tree for the full protein kinase set, with branches scaled according to mean branch length (left), and mean σ_k (right), taken across all trees containing a clade that appears in the consensus (σ_k rather than σ_k^2 used for plotting the tree for ease of visualisation). Taxons are colour-coded according to the scheme in Manning et al. (2002): red = tyrosine kinases, blue = calmodulin-dependent kinases, light green = yeast sterile kinases, dark green = (PKA,PKC,PKG), orange = (CDK,MAPK,GSK3,CLK), brown = tyrosine kinase-like, grey = uncategorised.	81

3.12	Summary of distributions for diffusivity coefficients at the leaf branches for the tree given in Figure 3.11. Taxons are colour-coded according to the scheme in Manning et al. (2002): red = tyrosine kinases, blue = calmodulin-dependent kinases, light green = yeast sterile kinases, dark green = (PKA,PKC,PKG), orange = (CDK,MAPK,GSK3,CLK), brown = tyrosine kinase-like, grey = uncategorised. Grey boxes in the background indicate boundaries between clades based on the consensus tree. Median and highest posterior density interval for the global σ_g^2 is shown by the dotted lines running across the boxplot.	82
3.13	Structures for 1mem (left) and 8pch (right), with charged residues highlighted in red (positive) and blue (negative), showing a large number of differences between the two proteins.	83
4.1	Actual total variation (solid lines) vs asymptotic approximation (dashed lines) for standard symmetric evolution on sequences of length $n = 100$ with $m = 2$ (left) and $m = 20$ (right). The approximation matches closely in both cases, but not as well for $m = 20$ at this sequence length.	94
4.2	Location of τ_n (dashed) and $\frac{m-1}{2m}\log(n(m-1))$ (solid) against total variation for standard symmetric evolution with $n = 200$ and $m = 20$. While τ_n is still a valid cutoff for this family, it does not identify the cutoff region as accurately as the more specific expression for symmetric evolution.	99
4.3	Total variation distance and eigenvalue bounds for symmetric random walk with $m = 20$ for $n = 100, 1,000, \text{ and } 10,000$. In all cases, the eigenvalue bound (dashed line) approximates the total variation distance closely toward the end of the cutoff region. The cutoff begins significantly sooner than demanded by the eigenvalue bound, although the cutoff grows sharper as n increases.	102
4.4	Hellinger distance bounds on total variation for different initial sequences \mathbf{x}_0 of length $n = 150$ with the adjusted JTT model. The dashed bounds correspond to an initial sequence composed entirely of serine, one of the fastest evolving amino acids under the model, while the dotted bounds result from an initial sequence of tryptophan. The solid bounds were computed for a sequence drawn from the equilibrium distribution. The vertical line is τ_n and falls well within the solid bounds.	104

4.5	Mean of 2.5%, 50%, and 97.5% percentiles for posteriors of 100 trajectories simulated at several time points, with exponential prior with mean 10 (left) and mean 1 (right). The solid line ($y = x$) gives the true simulated evolutionary distance. With the mean 10 prior, posterior intervals begin to widen at $t = 3$ and by $t = 5$ are very broad, while the mean 1 prior consistently underestimates the true distance.	106
4.6	Solid lines: proportion of simulations at each time point with 95% interval length less than 10 (lower) and 25 (upper). Dashed lines: Hellinger distance bounds on total variation distance. The proportion of simulations without broad intervals follows the lower bound much more closely than the upper, and for intervals of length 10 falls considerably below even the lower bound on total variation distance, indicating that the lower Hellinger bound should be used conservatively.	107
4.7	Simulations with mean 10 exponential prior with width of credible interval less than 3 at $t = 4.4$. This subset maintains tighter intervals only because they appear closer to the initial sequence than they actually are, resulting in underestimation of t from the beginning of the trajectory. This set also makes the transition to wide intervals soon afterward.	108
4.8	Left: The shape of the likelihood for distance estimation of binary sequences. The maximum occurs at $t = -\frac{1}{2}\log(\frac{2k}{n} - 1)$ for $2k > n$ where k is the number of identical positions, after which the likelihood decays asymptotically toward 2^{-n} (curve shown is proportional). The likelihood plateau as $t \rightarrow \infty$ causes the increase in posterior variance as sequence identity decreases. Right: The ratio of equilibrium likelihood to maximum likelihood as the number of observed substitutions increases for a binary sequence of length 500. The ratio is very small while the number of matches is > 280 , but transitions quickly toward 1 beyond this point. The likelihood for more sophisticated models does not depend solely on the number of matches, but exhibits similar behavior.	109
4.9	Mean-one exponential prior (solid line) vs uniform-normal prior with $p = .8$ and $c = 4.5$. The uniform-normal prior results in no shrinkage until c , then drops off more quickly than the exponential prior.	110

4.10	Posterior distributions of evolutionary distance between two protein sequences with mean-one exponential (solid lines) and uniform-normal prior with $p = .8$ and $c = 4.5$. The maximum likelihood estimate is given in each plot as a vertical line. The exponential prior shrinks intermediate distances more than the uniform-normal prior, but applies less shrinkage to extreme values. The normal tail of the uniform-normal prior forces more appropriate decay beyond the plausible region.	112
4.11	Three trees with the same number of taxa and same diameter but widely varying tree lengths, illustrating the difficulty of specifying a prior on tree length.	114
4.12	Marginal prior on tree diameter for trees with 1, 2, 3, 5, and 10 branches when tree prior is chosen such that $\pi(\Upsilon) \propto f(d(\Upsilon))$. The increasing volume of longer diameters causes the distribution to shift to the right as the number of branches increases. The additional shrinkage term of d^{-b+1} results in a marginal piecewise-uniform marginal for any number of branches (solid line).	116
4.13	Examples of the constrained space of trees for a fixed diameter $d = 1$. Left: a full depiction of the constrained 3-branch tree space, where each axis is a branch of the tree. The space consists of three congruent triangles joined at edges. Right: In four dimensions the full space cannot be depicted, but it consists of the union of 3-dimensional polyhedra residing in \mathbb{R}^4 . The figure depicts one of these polyhedra, the 3-dimensional volume bounded by four triangles and a trapezoid.	117
4.14	Tree topology used for simulations. Inner branches are half the length of outer branches. Simulations were performed with the outer branches scaled to 0.5, 1, 1.5, and 2.	118
4.15	The box plots in each figure employ the same methods, from left to right. The y -axis gives the total tree length. Left: MLE of tree length over 20 simulations, as calculated by PhyML. Center: average quantiles of posterior distributions with diameter prior. Right: average quantiles of posterior distributions with mean 1 exponential prior. Horizontal line: true tree length used in simulations. The diameter prior allows inference to closely match the MLE for trees of reasonable length, and appropriately scales back estimates for the longest trees. With the exponential prior, tree length is overestimated even for the smallest trees, and this overestimation becomes more pronounced as the tree grows longer.	120

B.1 Consensus tree for the 5-globin dataset, derived using BALi-Phy with default settings, running until convergence (10,000 iterations, roughly 30 minutes' runtime on a 2.13Ghz Intel core, with burn-in set to 365 as recommended by the `statreport` utility). 133

B.2 Average pairwise mean squared deviation (MSD) for each column plotted against $3\epsilon_i$ [cf. equation (3.12) in the main text] for the maximum likelihood MCMC sample for the 12-globin set under the drift model, showing that for most columns the B -factor-derived information is a good predictor of the MSD (variance), which supports the use of B -factors as a measure of baseline variability. The multiplication by 3 is necessary because MSD contains a contribution from x,y and z . The surplus variability beyond the baseline is modelled by the diffusion component of the drift model. 134

B.3 95% highest posterior density intervals for structural model parameters estimated on simulated data, on a 4-leaf tree. 135

B.4 95% highest posterior density intervals for structural model parameters estimated on simulated data, on a 8-leaf tree. 136

B.5 95% highest posterior density intervals for structural model parameters estimated on simulated data, on a 10-leaf tree. 137

Acknowledgements

I'd like to thank my advisor, Scott Schmidler, for many hours of discussion and deliberation, and for encouragement to develop and complete challenging ideas. I'm grateful to Jeff Thorne for his accessibility and willingness to review my work, despite having no obligation to do so. More than anyone else, I'm thankful to Melissa Challis for encouraging me to apply to graduate school, for moving her life across the country in order for me to pursue this, and for weathering the ups and downs of research for years.

This work was supported by National Institutes of Health grant NIH-1R01GM090201-01.

1

Introduction

Phylogenetics is the study of evolutionary relationships between organisms through molecular sequences or morphological features. There is considerable interest in studying evolution for its own sake, with the eventual goal of establishing a ‘tree of life’ containing all known organisms, tracing the evolutionary path which generated life on earth. In addition, there are several other scientific fields which benefit from understanding the evolutionary relationships between organisms and populations. In epidemiology, phylogenetics provides understanding of relationships between disease strain, thus illuminating ways in which diseases spread. In drug design, phylogenetic studies are becoming increasingly important, particularly for responding to rapidly evolving pathogens, where short evolutionary distances make inference of relationships more difficult. Fields such as anthropology may rely on phylogenetics to identify subpopulations and migration histories. In short, phylogenetics may be applied wherever there is interest in relationships between groups of organisms at any scale.

A key complication in understanding evolutionary relationships occurs due to the presence of both divergent and convergent evolution. In general terms, divergent

evolution occurs when subpopulations of a species begin to evolve under different environmental pressures, resulting eventually in two separate species. Convergent evolution describes the process by which different species acquire similar traits due to similar selection pressures. The same evolutionary processes occur at the level of proteins. Divergence points of a protein molecule in particular will generally correspond to speciation events (although events such as gene duplication can give rise to a protein divergence point within a species), while convergent evolution of proteins depends on evolutionary pressure at the level of protein function. A pair of proteins related through divergent evolution (thus sharing an ancestor) is termed *homologous*, while those that have converged in evolution are *analogous*. The methods in this work deal with establishing the relationships between homologous proteins only.

Molecular phylogenetic reconstruction is the task of estimating the evolutionary tree that generated a set of present-day molecular sequences. This can be done with methods ranging from a simple similarity score calculated for each pair of sequences in the set, to use of sophisticated stochastic process models attempting to capture the evolutionary process as realistically as possible. All but the most rudimentary methods of molecular phylogenetic reconstruction depend heavily upon a sequence alignment. The alignment is a correspondence between sequences identifying homologous residues in the sequences. Thus the alignment defines evolutionary relationships to individual residues, with homologous residues sharing an ancestral residue in the ancestral sequence. Inference of evolutionary distance is therefore sensitive to the alignment, as different alignments may imply the occurrence of a very different set of evolutionary events.

The first alignment methods relied upon a score for each possible residue pairing and a fixed penalty for gaps to allow for inserted and deleted residues (indels) (Smith and Waterman, 1981). Dynamic programming methods then allowed the optimal alignment to be found. However, it is preferable to represent insertions and deletions

(indels) occurring according to an evolutionary process, rather than utilizing a fixed gap penalty, as evolutionary models imply that more indels are expected to occur over longer evolutionary times, allowing indel information to also inform evolutionary distances. An explicit probabilistic model can also be used to produce a distribution over alignments (Thorne et al., 1991, 1992; Holmes and Bruno, 2001).

Protein alignment and phylogenetic reconstruction are both inherently statistical problems because of the apparently random nature of the evolutionary process, and because the true alignments and phylogenies can never be observed. We are only able to observe extant proteins and infer relationships from them. For this reason it is desirable to formulate formal stochastic models for the evolution of proteins, so that estimates are accompanied with appropriate measures of uncertainty.

Many commonly used methods for phylogenetic inference today separate the estimation of alignments and phylogenies into two separate problems. Researchers use one software package to align sequences (Thompson et al., 1994; Katoh et al., 2005), then turn to another to estimate the phylogeny, treating the alignment as known (Huelsenbeck and Ronquist, 2001; Drummond et al., 2013). This ignores the uncertainty in alignment estimation, resulting in overstated confidence in phylogeny estimation (Wong et al., 2008; Lunter et al., 2008). Often the alignment itself is performed on the basis of a ‘guide tree,’ resulting in bias toward the guide tree (Redelings and Suchard, 2005). Ideally, the alignment and phylogeny should be estimated simultaneously, as they are correlated and both uncertain (Drummond et al., 2013; Redelings and Suchard, 2005; Bouchard-Côté and Jordan, 2013).

An important point motivating the developments in this dissertation is that it can be difficult to estimate long evolutionary distances and resolve branching points deep within phylogenies using only sequence information. Homologous proteins need not have similar sequences, because protein function is determined primarily by structure, and many sequences may fold into similar structures (Krissinel, 2007).

Thus there is more evolutionary pressure to preserve protein structures than protein sequences, resulting in many mutations which alter the sequence without significantly altering the structure (Griffiths et al., 1999). For these reasons, protein structures are conserved over much greater time spans than protein sequences (Ingles-Prieto et al., 2013; Russell et al., 1997).

The focus of this dissertation is to extend formal stochastic evolutionary models to protein structure and provide statistical and computational tools for performing joint inference on alignment and phylogeny. I also examine fundamental limits on inference by applying probabilistic theory to sequence models and quantifying evolutionary distances over which they retain information. Chapter 2 introduces a stochastic evolutionary model for protein structures and demonstrates its utility in pairwise distance and parameter estimation. With a drastic reduction in alignment uncertainty, the model allows pairwise evolutionary distance estimation over much longer distances than sequence alone. Chapter 3 extends the model to a full phylogenetic tree to allow proper inference of alignment and tree topology. Model variants which improve agreement between sequence and structure information are introduced, which also allow for estimation of heterogeneous rates of structural evolution throughout the tree. In addition, tree topology Markov chain Monte Carlo (MCMC) proposals are introduced which greatly improve mixing over the complicated topology space. The inclusion of structural information is shown to reduce both alignment and topology uncertainty. All of these capabilities are implemented in the software package StatAlign and associated plugin StructAlign, which is available at <http://statalign.github.io/>. Chapter 4 discusses the ‘cutoff phenomenon’ of Markov chains and its relationship to models of sequence evolution, and shows that it can explain the timeframes over which sequence models lose information about evolutionary relationships. Theoretical results are given for symmetric models, with simulations indicating that results apply to more commonly used models as well.

The implications of this theory for common default priors for phylogenetic analysis are discussed, and an improved class of priors is developed.

Stochastic Evolutionary Model of Protein Structure

2.1 Introduction

This chapter introduces the basic evolutionary model for pairwise alignment of protein sequence/structure pairs, which will be used in later chapters. It builds on, and significantly extends, ideas from Rodriguez and Schmidler (2013). The material in this chapter has been published as “A stochastic evolutionary model for protein structure alignment and phylogeny,” by Challis and Schmidler in *Molecular Biology and Evolution* 29, 3575 - 3587.

Study of biopolymers has long relied heavily on alignment. Alignment algorithms identify regions of similarity between proteins and nucleic acids as a means of identifying common function and inferring homology. Alignment is vital for reconstruction because when sequences share a common ancestor the degree of similarity between them can be used to estimate evolutionary distances. In such situations, formal statistical inference and proper accounting for uncertainty rely on a model of the evolutionary process. Incorporation of alignment uncertainty has been shown to be crucial for proper characterization of uncertainty in phylogenetic reconstruction

(Wong et al., 2008; Lunter et al., 2008). Improved phylogenetic estimation therefore relies in part on reducing alignment uncertainty through more informative evolutionary modeling.

An enormous literature on statistical alignment and phylogeny exists, and we do not attempt a comprehensive summary here. Felsenstein (2003) provides a broad overview. Evolutionary models involve stochastic processes for mutation (Dayhoff et al., 1978; Jones et al., 1992) and insertion/deletion (Thorne et al., 1992, 1991; Miklós et al., 2004), and combined these provide a model suitable for use in Bayesian or maximum likelihood alignment calculations (Bishop and Thompson, 1986; Hein et al., 2000). Use of such models for Bayesian phylogenetics is widespread (Holmes and Bruno, 2001; Huelsenbeck et al., 2002; Lunter et al., 2005b).

Existing evolutionary models for proteins focus on primary structure, treating each protein as a sequence of amino acid characters. (Some work has attempted to incorporate structure-induced dependence among sequence positions - see e.g. Robinson et al. (2003); Rodrigue et al. (2009) - but these models nevertheless operate at the sequence level.) However, it is well known that protein tertiary structure is conserved over much longer time scales than sequence. This is because selective pressure occurs at the level of *function*; because a large percentage of sequence positions contribute to function only through their role in structure formation; and because of the significant redundancy in sequence space of protein folds. As a result, many homologous proteins may share limited sequence similarity, placing them in the “twilight zone” for sequence alignment.

When protein tertiary structure information is available, structural alignment algorithms can often be used to obtain highly accurate alignments in the absence of significant sequence similarity. Many such algorithms have been developed, typically based on optimizing a similarity score, including minimization of the sum of squared distances between aligned C_α coordinates or corresponding pairwise C_α distances.

See Eidhammer et al. (2000); Hasegawa and Holm (2009) for comprehensive reviews. However, as these algorithms are entirely based on optimization of heuristic score functions, most provide little or no accounting for uncertainty or confidence in the resulting alignment, and no possibility of formal statistical inference procedures. In addition, structural scores such as RMSD give only indirect information about evolutionary distance (Chothia and Lesk, 1986; Panchenko et al., 2005; Zhang et al., 2010).

Rodriguez and Schmidler (2013) have developed a probabilistic approach to structure alignment (see also Schmidler (2006); Wang and Schmidler (2013)), and shown that some other structural alignment algorithms are special cases of their model. This provides many advantages, including full accounting for uncertainty in the alignment, enabling adaptive estimation of alignment parameters, and making explicit the statistical assumptions implicit in commonly used score functions. Rodriguez and Schmidler (2013) also provide a joint sequence-structure model, and show significant improvements over a sequence-based approach alone in approximate estimation of evolutionary distances via selecting PAM distances. However, these approaches utilize a gap-penalty formulation, and as such do not serve as a formal, reversible evolutionary stochastic process suitable for use in phylogenetic applications. Gutin and Badretdinov (1994) and Grishin (1997) explore spatial diffusion processes to describe structural evolution and derive equations relating RMSD to sequence identity and evolutionary distance, but in both cases the alignment is assumed to be given. In the absence of an indel process these methods do not provide an explicit evolutionary model for alignment or phylogeny.

In this chapter, we build on these approaches to develop the first stochastic evolutionary process for protein sequence and structural drift simultaneously, suitable for protein alignment and phylogenetic estimation. We show that the inclusion of structural information effectively stabilizes inference of alignments and evolutionary

distances for distant relationships. We conclude with a discussion of several possible extensions to the model to incorporate greater biophysical realism. In this chapter we explore the model through pairwise analyses only; in Chapter 3 we give the extension to a fully Bayesian approach on phylogenetic trees.

2.2 Materials and Methods

2.2.1 *Evolutionary Model*

Our evolutionary model is formulated as a continuous time Markov process composed of three components: an insertion/deletion (indel) model, an amino acid substitution model, and a structural drift model. The indel component follows the Links model of Thorne, Kishino, and Felsenstein (1991). The sequence mutation component follows a standard substitution rate matrix. Finally, the structural component models the evolutionary drift of individual amino acids (represented by C_α coordinates) in three-dimensional space using an Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930; Karlin and Taylor, 1981). In what follows we denote by S^X the sequence of amino acid characters, and C^X the 3D atomic coordinates, of protein X .

Indel Model

Let X and Y represent two proteins, with X an evolutionary ancestor of Y . The indel model describes the process of residues being added to and deleted from X . Thorne, Kishino, and Felsenstein (1991) have previously developed a birth-death model for this process known as the Links model. The model assumes a constant birth rate λ and death rate μ through time and across the length of the protein chain, with independence from site to site. Amino acid survival probabilities can be determined from the Links model for any values of λ , μ , and time interval t (see e.g. Holmes

and Bruno (2001)):

$$\alpha(t) = e^{-\mu t} \tag{2.1}$$

$$\beta(t) = \frac{\lambda(1 - e^{(\lambda-\mu)t})}{\mu - \lambda e^{(\lambda-\mu)t}} \tag{2.2}$$

$$\gamma(t) = 1 - \frac{\mu(1 - e^{(\lambda-\mu)t})}{(1 - e^{-\mu t})(\mu - \lambda e^{(\lambda-\mu)t})} \tag{2.3}$$

Here $\alpha(t)$ is the probability of ancestral survival, $\beta(t)$ is the probability of insertions given at least one surviving descendant, and $\gamma(t)$ is the probability of insertions given ancestral death. These probabilities can be represented as a transition matrix for a pair hidden Markov model (Durbin et al., 1998) with emitting states Match, Insertion, and Deletion, and null Start and End states (Holmes and Bruno, 2001). (See Appendix A for details.) Let M denote the alignment matrix between X and Y , defined as the adjacency matrix of an order-preserving bipartite matching; then $P(M|\mu, \lambda, t)$ is given by the corresponding product of probabilities in this transition matrix.

Although the Links model is the most commonly used, alternative models that allow for larger indel events (Thorne et al., 1992; Miklós et al., 2004) may also be substituted.

Sequence Model

Using the Links model for indels, a complete evolutionary sequence model is obtained by specification of an amino acid substitution rate matrix. Several such matrices exist in the literature; for the examples in this paper we employ the JTT 1992 matrix (Jones, Taylor, and Thornton, 1992) as adjusted by Kosiol and Goldman (2005). We make the standard assumption that the substitution process is in equilibrium and that insertions arise according to the equilibrium distribution. Letting S^X and S^Y represent the sequences of X and Y , the joint likelihood of S^X, S^Y and an alignment

M is:

$$\begin{aligned} P(S^X, S^Y, M | \lambda, \mu, t, Q) &= P(S^X, S^Y | M, t, Q) P(M | \lambda, \mu, t) \\ &= P(S_M^Y | S_M^X, t, Q) P(S_M^Y | \pi) P(S^X | \pi) P(M | \lambda, \mu, t) \end{aligned}$$

where S_M^X and S_M^Y denote the matched (aligned) positions of S^X and S^Y , S_M^Y the unmatched positions of S^Y , Q the substitution rate matrix, and π the equilibrium distribution of characters. $P(S_M^Y | S_M^X, t, Q)$ is given by a product of independent substitution probabilities at each site, obtained by exponentiation of tQ ; $P(S_M^Y | \pi)$ and $P(S^X | \pi)$ are products of the appropriate entries of π ; and $P(M | \lambda, \mu, t)$ is described in the preceding section. This specifies a complete model for sequence evolution of the type employed by many researchers (see e.g. Holmes and Bruno (2001) and references therein).

Structural Model

We define a model for protein structure evolution analogously, building a structural drift process on top of the Links indel process. Let C^X and C^Y be $n_X \times 3$ and $n_Y \times 3$ matrices containing the Euclidean coordinates of the C_α 's of X and Y respectively, where n_X is the number of amino acids in X . Where the sequence model employs a continuous-time, finite-state Markov process, the structure model utilizes a reversible diffusion process in 3D space modeling drift and fluctuation in the amino acid positions (represented by their C_α coordinates). We model positions as drifting independently in space according to an OU process, or Brownian motion with a mean reversion coefficient. (Unlike standard Brownian motion, the OU process has a stationary distribution and thus can be used as a component in a reversible stochastic process.) If $C_{ij}^{(t)}$ is the j th coordinate of the i th C_α at time t , this process is described by the stochastic differential equation

$$dC_{ij}^{(t)} = \theta(\zeta_j - C_{ij}^{(t)})dt + \sigma dB \quad (2.4)$$

where dB is standard Brownian motion, ζ is the mean of the process, and θ represents the strength of the reversion toward the mean. We set $\zeta = 0$ for convenience, as we are concerned with shape and thus location is arbitrary (see Section 2.2.1). This process has the advantage of permitting closed-form expression of the equilibrium distribution

$$C_{ij}^{(t)} \sim N\left(0, \frac{\sigma^2}{2\theta}\right) \quad (2.5)$$

and conditional distribution at time t , given time s :

$$C_{ij}^{(t)} | C_{ij}^{(s)} \sim N\left(C_{ij}^{(s)} e^{-\theta(t-s)}, \frac{\sigma^2}{2\theta}(1 - e^{-2\theta(t-s)})\right). \quad (2.6)$$

Therefore, again assuming that the parent structure C^X and insertions in C^Y follow the equilibrium distribution, the joint likelihood of two structures and an alignment between them can be expressed in a form analogous to the sequence model:

$$\begin{aligned} P(C^X, C^Y, M | \lambda, \mu, t, \sigma^2, \theta) &= P(C^X, C^Y | M, t, \sigma^2, \theta) P(M | \lambda, \mu, t) \\ &= P(C_M^Y | C_M^X, t, \sigma^2, \theta) P(C_M^Y | \sigma^2, \theta) P(C^X | \sigma^2, \theta) P(M | \lambda, \mu, t) \end{aligned} \quad (2.7)$$

with $P(C_M^Y | C_M^X, t, \sigma^2, \theta)$ calculated according to (2.6), $P(C_M^Y | \sigma^2, \theta)$ and $P(C^X | \sigma^2, \theta)$ according to (2.5), and $P(M | \lambda, \mu, t)$ as the appropriate product of transition probabilities from matrix (A.1) in the Appendix. In addition, the marginal likelihood of the observed structures, $P(C^X, C^Y | \lambda, \mu, t, \sigma^2, \theta)$, can be obtained by summing across all possible alignments M using a dynamic programming forward algorithm for pair HMMs (Durbin et al., 1998).

Note that this diffusion process assumes no significant structural reorganization and is best viewed as a model of structural drift *within* the basin of attraction of a particular fold. Evolution between folds is likely a discontinuous event with slowly accumulating sequence changes suddenly crossing into the basin of an alternative fold; our model currently does not account for such between-fold evolutionary events.

The model also assumes independence among sites, as with most commonly used sequence evolution models. Site independence is necessary to maintain analytical tractability of (2.5) and (2.6) after convolving with the indel process, while mean reversion of the OU process (as opposed to Brownian motion) ensures existence of the equilibrium distribution (2.5). Independence does mean that the insertion distribution is diffuse, allowing insertions to arise anywhere in the protein (as dictated by the variance of C^X), without regard to the locations of neighboring amino acids. As a result of these assumptions the model is inadequate as a generative model for physically realistic protein structures, but behaves well for inference conditional on observed structures. Possible extensions of the model toward additional biophysical realism are described in Section 2.4.

Rotation and Translation

For simplicity, we have introduced the structural component of the model under the assumption that X and Y share a common coordinate frame. In practice, the coordinates C^X and C^Y are obtained through experimental methods in which the coordinate frame is arbitrary. Thus when comparing C^Y to C^X we should not distinguish between elements of the set:

$$\{C^Y R + \mathbf{1}\eta : R \in SO(3), \eta \in \mathbb{R}^3\}$$

containing all possible rotations and translations of C^Y , where $SO(3)$ denotes the special orthogonal group of 3×3 rotation matrices. It is possible to resolve this by treating equivalence classes of protein coordinates (shape spaces) using Procrustes transformations (Rodriguez and Schmidler, 2013). However, as the optimal transformation depends upon the full alignment, the likelihood over all alignments cannot be decomposed recursively as required for the HMM forward-backward algorithms. Instead, we treat R and η as uncertain parameters to be estimated (Green and Mar-

dia, 2006; Schmidler, 2006), and calculate likelihoods conditional on a given rotation and translation. Then (2.7) becomes

$$P(C^X, C^Y, M|\Theta) = P(C_M^Y|C_M^X, t, \sigma^2, \theta, R, \eta)P(C_M^Y|\sigma^2, \theta)P(C^X|\sigma^2, \theta)P(M|\lambda, \mu, t)$$

with Θ representing the entire parameter set $(\lambda, \mu, t, \sigma^2, \theta, R, \eta)$.

Joint Sequence and Structure Model

The combined model is obtained by assuming independence between the sequence substitution and structural diffusion processes, conditional on the indel process. Thus the full likelihood of the combined model is simply the product of the individual model likelihoods.

$$P(X, Y|\Theta) = \sum_M P(C^X, C^Y|M, t, \sigma^2, \theta, R, \eta)P(S^X, S^Y|M, Q, t)P(M|\lambda, \mu, t) \quad (2.8)$$

with Θ again representing the entire parameter set. Each factor of the product in (2.8) is provided by one of the preceding sections.

2.2.2 Parameter Estimation and Computation

We take a Bayesian approach to parameter estimation, with the posterior distribution obtained via Markov chain Monte Carlo (MCMC) simulation. Parameters are updated via a random walk Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970), with acceptance probability involving the marginal likelihood, equal to $P(X, Y|\lambda, \mu, t, \sigma^2, \theta, R, \eta)$ given by (2.8). In practice, it is best to update λ and μ together, likewise for R and η , to account for dependence in the posterior. All examples reported below use vague Gamma(1.01, .01) priors for $t, \lambda, \mu, \sigma^2$, and θ , a uniform distribution on rotations for R , and an improper uniform prior for η .

Rotation/Translation Sampling

A random walk for R and η can be constructed as follows. Propose R' from R by generating an axis v uniformly from the unit sphere and angle ϕ from a von Mises distribution with high concentration around 0, and form R' as the composition of R and (v, ϕ) . Then propose $\eta' \sim N(\eta, \tau^2 I)$, and accept or reject the pair R', η' together.

The mixing of R and η can be slow. To remedy this, an independence step is interspersed with the random walk, with proposal distribution constructed as a mixture with components centered at a “library” of plausible transformations. This library is created by computing the least-squares transformation between each pair of consecutive n -residue subsequences between X and Y (Rodriguez and Schmidler, 2013), and excluding all such transformations with $\text{RMSD} > \delta$, where the threshold δ is chosen to arrive at a manageable number of mixture components. Each component of the mixture is the product of a von Mises-Fisher distribution centered on the axis of rotation, a von Mises distribution centered on the angle of rotation, and a normal distribution centered upon the translation. Then the probability density of this distribution at any rotation R' and translation η' is

$$\frac{1}{k} \sum_{i=1}^k \text{vMF}(v'; v_i, \kappa_1) \text{vM}(\phi'; \phi_i, \kappa_2) \text{N}(\eta'; \eta_i, \tau^2 I)$$

where $\text{vMF}(v'; v_i, \kappa_1)$ is the density of the von Mises-Fisher distribution evaluated at v' , the axis of rotation of R' ; $\text{vM}(\phi'; \phi_i, \kappa_2)$ is the density of the von Mises distribution evaluated at ϕ' , the angle of rotation of R' ; $\text{N}(\eta'; \eta_i, \tau^2 I)$ is a multivariate normal distribution centered at η_i and evaluated at η' ; and k is the number of components in the mixture. Mardia and Jupp (2000) provide general information regarding spherical distributions. An algorithm for generating samples from the von Mises-Fisher distribution is provided by Wood (1994). The proposed pair (R', η') is then accepted or rejected according to the Metropolis-Hastings criterion.

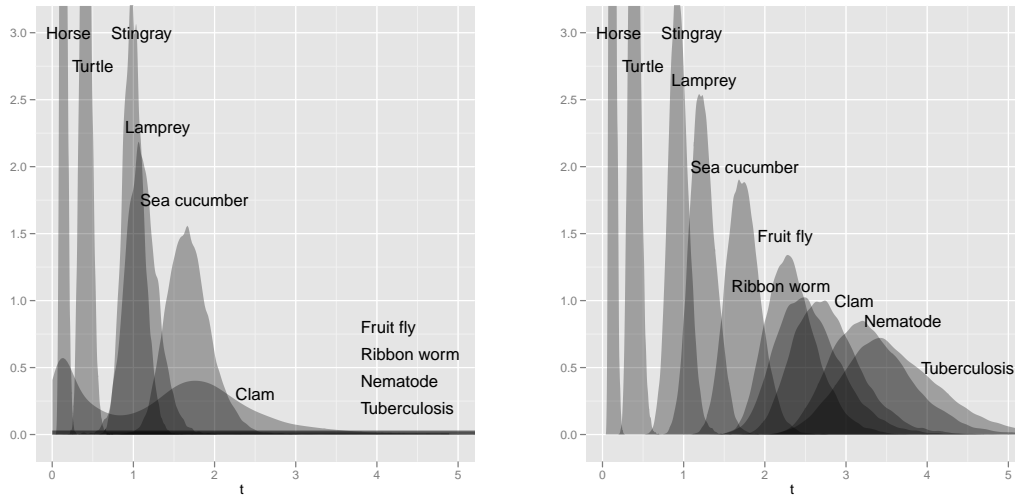
Monitoring convergence

Convergence of the MCMC algorithm was established by the following protocol in all analyses reported in the Results section below. Multiple independent MCMC chains of 50,000 iterations were run from overdispersed starting points, with 10,000 iterations discarded as burn-in. We used 8 chains for the sequence model and 16 chains for the combined model (to account for larger state space due to additional parameters). Convergence was tested by the Gelman and Rubin (1992) diagnostic on the marginal posterior distribution for each parameter.

2.3 Results

2.3.1 Inference for Distantly Related Proteins

The joint sequence-structure evolutionary model described in Section 2.2.1 enables improved alignment and estimation of evolutionary distance and rates between distantly related proteins. To illustrate this on a well-understood protein family, we applied both the sequence-only model and the combined sequence-structure model to estimate the evolutionary distance between the human hemoglobin α subunit and globins from a series of increasingly distant species (Table B.1 in Appendix). Figure 2.1 shows the resulting marginal posterior distributions for evolutionary distance t . In both models, the posterior distribution of t accounts for alignment uncertainty, which is critical for phylogenetic applications (Wong et al., 2008; Lunter et al., 2008). The two models yield comparable results for the pairs with short evolutionary distances and hence high sequence similarity, but as similarity decreases the uncertainty in sequence alignments grows. For sequences with very low similarity, many alignments have virtually equal probability, and the sequence-only likelihood becomes essentially flat for sufficiently large t . The inclusion of structural information via the combined model dramatically reduces this alignment uncertainty, allowing better use



Sequence-only

Sequence-structure

FIGURE 2.1: Posterior distributions for evolutionary distance between human hemoglobin α and a series of increasingly distant globins, obtained by (a) sequence-only model, and (b) combined sequence-structure model. Distributions obtained from both models are nearly identical for the closest three orthologs (horse, turtle, stingray), but begin to diverge beyond this point. The sequence-structure model stochastically orders the proteins according to generally accepted taxonomy, while the sequence model begins to underestimate distances with the lamprey and sea cucumber, and yields completely flat, uninformative posteriors for the fruit fly, ribbon worm, nematode and tuberculosis.

to be made of the sequence information, and also contributes additional information about evolutionary distance through the simple model of structural drift.

This ‘range’ extension of the model through the addition of structure is significant. The sequence-only model begins to differ from the combined model at distances of only 1.5 expected substitutions per site, becoming completely uncertain by 2.5 expected substitutions, while the combined model continues to provide informative posteriors to distances of at least 3.5 expected substitutions. In addition the sequence model parameters (t, λ, μ) become confounded even at modest evolutionary distances (see also Figure 3 below). In contrast, the combined model has no difficulty simultaneously estimating all parameters $(t, \lambda, \mu, \sigma^2, \theta, R, \eta)$ with no loss of precision in t .

Delaying the phase transition The sharp increase in entropy of the posterior distribution under the sequence model is suggestive of the phase transition discussed by Mossel (2003, 2004) (see also Daskalakis et al. (2011)), who shows that if the substitution rate is above a threshold, it is impossible to recover either ancestral sequences or phylogenetic topology over large evolutionary distances using sequence evolution models. Empirically we see the transition even earlier (at shorter distances) than suggested by Mossel’s bounds, between $t = 1.5$ and $t = 2$; this is explained principally by the fact that Mossel’s result assumes a fixed alignment, while accounting for uncertainty in the alignment (and indel rates) causes the uncertainty to grow much faster.

To examine the effect of alignment uncertainty on evolutionary distance estimation, we simulated (under the JTT substitution model, with no indels) the evolution of 100 independent sequence descendants from human hemoglobin α up to time $t = 4$, and another 100 descendants involving indels (using the Links model with rates $\lambda = 0.05$ and $\mu = 0.0504$). We estimated the evolutionary distance from the ancestral sequence to each of the 200 descendants, over the time interval $t \in [0, 4]$ at increments of 0.1, using the MCMC algorithm described above and treating all parameters as unknown, but with the alignment fixed for the first 100 (no indel) sequences. Figure 2.2 shows the quantiles of the posteriors averaged across the 100 simulations. When the alignment is known, the sequence model displays a sharp transition in mean credible interval width at $t = 3$. This transition occurs much earlier (around $t = 1.5$) when the alignment is unknown. In this case, when λ, μ and t are simultaneously estimated the model swiftly loses identifiability, resulting in completely uninformative posterior distributions. The apparent sharpness of the transitions shown here is due in large part to the extremely diffuse Gamma priors, employed intentionally to contrast the sensitivity of sequence and structure inference to the prior. This transition is discussed in much greater detail in Chapter 4.

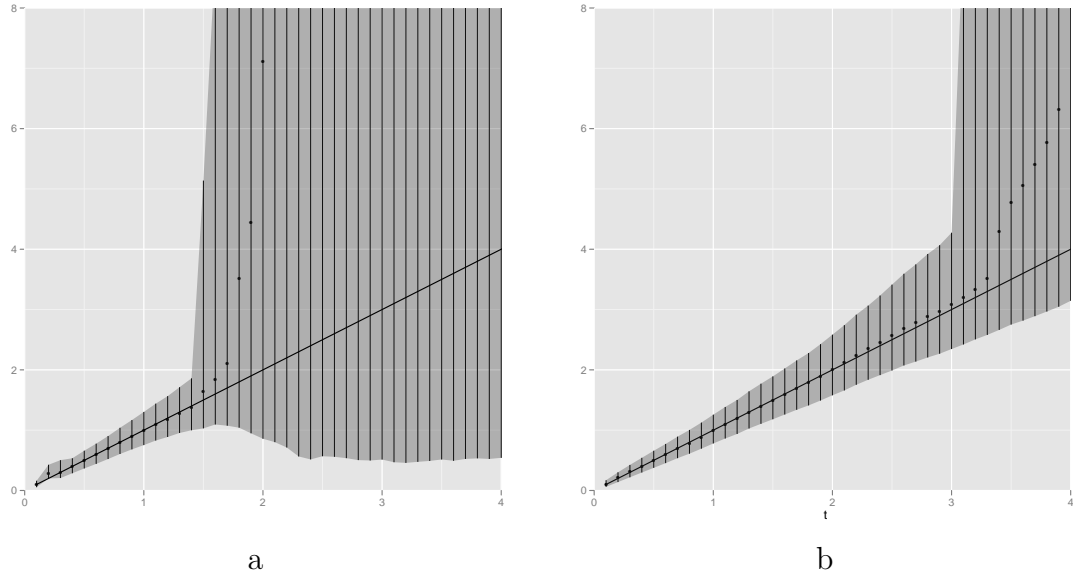


FIGURE 2.2: Average 95% credible intervals and medians from 100 simulated descendants of human hemoglobin α . The sequence model with unknown alignment (a) has a sharp transition at $t = 1.5$. Removal of alignment uncertainty (b) delays the transition to 3 expected substitutions. For our combined sequence-structure model we witness this transition still later, at times > 3.5 (see Figure 1).

The addition of structural information in the combined sequence-structure model dramatically reduces uncertainty in the alignment, which should therefore push the transition back to where it occurs for sequences with known alignment. The results in Figure 2.1 indicate that the transition for the combined model does not occur until after $t = 4$. As shown in Chapter 4, this is in line with the transition that occurs with sequence under less diffuse priors. The range of the model may be extended to even longer evolutionary distances by improving the realism of the structural diffusion model to include stronger information about t and not just M . The local- σ model of Chapter 3 provides a first step in this direction.

Estimating indel rates With the alignment known, the sequence model is able to provide a useful lower bound even after the transition, but this is no longer true when the uncertainty arising from an unknown alignment is accounted for (compare Fig-

ures 2.2a and 2.2b). In particular, underestimation of evolutionary distance occurs due to overestimation of the indel rates λ and μ : as sequence similarity decreases, differences become as likely to be explained by rapid insertions and deletions over a short time period as by substitutions, so deflated estimates of t can result. Around $t = 2$ in Figure 2.2a, approximately half of the simulated proteins exhibited high variance while the other half had narrower posteriors which underestimated the evolutionary distance; thus it is not enough to obtain a concentrated posterior from the sequence model, as larger values of t are likely to be underestimated.

Figure 2.1 contains three examples of this: 2LHB (lamprey), 1HLB (sea cucumber), and 1B0B (clam). For each of these, the sequence-only model gives significantly smaller estimates of distance than the combined sequence-structure model. Examination of the posteriors for λ (Figure 2.3) confirms that indel rates have been overestimated by the sequence model, with underestimation of t particularly extreme in the case of 1B0B as a result of a very diffuse posterior for λ . In fact, the long tailed posterior for λ leads to a second mode, near zero, in the posterior for t (Figure 1). A previous treatment of the Links model based on human α and β globins estimated the insertion rate at .03718 (Hein et al., 2000), and this value was confirmed by Knudsen and Miyamoto (2003); it is provided in Figure 2.3 for reference. Combined model estimates of indel rates are much more stable between protein pairs, and much closer to the results obtained by Hein et al. (2000).

2.3.2 *Phylogeny Estimation*

The uncertainty of evolutionary sequence models with respect to evolutionary distance can dramatically impact the ability to accurately estimate phylogenies (Wong et al., 2008; Lunter et al., 2008). As our joint sequence-structure model drastically reduces this uncertainty, we expect it will have significant impact on stabilizing phylogenetic estimation. Here we explore this impact by estimating pairwise evo-

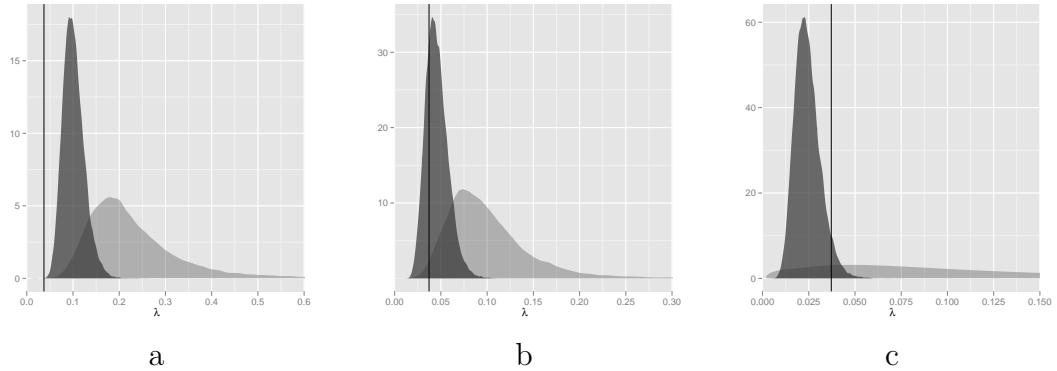


FIGURE 2.3: Posterior distributions of birth rate (λ) between globins of human and (a) lamprey, (b) sea cucumber, and (c) clam obtained under sequence-only (light) and sequence-structure (dark) models. Increasingly diffuse indel rate posteriors lead to underestimated evolutionary distance estimates; $\lambda = .03718$ estimated previously by Hein et al. (2000) is given as a reference (vertical line).

lutionary distances and applying neighbor-joining methods (Saitou and Nei, 1987; Howe et al., 2002). In the future the combined model will be integrated into a full Bayesian simultaneous alignment and phylogeny estimation model, for which it is naturally suited and directly applicable.

Figure 2.4a shows the estimated phylogeny for the hemoglobin α subunits of 24 organisms (Appendix Table B.1) including near and distant relationships (pairwise sequence identity 12-87%), obtained by applying neighbor-joining to the set of pairwise posterior mean distances. Commonly accepted taxonomy from the NCBI Taxonomy Database (Sayers et al., 2009; Benson et al., 2009) is given in Figure 2.4b. The phylogeny estimated similarly (neighbor-joining with posterior mean distances) under the sequence-only model is shown in Figure 2.4d, albeit with unit distances (see below).

The reconstructed phylogeny obtained using the combined sequence-structure model (Figure 2.4a) replicates the established taxonomy almost perfectly. All subgroups are correctly formed, including grouping of the only reptile (turtle) with the birds but as the most distant member. There are minor differences in the topologies

within groups where branch lengths are small and minor changes in length can result in topology changes. A fully Bayesian approach to phylogeny estimation would yield a posterior distribution over competing topologies as well – here our intent is merely to indicate the potential of our sequence-structure model for this purpose.

Using the sequence-only model, many of the pairwise distance posterior distributions remain essentially unchanged from the prior, resulting in broad posterior support and very large posterior means under diffuse gamma priors. In such situations point estimates have little meaning, and posterior intervals convey a near-total lack of information about the evolutionary distance between the two proteins. A phylogeny based solely on the sequence model therefore tends to form clusters of closely related proteins with very large inter-cluster distances, and arbitrary relative placement of the groups. Inter-group branch lengths are so long that visualization of the phylogeny is challenging; for this reason the sequence-based phylogeny is given with unit branch lengths (Figure 2.4d) so that topology can be easily examined. The topology contains multiple inconsistencies with the established taxonomy (Figure 2.4b). The lamprey is separated from other vertebrates, as well as the rockcod from other bony fishes. The mammals appear do not appear as a clade, but as zero-branch-length points between subtrees.

Comparison to Multiple Sequence Alignment

Our sequence-structure model dramatically outperforms the analogous evolutionary sequence model on a pairwise basis, as demonstrated. However simultaneous multiple sequence alignment (MSA) algorithms can also reduce alignment uncertainty, albeit to a lesser extent, through sharing of information. In addition, many phylogenetic methods in common use do not attempt to account for alignment uncertainty. We used MAFFT (Katoh et al., 2005) as a representative, widely used MSA algorithm, and compared the resulting tree with that estimated under our model for the group

of 24 globins of Figure 2.4. Default parameters were used for MAFFT. There are no major differences between the trees estimated by MAFFT and our model, indicating that the combination of MSA and selective use of multiply conserved positions used by MAFFT also does a good job of stabilizing the tree. Note however that these procedures, while adding robustness, do not correspond to an explicit evolutionary model as in our case.

More importantly, multiple sequence alignment algorithms rely on the presence of close homologs. There are several closely related groups in the set of 24 globins, making this well suited for an MSA approach. We performed the comparison again after removing the closely related proteins to arrive at a subset of eight mutually distant globins (pairwise sequence identity 12% - 43%); Figure 2.5 compares the resulting phylogeny under our sequence-structure model with that produced by MAFFT. The phylogeny from our model remains consistent with the established taxonomy and with the tree obtained using the full set of globins, with only a minor shift in the placement of the nematode. The MAFFT phylogeny, however, becomes unstable, separating the lamprey from the other vertebrates. Changing the MAFFT default substitution matrix from BLOSUM62 to BLOSUM30 (more appropriate for distant homologs) has little effect, while modifying the gap penalty parameter caused MAFFT to perform worse.

To further examine the different potential of the sequence-structure model and MSA approaches to analyze distantly related proteins, we simulated ten sets of six pairwise-distant descendants of the α subunit of the human globin at the leaves of a symmetric tree (top of Figure 2.6) with inner branch lengths .35 and outer lengths 1.2 (pairwise sequence identity 13% - 17% on average). Simulation parameters were ($\lambda = .03, \mu = .0302, \sigma^2 = 0.7, \theta = 0.005$) – values typically estimated from observed globins. To further challenge our structure model, insertions in simulated structures were placed at the midpoint of their neighbors, as the independence of

the insertion distribution (2.5) would otherwise make them easier to identify than naturally-occurring insertions. For each of the ten simulated data sets we estimated the underlying phylogeny using MAFFT with default parameters, and using our joint model as before (neighbor-joining on pairwise posterior means). The results are shown in Figure 2.6. The sequence-structure model arrives at the correct topology in six of the ten cases, and preserves correct nearest neighbors in three of the other four. MAFFT only estimates the topology correctly in one instance, and mismatches neighbors in all of the rest. The principal difficulty for the multiple sequence algorithm is insufficient sequence information to resolve the alignment when sequences are highly divergent. The problem is exacerbated by reliance upon a single optimal alignment, which is highly uncertain. Our model benefits from both the Bayesian averaging over all possible alignments, and also especially from the dramatic reduction of alignment uncertainty upon incorporating structural information, resulting in significantly improved phylogeny estimation.

Indeed, the effect of this stabilization extends to sets of proteins for which multiple sequence alignment fails completely. With a broader set of globin-*like* proteins (pairwise sequence identity 9-32%, Table B.2), MAFFT returns an error message that a reliable phylogeny cannot be produced. Our model continues to be effective at these distances; the phylogeny is given in Figure 2.7. The tree continues to correctly preserve the subtree containing the human globin, with the hagfish and sea cucumber as nearest neighbors. The extracellular giant hemoglobins of the earthworm and beardworm are placed together, and the nematode is the last multicellular organism before arriving at the microbes. This tree is not intended as a definitive estimate – a fully Bayesian treatment involving phylogeny sampling instead of neighbor-joining would be preferable to deal with the multiple near-polytomies in the tree – but these results nevertheless illustrate the significant improvement available from the joint sequence-structure model.

At extreme evolutionary distances (7% sequence identity) even the sequence-structure model becomes nearly unidentifiable, even when proteins share a common fold, for the following reason: as illustrated in Figure 2.2, there is a sharp threshold past which sequence information provides only a lower bound on evolutionary distance, even in the case of fixed alignment. Beyond this threshold, sequences are effectively in equilibrium and no longer provide any information for estimating t . At this point the structure component of the model provides all information about t , but the OU process by itself is identifiable only up to the product $\sigma^2 t$. (At shorter distances Q serves to determine the scale for t , making σ^2 and t simultaneously estimable.) Figure 2.8 demonstrates the relative precision of $\sigma^2 t$ to t on these time scales, for comparing the β subunit of phycocyanin from red alga with the α subunit of human hemoglobin. Thus at the farthest within-fold distances, a structure-only approach based on $\sigma^2 t$ as a measure of distance between proteins can still provide some information about evolutionary relationships, but we would need to fix σ^2 (analogous to scaling Q to one expected substitution per time unit) in order to estimate t itself.

Reconstruction regimes

Our results highlight the existence of multiple “regimes” of reconstructability, depending on divergence times of the input proteins. When sequence is sufficiently well-conserved that pairwise alignments are easily resolved, neighbor-joining works well. As divergence increases into the “twilight zone” of sequence similarity, pairwise alignments begin to fail but can be recovered by pooling information across the set of sequences using MSA. However, as demonstrated above, a third regime exists when sequence information is inadequate for even MSA. In this case, our model demonstrates that structural information can still resolve the alignment, and conditional on alignment the sequences still contain sufficient similarity to infer evolutionary dis-

tance. Finally, as sequences become widely divergent we enter a fourth regime where even though structural similarity may resolve the alignment, sequences are effectively in equilibrium and provide essentially no information about either the alignment or the evolutionary distance. In this last situation, our structural model may still be used to estimate divergence times t , but only if σ^2 is fixed by other means (see Fig 8), analogous to scaling sequence substitution models to one expected substitution per time unit.

2.4 Discussion

We have described a stochastic process model for combined protein sequence and structure evolution, suitable for use in likelihood-based alignment and phylogeny estimation. Results on example protein families indicate that the inclusion of structural information can dramatically decrease uncertainty due to alignment, and as a result significantly stabilize reconstructed phylogenies. The current model has certain shortcomings and we briefly describe them here, along with possible extensions for future investigation.

Availability of structural data. Clearly the benefits of our approach are reliant on availability of experimental structural data for the proteins of interest. However, the number of known structures continues to grow rapidly as a result of high throughput structure determination efforts. Moreover, our results suggest that availability of structures for even a subset of the sequences can significantly stabilize the reconstructed tree, by informing rate parameters (through a hierarchical model) and decreasing uncertainty in key evolutionary distances that may drive topology uncertainty. It may also be possible to incorporate high-accuracy predicted structures, such as those based on homology modeling, for sequences of unknown structure.

Improving the structural evolution model. Intuitively, the inclusion of structure adds quantitative information (compared to the discrete characters of sequence mod-

els): the diffusion process penalizes large displacements of atoms in Euclidean 3-space. This helps identify homologous residues by favoring indel scenarios that best preserve the relative positions of residues present in both ancestor and descendant.

As mentioned in Section 2.2.1, the diffusion model of structural drift does not account for significant structural reorganization leading to discontinuous changes in fold. Descendant proteins are centered around ancestral structures, slowly losing fold information, without the ability to significantly reorganize into new structurally distinct stable folds. Interesting preliminary work by Herman J, Taylor W, Hein J (personal communication) provides a possible approach to modeling such large scale events using transitions between discrete states, and may be useful in combination with our model to provide a process that diffuses locally but has potential for discrete transitions.

In addition, the independent-site assumption in the OU process lacks certain realistic biophysical features such as excluded volume/repulsion and bond length constraints, which give rise to dependence among positions. The challenge in incorporating such effects is analytical tractability: for a general (e.g. repulsive) potential $U(X)$ the stationary distribution is known only up to a normalizing constant, but that constant is required to evaluate changes in model size due to the indel process, and moreover the conditional distribution is generally not analytically tractable. Incorporation of some site-dependence may be achieved by the addition of a between-site covariance matrix to the OU process, but the conditional and stationary distributions again become problematic when convolved with the Links indel process. The current independent-site OU process was chosen to provide simplicity and computational tractability, at the expense of some physical realism. However, since inference is performed conditional on observed structures, these limitations may be less important. Still, it is worth noting that a more realistic evolutionary process model for the structure might help provide additional information about evolution-

ary distance, since as mentioned in Section 2.3 we believe that in the current model structural information serves primarily to dramatically reduce alignment uncertainty, with information about t coming primarily from the sequence model.

Structure specific indel and substitution processes. Currently the model assumes constant insertion/deletion rates (λ and μ), structural diffusion rate (σ^2), and substitution matrix (Q) at all sites along the protein. A more realistic model would take advantage of the known structure, by allowing different rates according to secondary structure, solvent accessibility, location in an active site or binding site, etc. Although this seems straightforward, some care is required to preserve reversibility under indels. Structure-specific substitution matrices have been used successfully in sequence alignment and sequence-structure alignment (threading) and should improve the realism and information content of the model.

Dependence among sequence & structure. Currently the sequence and structural information are combined by assuming conditional independence of substitutions and structural deviations given the alignment. This is easily extended to incorporate dependence. The magnitude of dependence may be explored by estimating the conditional mean and variance of atom coordinate changes given sequence substitution from a database of hand-alignments.

Fully Bayesian structural phylogenetic tree reconstruction. Finally, the results in Section 2.3 relate to pairwise evolutionary distances and phylogenies constructed using neighbor-joining methods. In Chapter 3 the model is incorporated into a fully Bayesian simultaneous alignment-and-phylogeny estimation, as done for sequence evolution models Lunter et al. (2005b); Redelings and Suchard (2005). The incorporation of structural data resolves the significant uncertainty reported in simultaneous estimation models involving sequence only (Wong et al., 2008; Lunter et al., 2008), particularly when the phylogeny involves long time scales.

Despite some shortcomings, results reported in Section 2.3 with the current model

show significant improvements over sequence-only models commonly used in current practice. As such, the model provides an additional tool for phylogenetic studies, especially those involving distant relationships or rapidly changing sequences, by extending the applicability of evolutionary protein models to longer time scales.

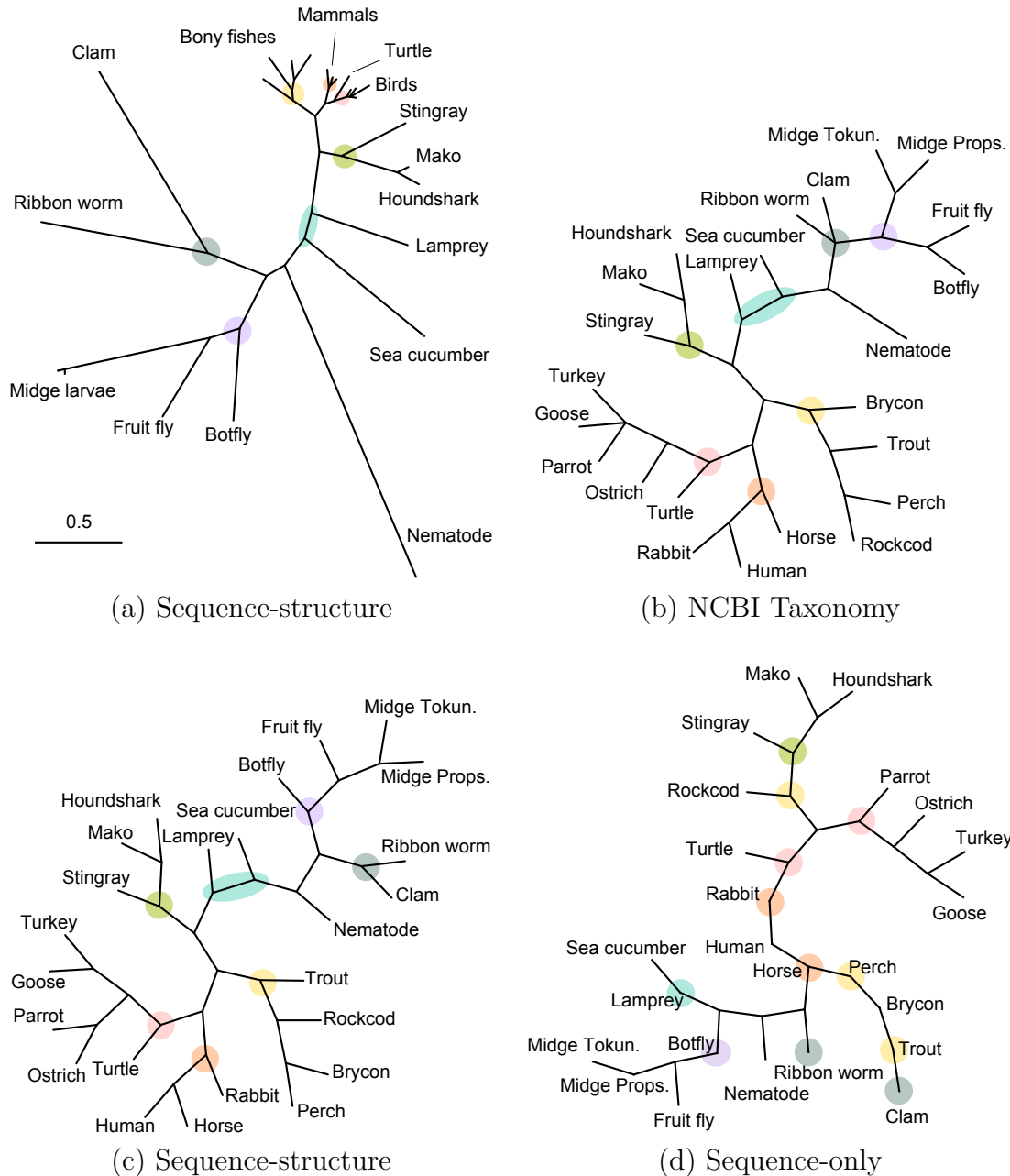
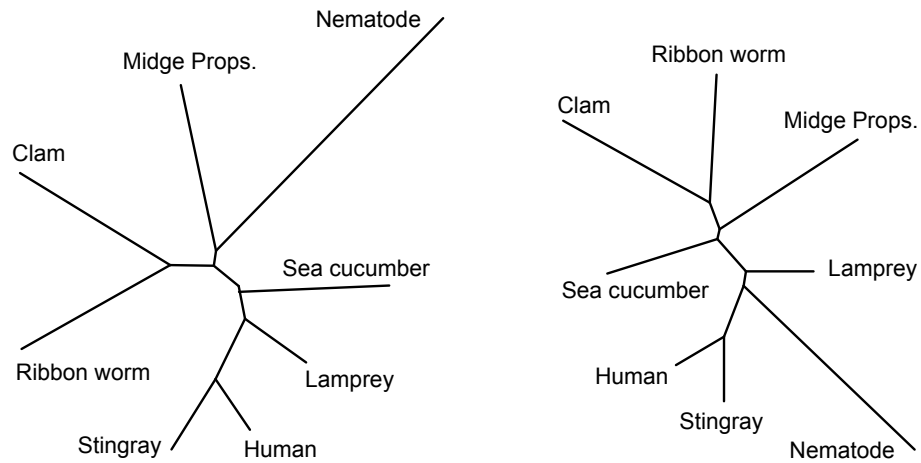


FIGURE 2.4: Phylogenies for a group of 24 globins (Table B.1, pairwise sequence identity 12-87%) obtained by different methods. Branch lengths in (b), (c), and (d) have been normalized for topology comparison. (a) Neighbor-joining tree using pairwise posterior mean evolutionary distances under sequence-structure model. (b) Accepted taxonomy (NCBI Taxonomy Database). (c) Topology of (a). Estimated topology closely matches NCBI taxonomy (b), with small differences. (d) Topology of neighbor-joining tree using pairwise posterior mean evolutionary distances under sequence-only model. Some groups are incorrectly separated and several species appear as zero-branch-length intermediate points. Figures created with TreeView (Page, 1996).



Sequence-structure

MAFFT

FIGURE 2.5: Estimated phylogenies for a subset of eight mutually distant globins (pairwise sequence identity 12-43%) . The sequence-structure model still closely matches the established NCBI taxonomy, while MAFFT begins to exhibit significant differences. Additionally, the MAFFT phylogeny has become more sensitive to parameter choice, while the sequence-structure model estimates appropriate parameters from the data.

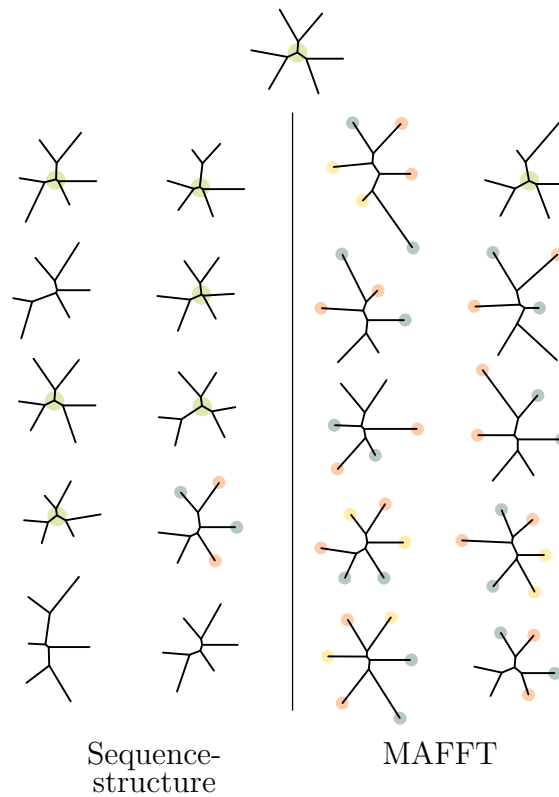


FIGURE 2.6: Phylogenies estimated from simulated highly divergent data sets (average pairwise sequence identity 13-17%). Top: True tree used to simulate data. Left: sequence-structure model estimates. Right: MAFFT estimates. Central green circle indicates correct topology, while red, blue and yellow identify correct pairs where mismatches are made. The sequence-structure model estimates the correct topology in 6 of 10 simulations, and preserves correct pairings of the proteins in all but one. MAFFT produces the correct topology in only one data set, and in all other cases matches pairs incorrectly.

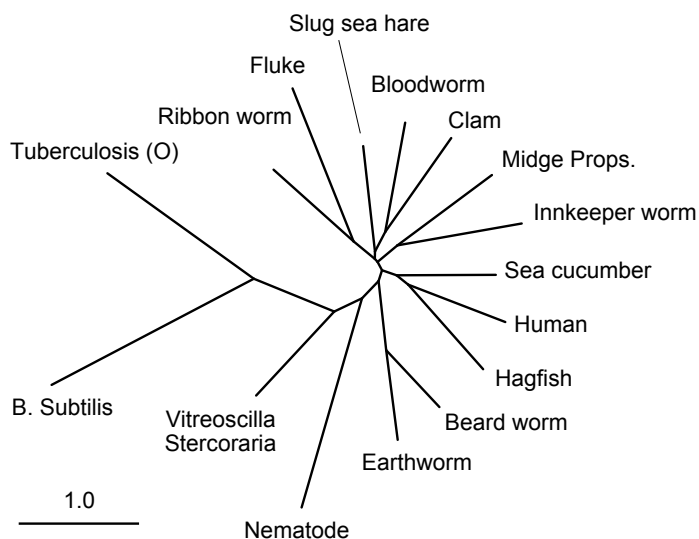


FIGURE 2.7: Phylogenetic tree estimated under the sequence-structure model on a highly divergent set of proteins (pairwise sequence identity 9-32%), from which MAFFT is unable to reconstruct a phylogeny.

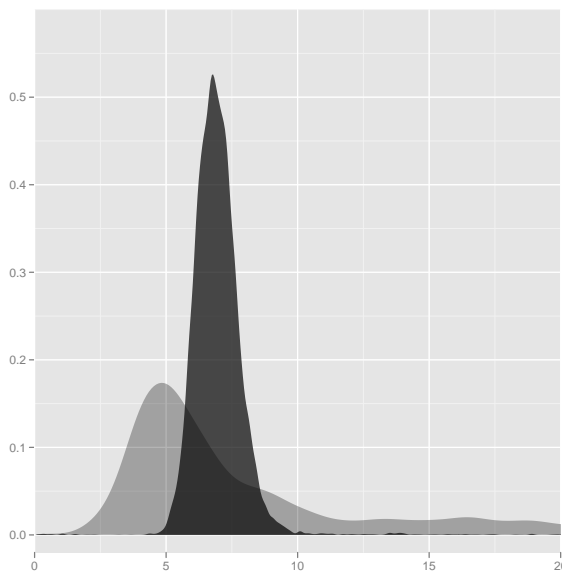


FIGURE 2.8: Posterior distributions of t (light) and $\sigma^2 t$ (dark) between phycocyanin β chain of red alga and human hemoglobin α obtained under sequence-structure model. At such large distances (7% sequence identity), sequence provides no information about t and only the product $\sigma^2 t$ may still be reliably estimated through structural information.

Joint Inference of Alignment and Phylogeny with Structure

Note

This chapter concerns extending the basic model described in Chapter 2 from pairwise comparison to a full Bayesian model for simultaneous tree estimation and alignment. The work presented in this chapter represents a collaborative effort with Joseph Herman, a PhD candidate in Statistics at Oxford University, and as such represents our shared contribution. It has been submitted for publication as “Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence structure” to *Molecular Biology and Evolution*. Although partitioning research contributions is rarely clear cut, in broad terms I was more responsible for model development and theory, while Joe handled more of the implementation in the software package StatAlign, as well as development of the section analyzing heterogeneity of structural evolution rates. However, the lines in the division of labor throughout the project were not strict. This document represents only those portions of the work in which I had a significant role. There were several other challenges

encountered in the development of this model where I did not deem my contribution significant enough to include here. Much of this additional work has been submitted as “StatAlign 3: Bayesian alignment and phylogenetics with protein structures” to Bioinformatics.

3.1 Introduction

In this chapter the pairwise structural model from Chapter 2 is extended to a fully Bayesian framework for estimation of phylogeny, alignment, and model parameters for sets of proteins structures (and sequences). There are several aspects of this problem that make it more challenging than a simple extension; the first of these is handling multiple alignments on trees and sampling new topologies.

Early methods for inferring alignments and phylogenetic trees were based on combinations of carefully tuned heuristic procedures, designed to optimise certain types of scoring metrics. Such methods have yielded many valuable insights; however, the results are often highly sensitive to user-specified parameters, and the focus on a single alignment and tree ignores much of the uncertainty associated with the analysis.

With the development of probabilistic models of molecular evolution, it has become possible to quantify this uncertainty in a statistically meaningful fashion. Bayesian methods for phylogenetic inference, such as MrBayes (Huelsenbeck and Ronquist, 2001) and BEAST (Drummond et al., 2013), address the issue of tree uncertainty by generating a distribution over phylogenies given a fixed alignment, although the choice of alignment may still heavily bias the resulting distribution on trees (Lake, 1991; Morrison and Ellis, 1997; Wong et al., 2008; Lunter et al., 2008; Blackburne and Whelan, 2013). A further set of methods have been developed to allow for joint sampling of alignments and trees, which allows this source of bias to be avoided (Redelings and Suchard, 2005; Lunter et al., 2005c; Miklós et al., 2008).

Such approaches are more computationally intensive, and analyses to date have been limited to tens rather than hundreds of sequences; however, these analyses are less prone to the misleading conclusions that can result from analysing a larger number of sequences under a biased model (Kumar et al., 2012).

3.1.1 Including structural information

However, as seen from examples in Chapter 2, for sequences that are highly divergent there may be a significant degree of uncertainty associated with the resulting alignments and trees. One way of addressing this issue is to combine multiple different types of data into a joint, or mixed, evolutionary model (Ronquist and Huelsenbeck, 2003). As well as offering a way of reducing uncertainty, this type of approach has the potential to lead to more robust and reliable results, since the resulting inference is based on multiple independent sources of information (cf. Kumar et al. (2012)).

For protein-coding genes, additional information regarding evolutionary relationships can be obtained from protein structures. Since tertiary structure is typically much more highly conserved than sequence, even over large evolutionary distances (Panchenko et al., 2005; Illergård et al., 2009), structural similarity is therefore a more reliable way to infer homology in the so-called *twilight zone* of low sequence identity, leading to more accurate alignments (Eidhammer et al., 2000; Hasegawa and Holm, 2009; Katoh and Standley, 2013), and potentially also phylogenies (Johnson et al., 1990; Bujnicki, 2000; Lundin et al., 2012).

In Chapter 2, I introduced a probabilistic evolutionary model describing the joint evolution of protein sequence and structure. In contrast to structurally-constrained sequence models that modulate substitution rates based on a fixed structure (Robinson et al., 2003; Rodrigue et al., 2005; Choi et al., 2007; Kleinman et al., 2010), this approach includes an explicit model for the evolution of structure, allowing for structural information to be used to help infer evolutionary distances.

In this work, the pairwise model is extended to a tree, and the utility of incorporating structural information into joint estimation of multiple alignments and phylogenies is explored. Since relatively little is known about structural evolutionary processes, we also introduce a model for heterogeneity in rates of structural evolution, which reduces the potential for conflict between structure- and sequence-based trees (Garau et al., 2005).

We also add a model of background (non-evolutionary) variability in structures, making use of prior information obtained from the x-ray crystallography experimental data, and drawing on aspects of other earlier probabilistic models of protein structure (Rodriguez and Schmidler, 2013; Green and Mardia, 2006; Schmidler, 2006; Green et al., 2010; Wang and Schmidler, 2013).

3.2 Probabilistic evolutionary models

In what follows, we deal with classes of probabilistic models on binary trees. Biologically these trees define phylogenetic relationships between a set of organisms; probabilistically, given the sequence at a particular parent vertex, evolution along each of its child branches is assumed to proceed independently.

3.2.1 Sequence and structure data

We consider a sequence evolving on a tree, Υ , with vertices \mathcal{V}_Υ and edges \mathcal{E}_Υ , according to an evolutionary model with parameters (Φ, Λ, Θ) , which describe rates of substitution, insertion and deletion (*indel*) events, and structural evolution processes, respectively. Associated with the K tips of the tree is a set of K homologous sequences $\mathcal{S} = \{S^{(1)}, \dots, S^{(K)}\}$, with $S^{(k)}$ of length $L^{(k)}$, and corresponding three-dimensional structures, $\mathcal{C} = \{C^{(1)}, \dots, C^{(K)}\}$, where $C^{(k)}$ is an $L^{(k)} \times 3$ matrix containing the Euclidean coordinates of the C_α atoms of structure k . In order to make use of the tree structure to permit tractable inference, each of the internal

nodes of the tree is augmented with an associated sequence and structure, the corresponding sets denoted by $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{C}}$ respectively. The structural coordinates and characters associated with these internal sequences will eventually be marginalised out analytically.

3.2.2 Representation of a multiple alignment

A multiple alignment can be represented as a set of pairwise alignments along the branches of a tree, $\tilde{\mathcal{M}} = \{M^{(k,l)}\}$, with $(k,l) \in \mathcal{E}_T$. Each pairwise alignment, of length $L^{(k,l)} \leq L^{(k)}L^{(l)}$, can be thought of as a series of columns in a $2 \times L^{(k,l)}$ matrix, indicating homology between characters in $S^{(k)}$ and $S^{(l)}$, i.e. the parent and child sequences along the branch. Each such column can take one of three possible states:

$$M_i^{(k,l)} \in \left\{ \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} x \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ y \end{pmatrix} \right\} \quad (3.1)$$

where $x \in \{1, \dots, L^{(k)}\}$ and $y \in \{1, \dots, L^{(l)}\}$ indicate the index of the characters aligned in this particular column, and 0 indicates an insertion or deletion. We will also denote by $M^{(k)}$ the row corresponding to sequence k in $M^{(k,l)}$, with the zero elements removed, equal to the vector $(1, \dots, L^{(k)})$; one of the requirements for a valid set of alignments, $\tilde{\mathcal{M}}$, is that all the pairwise alignments should be consistent in the sense that the mapping $M^{(k,l)} \mapsto M^{(k)}$ is the same for all l . Another requirement is that $L^{(k)}$ be equal to the length of $S^{(k)}$ when k is a leaf node. The full alignment, $\tilde{\mathcal{M}}$, can be projected down to a *leaf alignment* between the sequences at the leaves of the tree, \mathcal{M} , expressed in the familiar tabular format. We omit further notational details here for brevity.

3.2.3 Joint model for sequence and structure

The first phylogenetic evolutionary models to be developed allowed only for substitution events, assuming the alignment of the sequences to be known and fixed (Kimura,

1980; Felsenstein, 1981). However, work over the last two decades has shown that probabilistic modelling of insertion and deletion (indel) events can yield valuable additional information regarding evolutionary processes (Löytynoja and Goldman, 2005; Dessimoz and Gil, 2010), partly due to the rarity of such events (Lunter et al., 2003; Westesson et al., 2012). In this work we build on these existing approaches, adding a probabilistic model of protein structure to yield a joint Bayesian model for substitutions, indels, and structural evolution on a tree.

For reasons of tractability, we focus attention on models where the joint posterior of the unknown parameters of interest, given the observed (leaf) and augmented (internal node) data, can be factored as the product of substitution and structural contributions, and a stochastic indel process:

$$p(\tilde{\mathcal{M}}, \Phi, \Theta, \Lambda, \Upsilon \mid \mathcal{S}, \tilde{\mathcal{S}}, \mathcal{C}, \tilde{\mathcal{C}}) \propto p(\Upsilon) \underbrace{p(\tilde{\mathcal{M}}, \Lambda \mid \Upsilon)}_{\text{indel}} \underbrace{p(\Phi, \Theta \mid \mathcal{S}, \tilde{\mathcal{S}}, \mathcal{C}, \tilde{\mathcal{C}}, \tilde{\mathcal{M}}, \Upsilon)}_{\text{substitution/structure}} \quad (3.2)$$

The above factorisation will generally only be possible for independent-site models of substitution and structural evolution; insertions and deletions can change neighbourhood relationships, such that substitution, structure and indel processes are in general not separable in neighbour-dependent models.

In this work we also make the further assumption of separability between the substitution and structural evolutionary processes, such that

$$p(\Phi, \Theta \mid \mathcal{S}, \tilde{\mathcal{S}}, \mathcal{C}, \tilde{\mathcal{C}}, \tilde{\mathcal{M}}, \Upsilon) = \underbrace{p(\Phi \mid \mathcal{S}, \tilde{\mathcal{S}}, \tilde{\mathcal{M}}, \Upsilon)}_{\text{substitution}} \underbrace{p(\Theta \mid \mathcal{C}, \tilde{\mathcal{C}}, \tilde{\mathcal{M}}, \Upsilon)}_{\text{structure}}$$

Although it is also possible to formulate independent-sites models where there is some degree of dependence between sequence and structure (for example by allowing

for Θ to be a function of the amino acid content for a particular site), we leave such developments for future work.

3.2.4 *Marginal posterior*

Ultimately we are interested in the marginal posterior distribution over alignments, trees and model parameters obtained by integrating over the unobserved internal node data

$$p(\tilde{\mathcal{M}}, \Upsilon, \Phi, \Theta, \Lambda \mid \mathcal{S}, \mathcal{C}) \propto p(\Upsilon) p(\tilde{\mathcal{M}}, \Lambda \mid \Upsilon) \\ \times p(\Phi) p(\mathcal{S} \mid \Phi, \tilde{\mathcal{M}}, \Upsilon) \times p(\Theta) p(\mathcal{C} \mid \Theta, \tilde{\mathcal{M}}, \Upsilon)$$

We focus on cases where the observed data likelihoods $p(\mathcal{S} \mid \Phi, \tilde{\mathcal{M}}, \Upsilon)$ and $p(\mathcal{C} \mid \Theta, \tilde{\mathcal{M}}, \Upsilon)$ can be computed exactly by analytical summation and integration over ancestral characters and coordinates. Although, with some simplifying assumptions, certain indel models also allow for analytical summation over internal node alignments (Thorne et al., 1991; Bouchard-Côté and Jordan, 2013), for many models of interest this is not possible, yielding a problem of exponential complexity (Lunter et al., 2005a), hence we focus on the general case of inference for the full alignment $\tilde{\mathcal{M}}$ rather than directly targeting the marginal posterior for the leaf alignment \mathcal{M} .

Beyond the factorisability in equation (3.2), the statistical alignment framework we present here is not dependent on particular model choices for substitution and indel processes, but we will briefly describe the specific choices used in this work for the purposes of illustrating how they combine with the structural model. We introduce the structural model in more detail in the subsequent section, but note here that one of the key features of the approach we will present is that it allows the integration over unknown ancestral structures to be carried out analytically, greatly increasing the tractability of the resulting model.

3.2.5 Indel model

For a given tree, Υ , the contribution to the posterior for $\tilde{\mathcal{M}}$ from the indel model can be factored over the branches of the tree

$$\begin{aligned}
 p(\tilde{\mathcal{M}}, \Lambda \mid \Upsilon) &= \prod_{j \in \mathcal{V}_\Upsilon} p(M^{(j)}, \Lambda) \prod_{(k,l) \in \mathcal{E}_\Upsilon} \frac{p(M^{(k,l)}, \Lambda \mid \Upsilon)}{p(M^{(k)}, \Lambda)p(M^{(l)}, \Lambda)} \\
 &= p(M^{(\text{root})}, \Lambda) \times \frac{\prod_{(k,l) \in \mathcal{E}_\Upsilon} p(M^{(k,l)}, \Lambda \mid \Upsilon)}{\prod_{j \in \text{an}(\Upsilon)} p(M^{(j)}, \Lambda)^2} \tag{3.3}
 \end{aligned}$$

where \mathcal{V}_Υ and \mathcal{E}_Υ are the sets of vertices and, respectively, edges in the tree Υ , and $\text{an}(\Upsilon)$ is the set of ancestral (non-leaf) nodes of the tree. The vector $M^{(j)}$ is equal to one of the rows in the pairwise alignment $M^{(k,l)}$. The second line assumes that the tree is binary, which will be the case in all the examples we consider.

In this work we focus on the TKF92 model (Thorne et al., 1992) to generate the probability $p(\tilde{\mathcal{M}} \mid \Lambda, \Upsilon)$. This model is a birth/death process on fragments, each of which contains a contiguous run of characters (in our case amino acids). Fragments are inserted at rate λ and are deleted with rate μ ; the length of each fragment is geometrically distributed according to a probability r .

Each pair term in the numerator of equation (3.3) can be computed via dynamic programming using the pair-HMM representation of the indel model (Miklós et al., 2008), allowing the augmented likelihood to be computed in time linearly proportional to the number of branches in the tree, and the square of the average sequence length. The stationary probabilities for individual nodes are derived in Thorne et al. (1992), and take the form

$$p(M^{(k)} \mid \Lambda) \equiv p(L^{(k)} \mid \lambda, \mu, r) \tag{3.4}$$

$$= (1 - m) m(1 - r) [m(1 - r) + r]^{L^{(k)} - 1} \tag{3.5}$$

where $L^{(k)}$ represents the length of the k th sequence, equivalent to the length of $M^{(k)}$, and $m = \lambda/\mu$.

3.2.6 Substitution model

Under the independent-sites assumption, the substitution process is modelled as a collection of independent processes on individual amino acids. This allows the marginal likelihood of the leaf sequences, given a particular alignment $\tilde{\mathcal{M}}$, to be calculated using the familiar sum-product algorithm of Felsenstein (1981), yielding the quantity $p(\mathcal{S} | \Phi, \tilde{\mathcal{M}}, \Upsilon) = \sum_{\tilde{\mathcal{S}}} (\mathcal{S}, \tilde{\mathcal{S}} | \Phi, \tilde{\mathcal{M}}, \Upsilon)$.

The analyses conducted here employ the Dayhoff et al. (1978) matrix of amino acid substitution to parameterise Φ , although other choices are possible.

3.3 Structural drift model

There is empirical evidence of correlation between evolutionary time and structural divergence, although the exact nature of this relationship has remained the source of much speculation (Chothia and Lesk, 1986; Illergård et al., 2009). Chothia and Lesk (1986) famously observed an exponential relationship between structural divergence of core homologous residues as measured by RMSD and sequence divergence as measured by sequence identity. This original relationship was proposed based on a small dataset that was available at the time: 32 pairs of homologous proteins, as well as 5 instances of the same protein crystallised under different conditions. More recently, several authors have observed a linear relationship when sequence identity is converted to a measure of substitutions per site (Illergård et al., 2009), or if sequence identity and RMSD are replaced by approximate measures of significance (Wood and Pearson, 1999), although in some families a non-linear relationship may still be observed (Panchenko et al., 2005). In all cases structural divergence is observed to increase as sequence similarity decreases.

3.3.1 Model specification

We briefly reiterate the properties of the model introduced in Chapter 2. The structural model utilises a reversible diffusion process in 3D space, modelling fluctuations in the amino acid positions (represented by their C_α coordinates). As discussed earlier, independence between atoms is assumed to retain tractability. Structural evolution is modelled using an Ornstein-Uhlenbeck (OU) process on each C_α atom. Unlike Brownian motion, the OU process has a well-defined stationary distribution and so is reversible, allowing the combined structural, indel, and substitution processes to form a reversible model.

With $C_{ij}(t)$ representing the j th coordinate of the i th C_α at time t , the structural drift model describes the change in coordinates over time according to the following stochastic differential equation

$$dC_{ij}(t) = -\theta C_{ij}(t) dt + \sigma dB \quad (3.6)$$

where dB is standard Brownian motion, and θ is the rate at which a structure loses memory of its previous configuration, which we term the *structural drift rate*. The equilibrium distribution and conditional distributions of this process are Gaussians

$$C_{ij}(\infty) \sim \mathbf{N}(0, \tau) \quad (3.7)$$

$$C_{ij}(t) \mid C_{ij}(s) \sim \mathbf{N}(C_{ij}(s)e^{-\theta(t-s)}, \tau(1 - e^{-2\theta(t-s)})) \quad (3.8)$$

with the marginal variance $\tau = \sigma^2/(2\theta)$ proportional to the expected radius of gyration multiplied by the length of the structure. The quantity $\sigma^2/2$ can be thought of as a diffusion coefficient, with the expected mean square deviation after a time t approximately equal to $\sigma^2 t$ (see Section B.7). As such, we will refer to σ^2 as the *structural diffusivity*.

3.3.2 Structural diffusion on a tree

When extending this process to a set of structures related by a phylogeny, we must contend with an unknown ancestral structure at each internal node. Fortunately, the OU process allows for analytical integration over the unknown ancestral structure coordinates, such that the joint likelihood of the observed structures at the tips of the tree, $p(\mathcal{C} \mid \tilde{\mathcal{M}}, \Theta, \Upsilon)$, can be computed very efficiently. As discussed by Hansen and Martins (1996), for an OU process on a tree, the joint distribution for the data at the leaves is a multivariate Gaussian, in our case with a zero mean. The Markovian nature of the OU process means that the elements of the covariance matrix can be computed analytically, with $\Sigma_{kl}[\tau, \theta, \Upsilon] = \tau e^{-\theta d_{kl}(\Upsilon)}$, where $d_{kl}(\Upsilon)$ is the distance between leaves k and l along branches of Υ .

Denoting by $C_j^{(\mathcal{M}_i)}$ the length- $|\mathcal{M}_i|$ vector obtained by taking the j th coordinate of each observed (leaf) structure containing a character at the i th column, the marginal likelihood of the observed structures is then given by a product over the L columns of the alignment and the three spatial dimensions:

$$p(\mathcal{C} \mid \tilde{\mathcal{M}}, \Theta, \Upsilon) = \prod_{i=1}^L \prod_{j=1}^3 \mathbf{N}_{|\mathcal{M}_i|} \left(C_j^{(\mathcal{M}_i)} \mid \mathbf{0}, \Sigma_{\mathcal{M}_i}[\tau, \theta, \Upsilon] \right) \quad (3.9)$$

where $\Sigma_{\mathcal{M}_i}$ is a submatrix of Σ of dimension $|\mathcal{M}_i|$ formed by selecting the columns and rows corresponding to ungapped positions in the alignment column \mathcal{M}_i .

Figure 3.1 illustrates a set of samples on a tree drawn from the structural drift model with $\sigma^2 = 0.7\text{\AA}^2/\text{substitution per site}$, and $\tau = 70\text{\AA}^2$, evolving from structure 2DN2 (human haemoglobin) at the root.

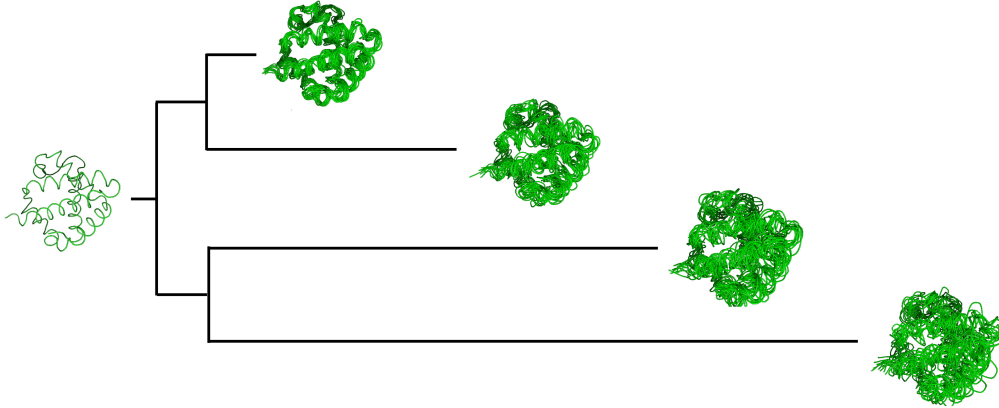


FIGURE 3.1: Ten samples from the structural drift model on a tree, with $\sigma^2 = 0.7\text{\AA}^2/\text{substitution per site}$, and $\tau = 70\text{\AA}^2$. With σ^2 set to zero we would see equal variability at each leaf, whereas the structural drift model proposes that structural divergence will be larger over greater evolutionary distances, in accordance with empirical observations.

3.3.3 Branch-specific structural drift rates

The model thus far assumes a constant structural diffusion coefficient, σ^2 , throughout the phylogenetic tree. This assumes that structures respond to sequence mutations in a homogeneous fashion, leading to an approximately linear relationship between evolutionary time and mean-square-deviation (*see Section B.7*). In order to allow for more general relationships between structural and sequence deviation, as well as reducing potential conflict between sequence- and structure-based trees, we relax this assumption and allow the structural diffusivity to vary over the tree. Following the approach of Thorne et al. (1998) and Aris-Brosou and Yang (2002) with regards to variable rates of sequence evolution, we allow σ^2 to vary by branch, which provides additional flexibility while allowing important properties such as infinite divisibility and reversibility to be maintained across the tree.

There are many ways in which this can be done; here we consider a model formulation that limits the number of additional parameters required. Let \mathcal{E}_Υ be the

set of branches of tree Υ , with $\{\sigma_k^2, \theta_k \mid k \in \mathcal{E}_\Upsilon\}$ the associated set of structural parameters. Allowing both σ_k^2 and θ_k to vary by branch does not preserve a common stationary distribution at each node of the tree, making the joint distribution difficult to specify. To solve this issue, we instead consider the alternative parameterisation $\tau_k = \sigma_k^2/(2\theta_k)$ with $\tau_k = \tau$ for all k , such that τ represents the equilibrium variance common to all nodes of the tree, while σ_k^2 is the local structural diffusivity, which is allowed to vary by branch. Since $\sigma_k^2 = 2\tau\theta_k$, the diffusivity of a branch is proportional to its structural drift rate, hence when describing heterogeneity across the tree, we will refer to these quantities interchangeably. The joint distribution of leaf nodes under this model remains simple and easy to obtain. The marginal distribution for each coordinate is then $\mathbf{N}(0, \tau)$ as before, while the covariance between coordinates of leaves k and l becomes

$$\Sigma_{kl}[\tau, \theta, \Upsilon] = \tau \exp \left\{ \sum_{m \in \pi(k,l|\Upsilon)} t_m(\Upsilon) \frac{\sigma_m^2}{2\tau} \right\} \quad (3.10)$$

where $\pi(k, l \mid \Upsilon)$ represents the set of branches lying on the unique shortest path from leaf k to leaf l , and $t_m(\Upsilon)$ is the length of branch m in tree Υ .

3.3.4 *Non-evolutionary sources of structural variability*

With sequence data, sequencing errors are relatively rare, such that any differences between sequences can generally be attributed to mutation events. However, for structural data, other sources of variability in the coordinates arise from factors such as flexibility of polypeptide chains, variable conformations, and measurement error (Gutin and Badretdinov, 1994; Grishin, 1997; Illergård et al., 2009). Moreover, this uncertainty may vary across the protein, with surface residues and loops exhibiting increased flexibility over buried core positions.

Information about this uncertainty for high-resolution structures solved by x-

ray diffraction is contained in crystallographic B -factors for each atomic coordinate. These values, reported by the crystallographer, are intended to summarise a combination of experimental uncertainty and thermal fluctuations, and are often strongly correlated with intrinsic structural flexibility measured by nuclear magnetic resonance and molecular dynamics simulations (Rueda et al., 2007). B -factors can be converted to units of coordinate uncertainty using approximate formulae such as the *diffraction-component precision index* (Cruickshank, 1960, 1999). This can be combined with additional assumptions (Schneider, 2000) to obtain a linear relationship between the B -factor and the standard deviation of the coordinates for each atom. We therefore model the variance for the i th atom of structure k (with B -factor B_{ki}) as

$$\epsilon_{ki} = \epsilon \frac{B_{ki}^2}{\left(\sum_j B_{kj}\right)^2} \quad (3.11)$$

where ϵ is a global scale parameter for background variance, to be estimated from the data. For the i th column, we compute the expected variance for the column as the average over the atoms aligned to the column

$$\epsilon_i = \frac{1}{|\mathcal{M}_i|} \sum_{k \in \mathcal{M}_i} \epsilon_{k\mathcal{M}_i} \quad (3.12)$$

Incorporating this into the structural drift model leads to a variance components model, with column i having covariance $\Sigma^{(i)} = \Sigma_{\mathcal{M}_i} + \epsilon_i I_{|\mathcal{M}_i|}$.

Uncorrelated structural perturbations (ϵ -only model)

In the limiting case as $\sigma_k^2, \theta_k \rightarrow 0$, keeping the ratio $\frac{\sigma_k^2}{2\theta_k} = \tau$ fixed, all structural deviation is explained via ϵ , and the marginal distribution of the observed data in the i th column is

$$C_{ij}^{(\mathcal{M}_i)} \mid \mathcal{M}, \tau, \epsilon, \Upsilon \sim \mathbf{N}_{|\mathcal{M}_i|}(0, \Sigma^{(i)}) \quad (3.13)$$

where $\Sigma_{kl}^{(i)} = \tau$ if $k \neq l$, and $\Sigma_{kk}^{(i)} = \tau + \epsilon_i$. This is similar to the non-evolutionary Bayesian structure alignment models described above (Wang and Schmidler, 2013), where structural perturbations are independent of evolutionary distance. In this limiting model, the structural likelihood does not depend on the tree nor on the evolutionary parameters, and structural information only indirectly affects the distribution over trees via the effect on the alignment.

3.4 Rotations and translations

As in Chapter 2 we have thus far avoided mention of rotating and translating coordinates. However, the coordinates of each structure are recorded with respect to an arbitrary reference frame, and the likelihood is not invariant to such transformations. This can be addressed without compromising the reversibility of the model by introduction of auxiliary rotation and translation random variables for each structure, as discussed in Chapter 2. Since the OU process is symmetric and hence invariant to rotations of the coordinate system, we can omit the rotation for an arbitrarily chosen reference protein; this reference protein still has an associated translation, such that the likelihood is independent of the choice of reference.

While the OU process specified on structural coordinates is reversible, it is not obvious if this still holds with the introduction of rotations and translations. That is, by introducing nuisance parameters with associated priors, it is unclear if the marginal probability of a group of proteins is the same regardless of the choice of the root. The posterior distribution should also be independent of the choice of (an unrotated) reference protein. We show that fixing both the translation and rotation of the reference protein does cause the posterior to be dependent on this choice; fixing only the rotation does not.

Given protein structures X and Y , we can define a procedure for calculating

the likelihood of X and Y from the OU model by fixing the coordinates of either X or Y and rotating/translating the other. Call the procedure $\bar{p}(X, Y)$ where the coordinates of X are fixed. According to this procedure, it should be clear that when X is near the origin and Y is not, $\bar{p}(X, Y) > \bar{p}(Y, X)$

$$\begin{aligned}
\bar{p}(X, Y|M, \sigma, \theta) &= \iint p(X, Y|M, \sigma, \theta, R, \eta)p(R)p(\eta)dRd\eta \\
&= \iint p(X|\sigma, \theta)p(YR + \mathbf{1}\eta|M, X, \sigma, \theta, R, \eta)dRd\eta \\
&= p(X|\sigma, \theta) \iint p(Y'_M|M, X_M, \sigma, \theta, R, \eta)p(Y'_M|M, \sigma, \theta, R, \eta)dRd\eta \\
\bar{p}(Y, X|M, \sigma, \theta) &= p(Y|\sigma, \theta) \iint p(X'_M|M, Y_M, \sigma, \theta, R, \eta)p(X'_M|M, \sigma, \theta, R, \eta)dRd\eta
\end{aligned}$$

Here $Y' = YR + \mathbf{1}\eta$. When comparing the last two lines above, we notice that clearly $p(X|\sigma, \theta) > p(Y|\sigma, \theta)$. The integral from $\bar{p}(X, Y)$ is also larger than its counterpart in $\bar{p}(Y, X)$ because the functions in the integrand are maximised over the same region in terms of η , while in the second integral this is not the case. Our initial approach was to first centre X (or Y) and then fix the coordinates, which clearly reduces the difference between $\bar{p}(X, Y)$ and $\bar{p}(Y, X)$, but in general cannot guarantee their equality.

Given X, Y, R , and η , we can instead define $X'' = X - \mathbf{1}\bar{c}$ and $Y'' = Y' - \mathbf{1}\bar{c}$, where $\bar{c} = \frac{\mathbf{1}^T X + \mathbf{1}^T Y'}{n_X + n_Y}$ is the centroid of X and Y' . Thus X and Y' are centred together for each rotation/translation pair.

$$\hat{p}(X, Y) = \iint p(X''|\sigma, \theta)p(Y''_M|M, X''_M, \sigma, \theta, R, \eta)p(Y''_M|M, \sigma, \theta, R, \eta)dRd\eta$$

We can now show that $\hat{p}(X, Y) = \hat{p}(Y, X)$. The key is that the OU process is symmetrical around the origin, and so the probability densities in the integrand are invariant to rotations. First observe the effect of counter-rotating both X'' and Y'' :

$$\begin{aligned}
X''R^T &= XR^T - \mathbf{1} \frac{\mathbf{1}^T XR^T + \mathbf{1}^T Y + \mathbf{1}^T \mathbf{1} \eta R^T}{n_X + n_Y} \\
&= (X - \mathbf{1} \eta)R^T + \frac{(n_X + n_Y) \mathbf{1} \eta}{n_X + n_Y} - \mathbf{1} \frac{\mathbf{1}^T (X - \mathbf{1} \eta)R^T + \mathbf{1}^T Y + n_Y \eta R^T}{n_X + n_Y} \\
&= (X - \mathbf{1} \eta)R^T - \mathbf{1} \frac{\mathbf{1}^T (X - \mathbf{1} \eta)R^T + \mathbf{1}^T Y}{n_X + n_Y} \\
&= (X - \mathbf{1} \eta)R^T - \mathbf{1} \hat{c} \\
&= X''' \\
Y''R^T &= Y - \mathbf{1} \hat{c} = Y'''
\end{aligned}$$

We arrive at X''' and Y''' , the centred values, by taking Y as the parent, rotating X by R^T and translating by $-\eta R^T$ (the inverse transformation that brings X and Y into exactly the same relative position to each other and the origin). Thus we have

$$\begin{aligned}
\hat{p}(X, Y) &= \iint p(X'')p(Y''|X'', R, \eta)dRd\eta \\
&= \iint p(X''R^T)p(Y''R^T|X''R^T, R, \eta)dRd\eta \\
&= \iint p(X''')p(Y'''|X''', R^T, -\eta R^T)dRd\eta \\
&= \iint p(Y''')p(X'''|Y''', R^T, -\eta R^T)dRd\eta \\
&= \hat{p}(Y, X)
\end{aligned}$$

The priors for R and η do not appear because we have assumed that they are both uniform. More generally, equality is maintained with symmetric priors such that $p(R) = p(R^T)$ and $p(\eta) = p(-\eta R^T)$, and it should be clear that this applies to an arbitrary number of structures related by the OU process. Thus by fixing only

the rotation of the the reference protein (and appropriate choice of priors), the model retains reversibility.

3.4.1 Shrinkage prior for branch-specific diffusivity

With a separate drift rate for each branch, there might be concern that the structural drift model could be overparameterised (Dutheil et al., 2012; Groussin et al., 2013). To address this possibility, we adopt a shrinkage-favouring mixture prior for the branch-specific σ_k^2 parameters:

$$\sigma_k^2 \mid \sigma_g^2, \nu \sim \gamma \delta(\sigma_k^2 - \sigma_g^2) + (1 - \gamma) \text{LogN}(\log \sigma_g^2, \nu) \quad (3.14)$$

with $\sigma_g^2 \sim \text{Gamma}(a_g, b_g)$ and $\nu \sim \text{Gamma}(a_\nu, b_\nu)$. This setup allows for pooling of information about σ_g^2 from all branches, while maintaining the flexibility of individual rates for each branch, as well as allowing for some degree of variable selection when appropriate. We set $a_g = 1, b_g = 2$, and $a_\nu = 1, b_\nu = 6$.

When $\gamma = 1$, all σ_k parameters are shrunk to the global mean, whereas $\gamma = 0$ yields the fully branch-specific model. For $0 < \gamma < 1$, the σ_k parameters that lie close to the global mean are shrunk strongly to σ_g . This additional shrinkage beyond the basic hierarchical prior is useful in larger trees where the internal branch drift parameters may have high uncertainty, particularly when the corresponding branches are very short.

For smaller trees we fix $\gamma = 0$; for larger trees γ is inferred from the data, using a $\text{Beta}(a_\gamma, b_\gamma)$ prior. When high levels of shrinkage are desired, we use $a_\gamma = 3.1$ and $b_\gamma = 1.1$, such that the prior favours shrinking all σ_k^2 to the global σ_g^2 with odds of approximately 3 : 1.

To carry out inference under this prior for γ , we employ a standard data augmentation scheme, with indicator variables z_k for inclusion of σ_k^2 . To improve mixing, we

can integrate out γ from this augmented model, yielding a Beta-Binomial prior for z

$$p(z \mid a_\gamma, b_\gamma) = {}^n\mathbb{C}_m \frac{B(a_\gamma + m, b_\gamma + n - m)}{B(a_\gamma, b_\gamma)}$$

where n is the number of branches in the tree, $m = \sum_k z_k$ is the number of local σ_k^2 parameters, and $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the Beta function.

3.5 Priors

3.5.1 Alignment and tree parameters

We assume a uniform prior on tree topologies, since we typically have no data-independent information about the topology. For branch lengths, we use a diffuse $\text{Exp}(0.01)$ prior. The prior on alignments is induced by the indel model parameters and their priors.

3.5.2 Substitution parameters and indel model parameters

In the analysis considered here use Dayhoff substitution rate matrix (Dayhoff et al., 1978). It is possible to estimate parameters of a more general substitution model during inference, but in the current analysis we keep these parameters fixed for reasons of computational efficiency.

The TKF92 model parameters are assigned the following prior specification

$$\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$$

$$\mu \sim \text{Gamma}(a_\mu, b_\mu)$$

$$r \sim \text{Beta}(a_r, b_r)$$

The hyperparameters are set to $a_\lambda = b_\lambda = a_\mu = b_\mu = 1$, resulting in $\text{Exp}(1)$ priors for λ and μ , and $a_r = b_r = 1$, resulting in a $\text{Unif}(0, 1)$ prior for r . Although λ and μ will typically have a value somewhat lower than 1, we favour the $\text{Exp}(1)$ prior over a prior more concentrated around zero in order to ensure that the effect of the prior be more similar across the range of probable values for λ and μ .

3.5.3 Priors for structural parameters

Rotations and translations are given uniform priors, as no rotation or translation is favoured *a priori*. Since the likelihood is not invariant to overall translations of the coordinates, the posterior remains proper despite the improper prior on translations. For the other structural parameters we use

$$\tau \sim \text{InvGamma}(a_\tau, b_\tau)$$

$$\epsilon \sim \text{Gamma}(a_\mu, b_\mu)$$

$$\sigma^2 \sim \text{Gamma}(a_\sigma, b_\sigma)$$

with hyperparameters $a_\tau = b_\tau = 0.001$, $a_\epsilon = b_\epsilon = 2$, and $a_\sigma = b_\sigma = 1$ yielding weakly informative priors reflecting our knowledge about the expected magnitude of structural fluctuations.

3.6 MCMC inference

Calculations of posterior distributions are performed by MCMC sampling. Since the joint posterior over alignments, topology, and parameters can be complicated, careful design of the MCMC algorithm is essential, and we have developed a number of specialised moves to increase the efficiency of convergence and mixing. Continuous parameters, i.e. (Θ, Φ, Λ) plus the branch lengths of the tree, are updated using random walk Metropolis updates after appropriate transformations, and tree topologies are proposed using a combination of stochastic nearest-neighbour interchanges and the LOCAL move of Larget and Simon (1999) with the acceptance ratio given in Holder et al. (2005). Alignments are resampled using a window-based progressive dynamic programming scheme to generate proposals, correcting the acceptance ratio by the ratio of likelihoods under the full model. The scheme is similar to the approach outlined in Miklós et al. (2008), augmented to include the structural likelihood. Although the rotations and translations would ideally be integrated out of the

model analytically, this typically leads to marginal likelihoods that are complicated functions of the unknown ancestral structures, even for uncorrelated Gaussian noise models (Goodall and Mardia, 1993). Hence we sample rotations and translations using the scheme described in Challis and Schmidler (2012).

3.6.1 Monitoring convergence

All MCMC simulations reported used four independent chains with randomised initial conditions. The overall likelihood and all scalar parameters were monitored for convergence using Gelman-Rubin potential scale reduction factors. For tree topologies, we monitored the stability of clade probabilities in the consensus tree; for alignments, we monitored convergence of alignment length and stabilisation of the maximum posterior decoding (MPD) alignment (Satija et al., 2009; Herman et al., 2013) and associated probabilities for each column.

3.7 Results and model comparison

To investigate the benefits of the structural model, we focused on datasets with highly divergent sequences, for which sequence-based analysis leaves significant uncertainty. We devote particular attention to the well-studied globins as a test case (*Table B.4*); previous attempts to reconstruct the evolutionary history for this family using sequence data have yielded trees with high uncertainty. Next we examine a set of cysteine proteinases (*Table B.5*), which further demonstrate the utility of structural information in reducing uncertainty in alignments and topologies, while also providing insight into patterns of structural divergence.

To assess the accuracy of parameter estimation (including topologies and alignments), data were simulated from the structural drift model, with $\sigma^2 = 0.7$, $\lambda = 0.03$, $\mu = 0.0305$, $r = 0.67$, and all B -factors equal to 1 for simplicity, using three different tree topologies, with 6, 8, and 10 leaves respectively. The structure at the root was

set to be equal to the haemoglobin 2DN2, and model parameters were chosen based upon typical values observed on test runs on small globin datasets. For each topology, branch lengths were multiplied by two different scale factors (1.0 and 2.0) in order to yield varying levels of divergence. Each parameter combination was simulated ten independent times, and results averaged over the ten repetitions.

For each dataset, we perform analysis using the sequence-only, ϵ -only and full structural drift model variants in order to assess the effect of including structural information.

3.7.1 Structural information improves alignments

For the simulated datasets the true multiple alignment is known, and we can measure the distance of the posterior alignment samples to this known alignment using the *column score* (proportion of correct columns), and the *sum-of-pairs score* (proportion of correct pairwise homology statements – see *Section B.7*). The alignment accuracy metrics are averaged over the ten repetitions for each tree. Under the sequence-only model alignment accuracy decreases markedly as branch lengths increase; in contrast, with the structural models, alignment accuracy remains high (*Figure 3.2*).

On the 5-globin and cysteine proteinase datasets, alignment accuracy was measured with respect to the alignment contained in the HOMSTRAD database (Mizuguchi et al., 1998), based 48 (globin) and 13 (cysteine proteinase) structures. In each case, the addition of structural information results in a consistent improvement in alignment accuracy and decreased variability (*Figure 3.2*), as with the simulated data.

3.7.2 Structure reduces topological uncertainty

The 5-globin dataset was chosen as a simple test case to explore the effect of structural information on topology uncertainty. The sequence-only model visits the most probable tree only 60.1% of the time, with 27.7% of the samples coming from a sec-

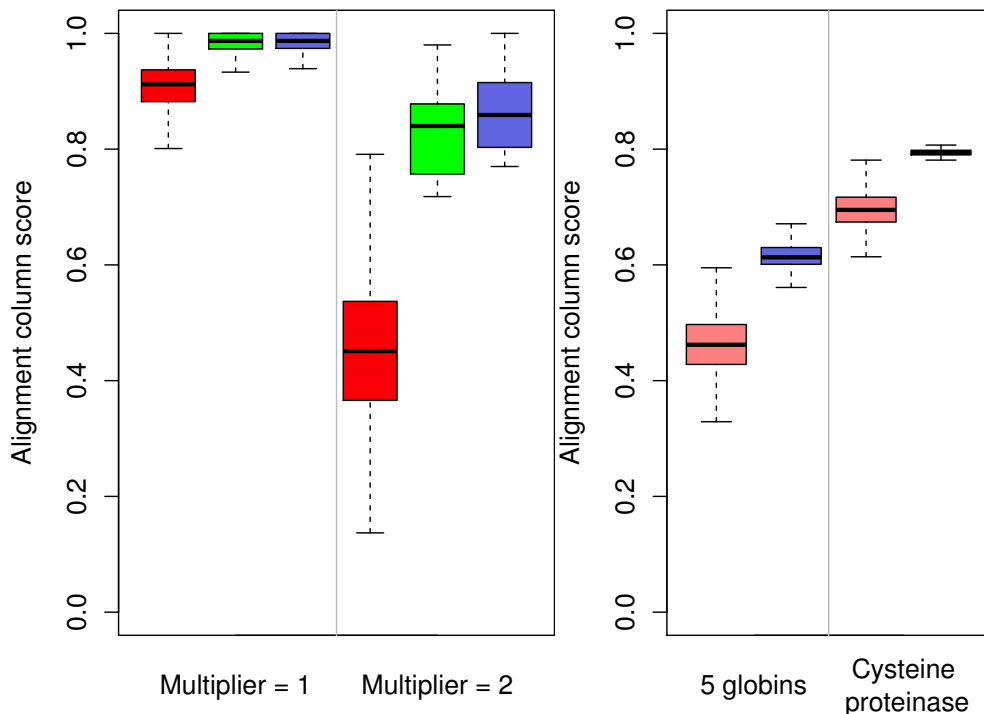


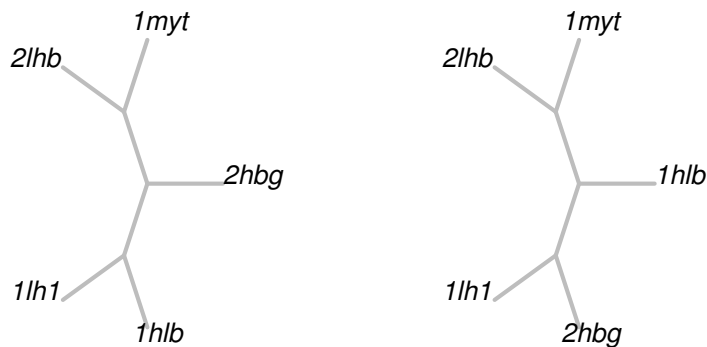
FIGURE 3.2: Alignment accuracy on simulated data (left two panels) for short branches (multiplier = 1) and long branches (multiplier = 2), and on the 5-globin and cysteine proteinase datasets (right panels). Shown are posterior distributions of distance to true alignment (simulated data) or HOMSTRAD alignment (globins and cysteine proteinases) obtained under the sequence-only model (red), and the structural model without (green) and with (blue) drift. In all cases structural alignments are more accurate than sequence-only, with a much lower spread of accuracy values. In many cases the drift model also offers an additional improvement in alignment accuracy. Simulated data results shown for ten realisations on an 8-taxon tree with $\sigma_k^2 = 0.7$ and $\epsilon = 0.5$, with branch lengths multiplied by the multiplier indicated. Similar results were seen with the sum-of-pairs alignment accuracy metric (not shown).

ond topology (Figure 3.3). In contrast, under both structural model variants there is virtually no uncertainty in the topology, with more than 99% of the samples coming from the most probable topology, placing 2hbg (*G. dibranchiata* haemoglobin) in between the other four structures. Results were generated from 100,000 samples, thinned from 20*m* iterations, after a 5*m* burnin. For sequence-only, on average around 165,000 topology switches were observed during the 20*m* iterations. For ϵ -only, around 4500 switches were observed, and for the drift model around 1400.

We also ran BAli-Phy (Suchard and Redelings, 2006) on this dataset, and the consensus tree yields a polytomy between 1lh1, 1h1b and 2hbg, indicating even higher posterior tree uncertainty under the BAli-Phy sequence-only evolutionary model (Figure B.1). These results clearly illustrate the improved concentration of the posterior under the structural model around the most likely topologies, with little additional computational cost: the three models required the same number of iterations to achieve convergence, with the runtime of the structural models around 1.3-1.5 times that of the sequence-only model.

Similar results are observed with the larger cysteine proteinase dataset (Figure 3.4). Again the structural consensus trees do not differ topologically from the sequence tree, and consensus branch lengths are very similar, but uncertain splits in the consensus tree are more highly resolved when structure is included.

As discussed earlier, structural information can reduce topology uncertainty in at least three ways: by increasing alignment accuracy, by reducing alignment uncertainty, and by providing direct information regarding the topology and branch lengths. In the above cases, a decrease in topology uncertainty is also observed with the ϵ -only model, suggesting that alignment inaccuracy and/or uncertainty is a principal cause of topology uncertainty in these examples. Nevertheless, additional reductions in alignment and topology uncertainty are also seen from adding the drift component to the model (Figures 3.2 and 3.4).



seq-only	0.606	0.277
ϵ -only	0.997	0.000
ϵ + drift	0.999	0.000

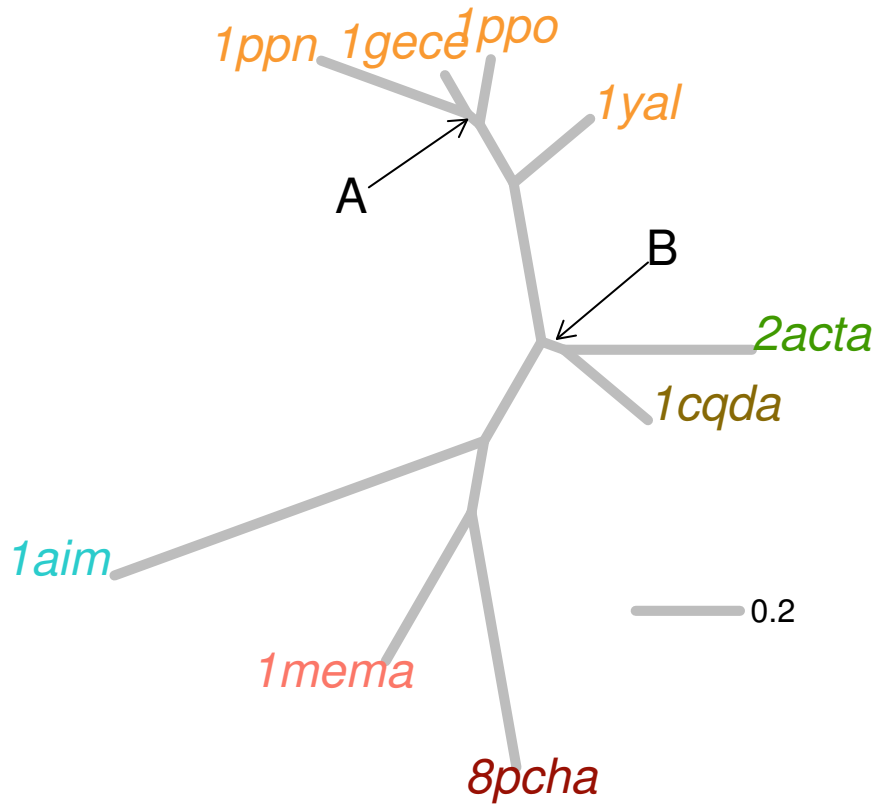
FIGURE 3.3: The two most frequently sampled tree topologies for the 5-globin data set under the sequence-only model, with posterior probabilities shown under sequence-only and structural models. Posterior probabilities were computed using the program `trees-consensus`, written by Benjamin Redelings.

3.7.3 Structural information reduces tree errors

For the simulated datasets where the true tree is known, we can also assess whether the structural model concentrates the tree posterior around the correct topology, using the Robinson-Foulds topology distance (Robinson and Foulds, 1981). For trees with smaller branch lengths, the sequence and structure models performed similarly, with the structural model only slightly more accurate. However, when branch lengths are doubled, the structural model not only reduces uncertainty, but also improves accuracy of the sampled topologies (*Figure 3.5*).

3.7.4 Structure helps select between alternative topologies

In cases where the majority of the tree is well resolved, the structural model often favours the same consensus tree as sequence. However, for trees with higher uncertainty, structure can also help to select between alternative hypotheses in regions that are difficult to resolve. Here we illustrate this by analysing a larger set of globins (*Table B.4*).



	A	B	C
seq-only	0.53	1.00	0.61
ϵ -only	0.81	0.97	0.96
ϵ + drift	0.97	1.00	1.00

FIGURE 3.4: For the cysteine proteinases the consensus topology was the same under all model variants. The labelled edges correspond to splits with significant uncertainty under the sequence-only model (the other three splits had posterior probability 1.00 in all cases). The table below the figure shows the posterior probability of each of these labelled splits under the different model variants.

The known set of vertebrate globin types was expanded relatively recently with the discovery of two additional globins: the neuroglobin (Burmester et al., 2000) and cytoglobin (Burmester et al., 2002). Neuroglobin tends to occur in neurons and endocrine cells, while cytoglobins appear in fibroblast-related cell types, and have been observed to be present in all vertebrates, suggesting an ancient split from other globin types. The function of both proteins is still somewhat unclear, although high levels of sequence conservation suggest a vital physiological function for cytoglobin (Hoffmann et al., 2012b).

Since these discoveries, there has been a surge of interest in establishing the likely evolutionary history of the four vertebrate globin types: haemoglobin (Hb), myoglobin (Mb), neuroglobin (Ngb), and cytoglobin (Cygb). All previous analyses have found Ngb to be the most distant outgroup, so we are primarily interested in the order in which the other globins split after diverging from the neuroglobins.

Initial phylogenetic studies of Cygb using maximum likelihood approaches suggested the topology (Ngb, (Hb, (Mb, Cygb))) (Burmester et al., 2002), although the support for this arrangement was found to be low. This topology may have initially appeared more plausible, since it requires O₂ transport to have evolved only once, along the branch to Hb. However, close homology was subsequently discovered between Cygb and the Hbs found in the jawless fishes known as cyclostomes (abbreviated as CycHbs). Accounting for this relationship requires either double evolution of O₂ transport function, or double loss of this functionality, as discussed by Hoffmann et al. (2010). Based on Bayesian phylogenetic analysis, the authors proposed the same phylogeny as Burmester et al. (2002), but with CycHb splitting from Cygb, i.e. (Ngb, (Hb, (Mb, (Cygb,CycHb))))), as shown in the top-left tree in Figure 3.6. Under this scenario, oxygen transport functionality is proposed to have developed independently in the cyclostome Cygb, the ancestor of the current CycHb, with the orthologues of the Mb and Hb genes subsequently lost (Hoffmann et al., 2010, 2012b;

Storz et al., 2013).

More recently Hoffmann et al. (2012a) conducted a Bayesian analysis on a larger dataset including globins from plants, and in this case reported a three-way split (Ngb, (Hb,Mb,(Cygb,CycHb))) (as shown in the bottom left tree in Figure 3.6, which contains a polytomy at the centre). Using a similar dataset including plant globins (but no CycHb), Ebner et al. (2010) were also unable to resolve this three-way split, reporting the same polytomy.

Here we compare the results obtained by Hoffmann et al. (2010, 2012a) with those from our structural model, as well as the sequence-only indel model. To do so, we construct smaller versions of the two datasets, containing one or two representatives from each of the clades of interest (*details in Table B.4*). The first dataset is the 8-globin set containing only Hb, Mb, Cygb, Ngb and CycHb, and the second dataset contains an additional four proteins, namely three plant globins and a recently-crystallised bacterial globin known as *Hell's gate*, which has been observed to show high structural homology with human neuroglobin (Teh et al., 2011; Vázquez-Limón et al., 2012).

Although the original analyses of Hoffmann et al. (2010, 2012a) used 68 and 110 sequences respectively, we obtain the same consensus tree from just 8 and 12 sequences using our sequence-only statistical alignment model (*see Figure 3.6*). However, as with the results of Hoffmann et al. (2012a), the addition of the plant globins appears to destabilise the consensus tree, favouring other topologies in the posterior.

Specifically, our sequence-only model shifts from having 94% posterior probability on the split (Cygb,CycHb), Mb | Hb in the 8-globin case, to favouring this less than 50% of the time when the plant globins are added. In the 12-globin case, the sequence-only model visits the following three topologies between the clades of interest:

1. (Mb,((Cygb,CycHb),Hb))
2. ((Cygb,CycHb),(Mb,Hb))
3. (Hb,((Cygb,CycHb),Mb))

with relative frequency 2:1:1. The third topology is the same as the consensus topology on the 8-globin set.

As noted by Hoffmann et al. (2012a), globins are relatively short proteins and thus limited in the information that can be provided about evolutionary history. Hence, there is good reason to believe that more accurate inference can be obtained by including other sources of information such as structure.

Indeed, as shown in Figure 3.7, the structural model favours topology 2 with almost 100% certainty regardless of whether the plants globins are added. This demonstrates that inference under the structural model is more robust to the choice of dataset. Moreover, we can see that the sequence-only model is shifting to increasingly favour the structural tree as more sequences are included, illustrating the fact that structures can contain additional evolutionary information beyond what can be obtained from sequences alone.

Both structural models favour (CycHb,Cygb) as the first split from the root. It should be emphasised that in the ϵ -only model, only the alignment is directly informed by structural information (rather than evolutionary distance), which reiterates the fact that the alignment can have a large impact on the resulting phylogenetic inference. When structural drift is also included in the model, the posterior probability of (CycHb,Cygb) diverging before the Mb-Hb split increases further (from 0.72 to 1.00), demonstrating that the structural drift model does indeed allow for additional structural information to be used in estimating tree topologies.

Table 3.1: Effective number of parameters, P_V , and model fit as measured by DIC for structural models with and without a drift component. Results averaged over four independent repetitions for each dataset.

	8-globins		12-globins		Cys proteinase	
	ϵ -only	ϵ + drift	ϵ -only	ϵ + drift	ϵ -only	ϵ + drift
P_V	150	140	258	229	226	213
DIC	16759	15959	25110	23743	18739	17075

3.7.5 Structural drift model improves fit

As shown by the results in the previous sections, structural information is able to reduce topology uncertainty, concentrating the topology distribution around the posterior mode, as well as offering improvements in alignment accuracy. These improvements are often greater with the drift model than the ϵ -only uncorrelated fluctuation model.

In order to measure whether the drift model also achieves a better model fit to the data, we make use of the *deviance information criterion* (DIC) (Spiegelhalter et al., 2002), given by $DIC = \mathbb{E}[D] + P_V$, where $D = -2\log L$ is the *deviance*, and $P_V = Var[D]/2$ is a measure of the effective number of parameters in the model (Gelman et al., 2003). Smaller values of DIC indicate a better model fit. The DIC measure is particularly suited to analysing the output of MCMC inference in hierarchical models when Bayes factors are not easily available (Spiegelhalter et al., 2002). It should be noted that the effective number of parameters includes a contribution from the alignment and the tree, such that lower posterior uncertainty in these parameters will reduce the effective dimensionality of the model.

As shown in Table 3.1, despite increasing the number of parameters, the addition of local drift rates for each branch reduces the overall uncertainty associated with the model, hence decreasing the effective number of parameters, P_V , and resulting in a substantial improvement in model fit, as measured by the DIC.

With complete shrinkage ($\gamma = 1$), the model retains only a single global σ_g^2 , decrease the effective number of parameters (on the 5-globin set this results in a reduction in average P_V from 148 to 140). However, the model fit generally suffers as a result (average DIC increases from 13,640 to 13,700 on the 5-globin dataset), and tends to result in trees with very different branch lengths from those obtained with sequence-only data. In contrast, the heterogeneous diffusivity model ($\gamma < 1$) results in a better model fit, and estimates branch lengths similar to those in the sequence-only trees. This suggests that branch-specific drift rates are indeed needed to explain the heterogeneity in the data. We examine this in more detail in the next section).

3.8 Heterogeneity in structural drift rates

3.8.1 Branch-specific drift rates result in better fit

With a separate drift rate for each branch, there might be concern that the structural drift model could be overparameterised (Dutheil et al., 2012; Groussin et al., 2013). To address this possibility, we also consider a shrinkage-favouring mixture prior for the branch-specific σ_k^2 parameters (*see Supplementary Section 3.5 for more details*). With medium levels of shrinkage, this allows borrowing of strength, and allows all the local diffusivity coefficients to be consistently estimated (Gelman-Rubin potential scale reduction factors close to 1).

With high shrinkage, all the local diffusivity coefficients are forced to be equal to the global σ_g^2 . Although shrinking all to the global σ_g^2 may decrease the effective number of parameters in the model (on the 5-globin set this results in a reduction in average P_V from 148 to 140), the model fit generally suffers as a result (on the 5-globin dataset we see an increase in average DIC from 13,640 to 13,700). More generally, forcing all σ_k^2 parameters to be equal tends to result in trees whose branch lengths are very different from those arising from the sequence-only setting, whereas

allowing for heterogeneity in diffusivity allows the branch lengths to remain similar to those in the sequence-only trees.

3.8.2 Specific examples of heterogeneity

On the 12-globin dataset, there are some striking examples of heterogeneity in the structural drift rates across the tree, consistent with the observations of Illergård et al. (2009). As shown in Figure 3.9, diffusivity is often higher along internal branches between proteins or clades that perform different functions. There is also strong evidence for positive selection (low σ^2) in six out of the 12 structures, corresponding to the haemoglobins and myoglobins, implying a high degree of selective pressure to preserve structure in these proteins. Given the importance of O₂ transport and storage function in vertebrates and cyclostomes

Equally notable are the highly increased rates of structural drift among the plant globins, particularly along the internal branch between the type-I non-symbiotic globin (nsGb) 2oif and the symbiotic leghaemoglobins (Lhb) 1bin and 1lh1. Although it was first hypothesised that Lhbs may have evolved from a bacterial ancestor, it is now thought that the Lhbs evolved from the nsGbs around 200mya, acquiring O₂ transport capability through the stabilisation of the open pentacoordinate haem configuration as opposed to the original, more stable hexacoordinate configuration (Landsmann et al., 1986; Vinogradov et al., 2005; Garrocho-Villegas et al., 2007; Hoy et al., 2007).

Although our structurally-based results support this same topology, there is a noticeable acceleration in the rate of structural evolution between the nsGbs and Lhbs. Previous studies have also uncovered a high rate of sequence variation in Lhbs than type-I nsHbs during the evolution of land plants, suggesting that different types of evolutionary pressures may have been involved along these two separate lineages (Vázquez-Limón et al., 2012). Since the purpose of O₂ transport functionality in

Lhbs is to sustain the symbiotic bacteria living in the root nodules of leguminous plants, it is conceivable that the this increased rate of structural divergence may be related to the emergence of symbiosis in legumes. Analysis of intermediate structures along this transition may help to uncover more of the mechanisms responsible for this major structural transition (Gopalasubramaniam et al., 2008).

With the cysteine and serine proteinases, the drift rates are generally much smaller, as might be expected given that function is largely conserved across most members of the datasets, with $\hat{\sigma}_g^2 = 0.05$ and 0.04 respectively. However, several of the branches have much larger diffusivity, for example for the porcine cathepsin (PDB code 8pch) we have $\hat{\sigma}_k^2 = 0.43$ (Gunçar et al., 1998) (*see Figure 3.10*).

3.8.3 Independence of drift rates and branch lengths

Since the default substitution model we use here posits a single substitution rate for the whole tree, one might wonder whether the variability in local drift rates merely reflects variability in rates of sequence evolution. However, under a global rate model of sequence evolution, local variations in rate will be encoded as longer or shorter branch lengths. Hence, if the drift rates are simply a proxy for variations in the rate of sequence evolution, and structure deviations are actually independent of evolutionary time, we would then expect to see a negative correlation between branch length and structural drift rate.

However, on all the datasets we examined there was essentially zero correlation between σ_k^2 and the branch length for all k , showing that these quantities contain separable sources of information (data not shown). This suggests that similar patterns of heterogeneity would be seen using a substitution model that allows for branch-specific substitution rates.

3.8.4 Patterns of structural divergence

The larger protein kinase dataset exhibits some further intriguing patterns of heterogeneity in the structural evolution rates, with several clades containing branches with very low as well as very high drift rates (Figures 3.11 and 3.12). In this case we can also see more clearly some different types of patterns in rates, which can be divided into the following categories of particular interest:

A Small σ^2 : high structural constraint (e.g. Hb and Mb)

B Large σ^2 : accelerated rate of structural drift (e.g. when developing new functionality, for example in symbiotic Lhb)

A+B Bifurcation where both branches are similar length, but one has a much smaller diffusivity coefficient

- suggests that there may have been a duplication event, and that the branch with the smaller diffusivity is closer to the ancestral structure, allowing the other structure to diverge since there is some redundancy (several examples, including α and β Hb); a form of *neofunctionalisation* (Hughes, 1994; Rastogi and Liberles, 2005).

A+A Bifurcation where both branches have very low diffusivity

- suggests strong selective pressure to preserve structure (e.g. α Hb in human versus fish); may be a form of *subfunctionalisation* (Rastogi and Liberles, 2005)

More generally, we also see patterns of the type $(\mathbf{A} + \mathbf{B})^n$, i.e. repeated bifurcations where one of the children of the pair has a very low diffusivity, and no descendants, for example in the top right of Figure 3.11, which may be a signature

of a series of duplication and neofunctionalisation events, whereby the ancestral protein retained its original function, and the new duplicate was either free, or perhaps under selective pressure to evolve a new functionality.

3.8.5 *Structural determinants of evolutionary drift rates*

There is theoretical and empirical evidence to suggest that more designable proteins (those with a higher contact density) may evolve faster on the sequence level, since destabilising mutations are more easily tolerated in such cases (England and Shakhnovich, 2003; Tiana et al., 2004; Bloom et al., 2006). Equivalently, in our framework these cases correspond to branches for which σ_k^2 is small, meaning that mutations to the sequence result in a smaller change to the structure along these branches.

On the other hand, Lukatsky et al. (2007) provided evidence to suggest that structurally similar proteins may exhibit a propensity to interact with each other; indeed, the globin family provides a particularly rich set of examples of oligomer formation, ranging from the familiar α - β Hb heterotetramer, to the large extracellular homo-oligomers found in insects (Terwilliger, 1992; Lamy et al., 1996). Although this may present a mechanism for the evolution of new binding partners Levy et al. (2008), it also poses a risk of unintentional homodimerisation. The need to avoid homodimer formation may give rise to what has been termed *negative design*, whereby a structure accumulates mutations that reduce its potential for self interaction Lukatsky et al. (2007). Such negative design may explain local accelerations in structural drift at certain branches in the tree, particularly after a duplication event, when the presence of two copies of a particular protein is likely to further increase the propensity for unwanted self-oligomerisation.

As discussed by Hughes (1994), one possible mechanism by which functional diversification to occur in enzyme families is to evolve new binding capabilities through

modulating the charge distribution on the surface of the protein. Among the cysteine proteinases, Hughes (1994) observed several regions of major charge difference among cathepsin B sequences, and devised a statistical test that suggested shifts of charge in certain regions of the structure were likely to have arisen as a result of selective pressure to diversify. The elevated rates of structural drift we observe in certain regions of the tree may be a signature of a similar mechanism of structural diversification.

In our case, we also see an elevated number of charge differences between the human cathepsin K, (PDB code 1mema) and the other sequences, including its nearest neighbour, 8pch (*see Figure 3.13*). When combined with the observation of an unusually high structural drift rate, this might suggest that charge modulation could play a role in the functional diversification of the cysteine kinase family.

3.8.6 Parameter inference

In addition to alignments and phylogenies, the model also provides the ability to estimate several scalar parameters of interest in the evolutionary process, such as indel rates and structural diffusivity coefficients.

On simulated data, the structural parameters are recovered to a high degree of accuracy, lying within the 95% highest posterior density interval in all cases, with the posterior median usually very close to the true value (*see Supplementary Figures B.3-B.5*). Importantly, we are able to clearly resolve the different contributions from ϵ and σ even without repeated observations at the leaves.

Table 3.2 shows posterior quantiles for ϵ and σ_g^2 (the global diffusivity) on two globin datasets (with 8 and 12 taxons), and the cysteine proteinase dataset, under the ϵ -only and structural drift models. The drift model estimates $\sigma_g^2 > 0$ even with ϵ in the model, indicating that there is always a time-dependent component to the structural variation. ϵ is a multiplicative scale factor (in units of \AA^2) for the

Table 3.2: Comparison of inference for global structural parameters on three datasets with and without drift, averaged over four repetitions from independent starting points. In the cysteine proteinase case, most of the variability is explained by baseline variance rather than evolutionary drift, although drift coefficients are significantly higher in certain regions of the tree (not shown).

		8-globins		12-globins		Cys proteinase	
		ϵ -only	ϵ + drift	ϵ -only	ϵ + drift	ϵ -only	ϵ + drift
$\hat{\epsilon}$	5%	3.23	0.762	5.16	1.54	1.03	0.239
	50%	3.53	0.902	5.78	1.76	1.09	0.275
	95%	3.81	1.046	6.37	1.99	1.14	0.310
$\hat{\sigma}_g^2$	5%	0	0.085	0	0.112	0	0.032
	50%	0	0.192	0	0.232	0	0.049
	95%	0	0.336	0	0.386	0	0.069

site-specific variance parameters, which in our case are proportional to normalised B -factors. Hence, $\epsilon = 1$ signifies that an atom with B -factor equal to the mean has baseline variance equal to 1\AA^2 . The parameter σ_g^2 has units of \AA^2 per substitution per site. For example, from the 12-globin set we expect structural drift to lead to an increase in mean square deviation of approximately 0.23\AA^2 per substitution per site (*see Table 3.2*), although there are also noticeable heterogeneities in drift rates across the tree.

In all cases Gelman-Rubin (GR) potential scale reduction factors were very close to 1, except for the ϵ -only model on the 12-globin dataset, since a single ϵ parameter struggles to explain the variability in this dataset, leading to slow convergence. In the cysteine proteinase case, although the global σ_g^2 is estimated to be very low (around 0.05), some branch-specific diffusivity coefficients are estimated to be substantially higher, hence there is still a substantial improvement in model fit using the drift model in this case (*Table 3.1*).

Table 3.3 also shows posterior distributions of the TKF92 parameters with and without structural information. Increasing the dataset from 8 to 12 sequences reduces the uncertainty associated with the parameter estimates in all cases, but a

Table 3.3: Posterior quantiles for alignment lengths (L), and TKF92 indel model parameters for globin datasets, aggregated from four independent MCMC chains in each case. All runs used a burn-in of $10m$ iterations, followed by a sampling period of $20m$ (sequence-only) and $40m$ (structural variants).

	8-globins		12-globins		
	seq-only	$\epsilon + \text{drift}$	seq-only	$\epsilon + \text{drift}$	
L	5%	167	177	174	184
	50%	173	182	184	188
	95%	183	186	194	194
R	5%	0.669	0.700	0.644	0.681
	50%	0.787	0.796	0.742	0.761
	95%	0.887	0.880	0.833	0.832
λ	5%	0.021	0.035	0.028	0.045
	50%	0.049	0.071	0.050	0.073
	95%	0.092	0.121	0.079	0.109
μ	5%	0.021	0.037	0.029	0.047
	50%	0.053	0.077	0.053	0.080
	95%	0.103	0.137	0.087	0.123

similar reduction in uncertainty in the alignment length and R is also observed when structural information included. Alignments are typically slightly longer with the structural model, and the indel rate parameters, λ and μ , are estimated slightly higher. This shows the estimation of these parameters can also be affected by alignment uncertainty, hence the inclusion of structural information also has the potential to improve estimates of insertion and deletion rates by improving alignment accuracy.

3.9 Discussion

3.9.1 Key conclusions

The main achievement of this work is the development of a tractable probabilistic model for joint evolution of sequences and structures on a phylogenetic tree. Our results demonstrate that inclusion of structural information reduces posterior uncertainty over alignments and topologies, improves alignment accuracy and reduces the number of tree errors, allowing for more reliable inference over larger evolutionary

distances. The structural model is also more robust to the particular dataset chosen for analysis, whereas sequence-only models can be highly sensitive to this choice.

Using this approach, we are able to provide structural insights into the evolutionary history of the globin family, whereas sequence-only methods encounter high uncertainty and sensitivity to choice of dataset, making it difficult to confidently characterise deep splits in the tree.

Structural information can reduce topology uncertainty both by reducing alignment uncertainty and by adding additionally information regarding divergence times for estimating topology and branch lengths. We observe that in some cases a large decrease in topology uncertainty can be obtained even with a model variant (the ϵ -only model) that affects the tree only via the alignment. This suggests that alignment inaccuracy and/or uncertainty can be a major cause of topology uncertainty, and further highlights the benefits of approaching alignment and topology inference in a joint framework, as we have done here.

3.9.2 Future work

As discussed, several modelling assumptions are made to ensure tractability of likelihood computations. These are likely to be reasonable for modelling local fluctuations around a particular fold, but may be less appropriate for modelling larger deviations. In particular, the assumption of independence between sites under the structural model becomes questionable when considering large displacements of secondary structure or other structural motifs. We are currently exploring extensions to allow for dependency between sites, although this is computationally very demanding, just as it is for sequence-based models.

The current model requires experimental structural data for all sequences included in the analysis. This is somewhat restrictive, and we are also developing extensions to the model to allow analyses when only a subset of the sequences have

structural data available. A number of other extensions to the model could be considered, including using mixture models in the diffusion process to increase flexibility of the model and potentially locate differing rates of evolution along the sequences, for example to identify structural features that are under strong selection.

Another modification that may improve model fit would be to allow the priors for each σ_k^2 to depend on the rate of the parent branch, as discussed by (Thorne et al., 1998; Aris-Brosou and Yang, 2002), to account for the fact that evolutionary rates are likely to diverge as a function of time. From a biophysical perspective, this might reflect the fact that the σ^2 parameters are related to the ability of a structure to accommodate sequence mutations, and this property is likely to be inherited to some extent from the parent structure.

Currently the model uses the magnitude of the crystallographic B -factor to estimate the expected standard deviation for each atom. In the cases we have examined, this relationship appears to hold very well (*e.g.* *Figure B.2*), but there may be cases where anisotropy and the presence of multiple conformers could lead to noticeable deviations from the expected behaviour (DePristo et al., 2004). By instead using the B -factor information to specify a prior distribution for each ϵ_{ki} , it would be possible to allow the data to override the B -factors where appropriate, although a larger number of structures may be needed to carry out parameter estimation in such a model.

Finally, as mentioned earlier, the structural model presented here is independent of the particular choice of indel model. By combining structural drift with other stochastic models of insertion and deletion, for example the recently developed Poisson indel model (Bouchard-Côté and Jordan, 2013), which allows for analytical marginalisation of indel histories as a result of some simplifying model assumptions, it may be possible to increase the size of datasets that can be analysed using this type of joint approach.

3.10 Availability

We have implemented the joint sequence-structure model as a plugin for the StatAlign software package (Novák et al., 2008), which can be downloaded, along with example datasets, from <http://statalign.github.io/>

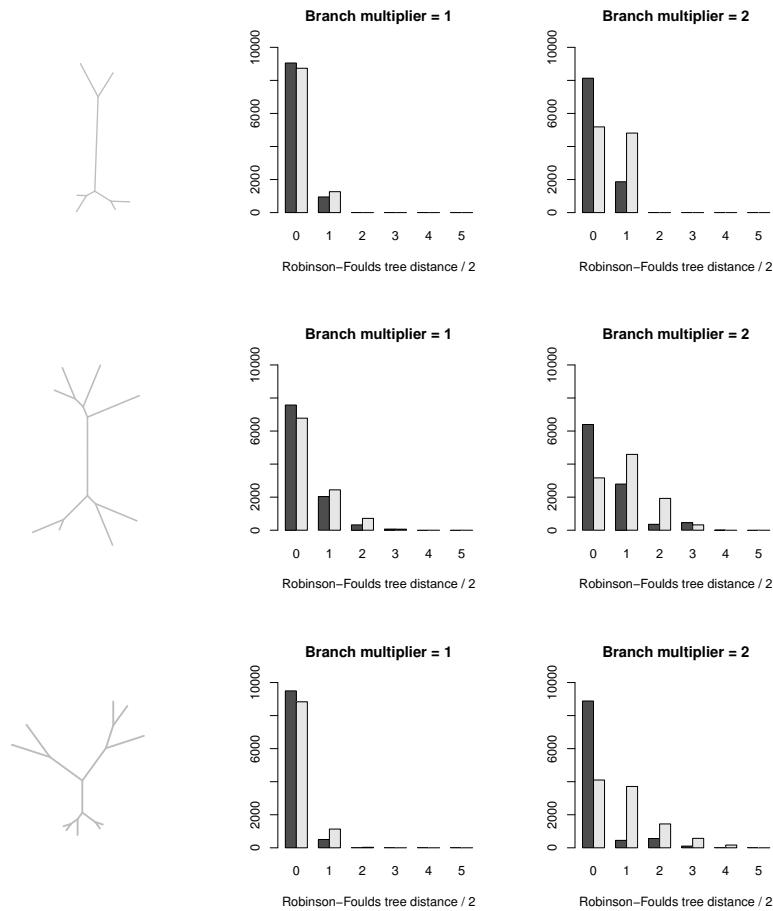


FIGURE 3.5: Posterior distribution of topology errors relative to the true tree for simulated data, analysed under the structural (black) and sequence-only (grey) models, as branch lengths are doubled (left to right). The inclusion of structural information allows the tree to be accurately inferred even for large evolutionary distances, whereas the trees inferred by the sequence-only model become much less accurate. Frequencies shown for the trees on the left, with 6 (top), 8 (middle), and 10 (bottom) leaves, aggregated from 10 independent samples from the model; the maximal half Robinson-Foulds distance for a tree with n leaves is $2(n - 3)$, i.e. 3, 5 and 7 for the three trees above.

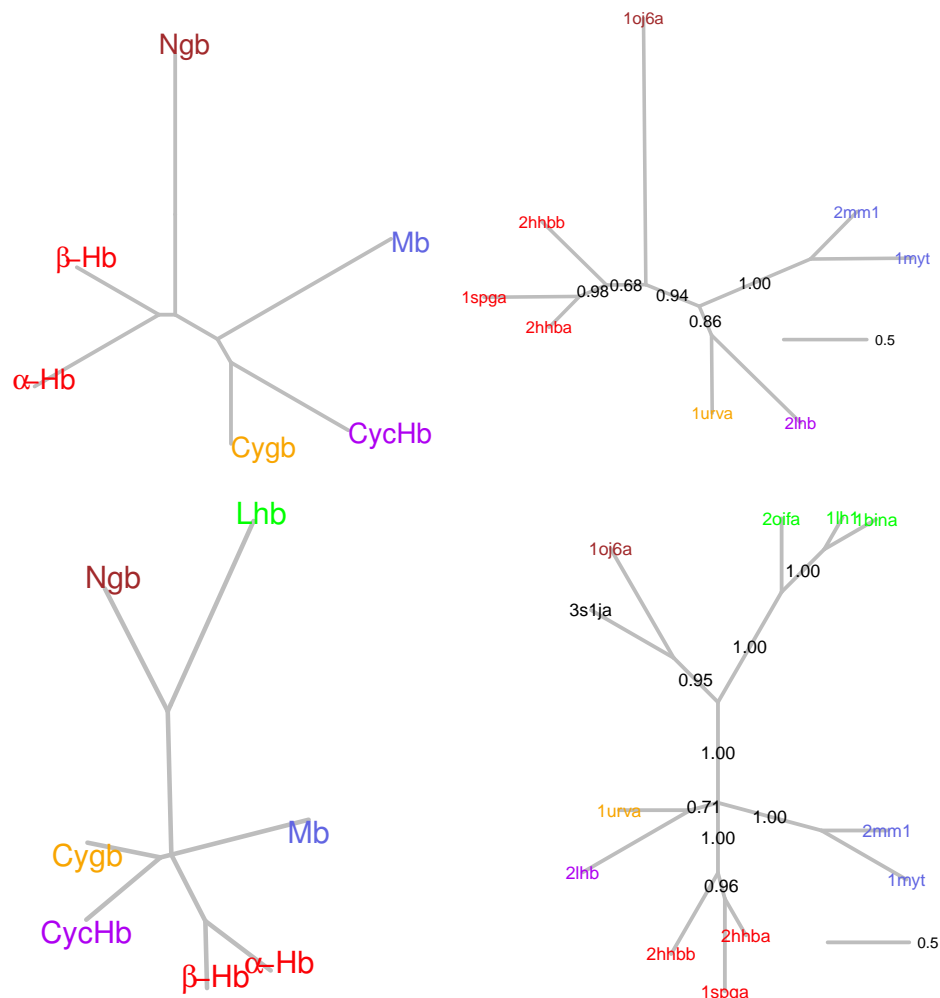


FIGURE 3.6: Consensus trees for globin datasets, taken from Hoffmann et al. (2010) and Hoffmann et al. (2012a) (top left and bottom left respectively), and inferred using the sequence-only evolutionary model of Miklós et al. (2008) (top right and bottom right). The bottom row features an augmented dataset containing plant globins, as well as a bacterial globin in our analysis. In both cases we obtain the same consensus tree as Hoffmann *et al.*, including the four-way polytomy in the 12-globin case.

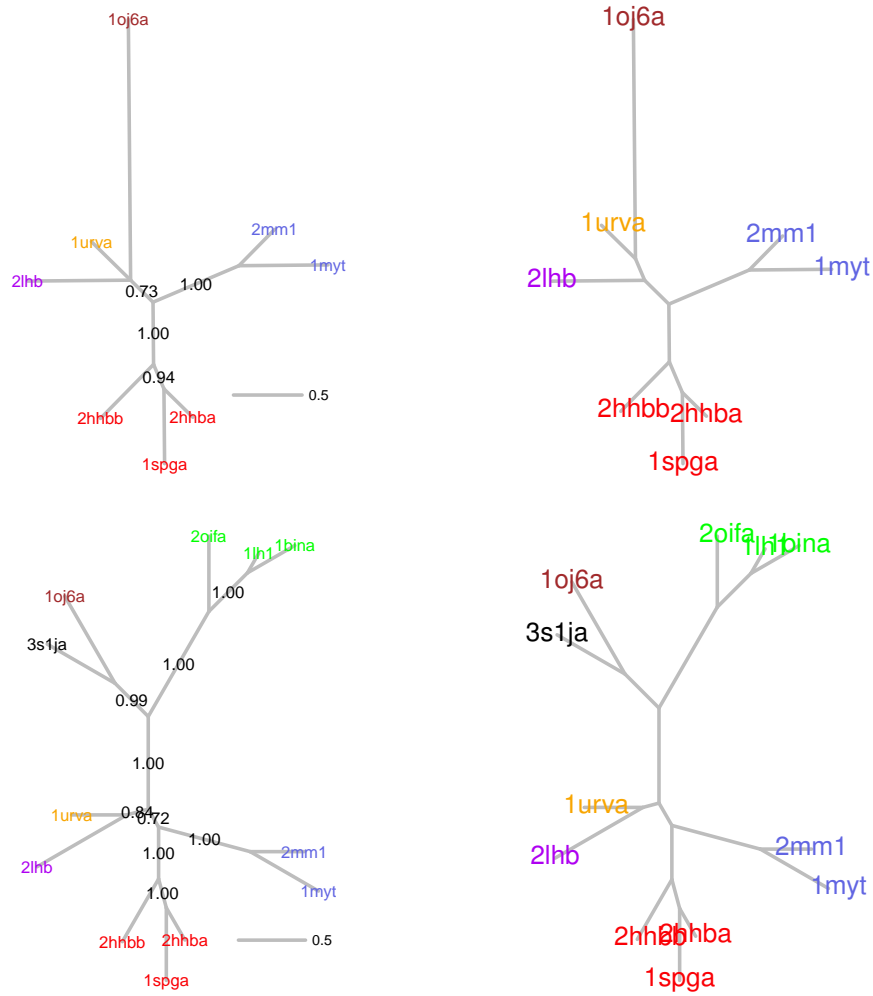


FIGURE 3.7: The structurally derived trees have very low uncertainty, and the order of the splits of interest is unchanged by the inclusion of additional sequences. Consensus trees derived under the ϵ -only model (top left and bottom left), and the full structural drift model (top right and bottom right).

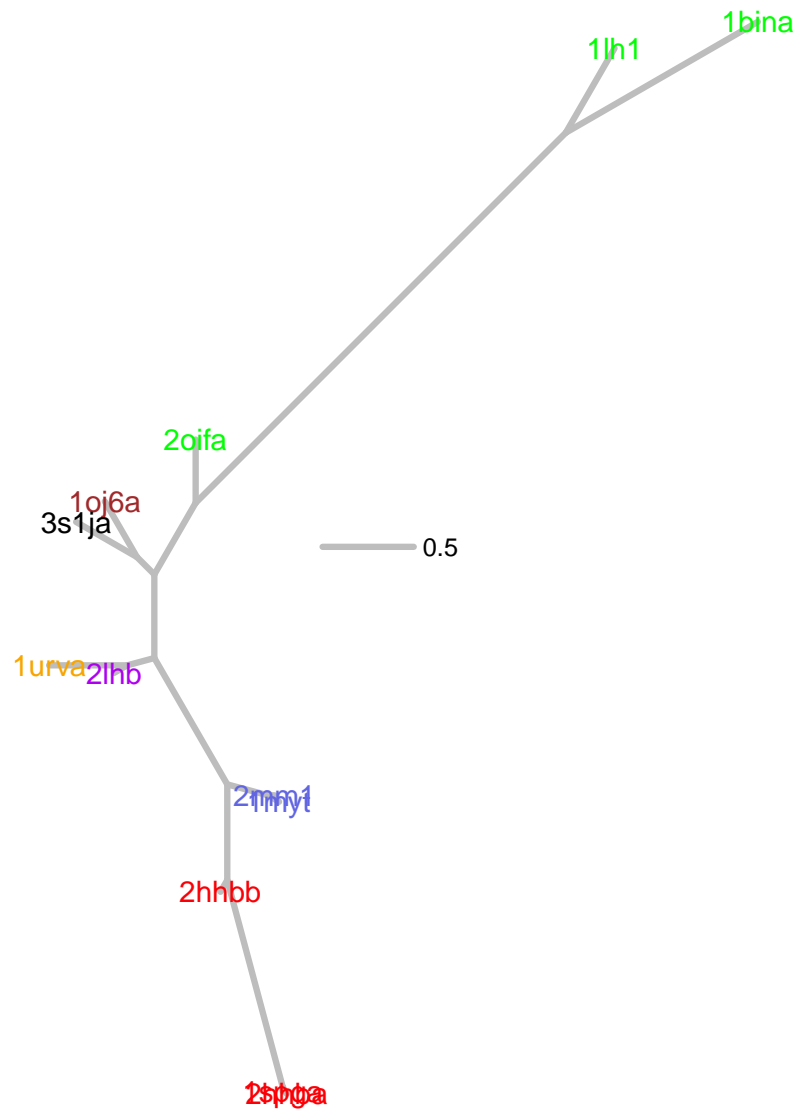


FIGURE 3.8: Consensus tree with branches scaled by local σ_k^2 parameters for the 12-globin dataset,

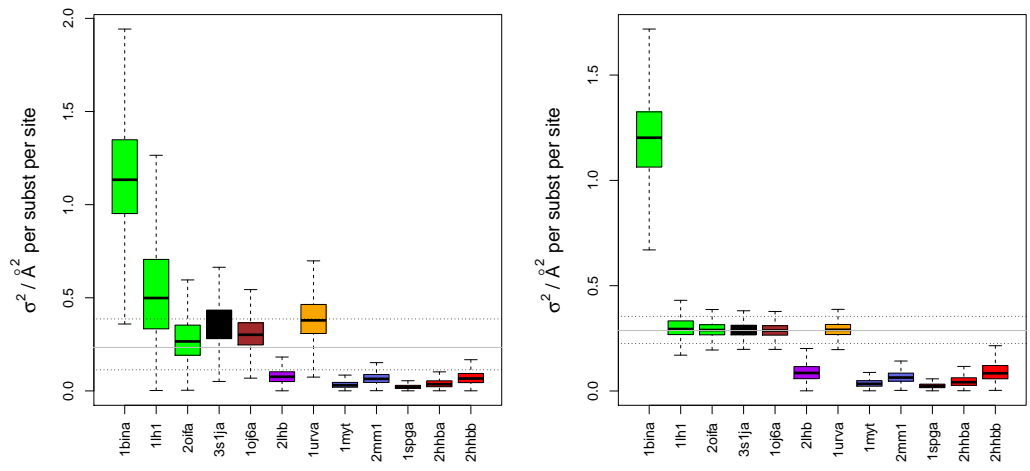


FIGURE 3.9: Distributions for σ_k^2 for leaf branches in the 12-globin dataset, estimated with low (top right) and high (bottom right) shrinkage to the global σ_k^2 , using the shrinkage mixture prior described in Supplementary Section 3.5.

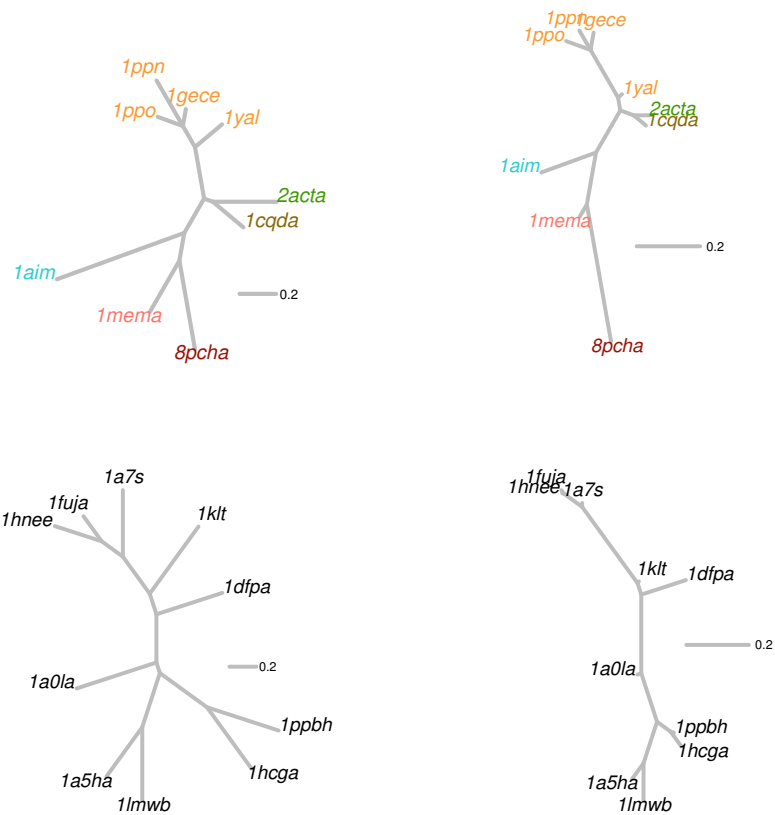


FIGURE 3.10: The consensus tree for the cysteine proteinase and serine proteinase datasets (top, and bottom respectively), with branches scaled according to mean branch length (left), and mean σ_k^2 (right), showing heterogeneity in structural diffusivity coefficients across the tree.

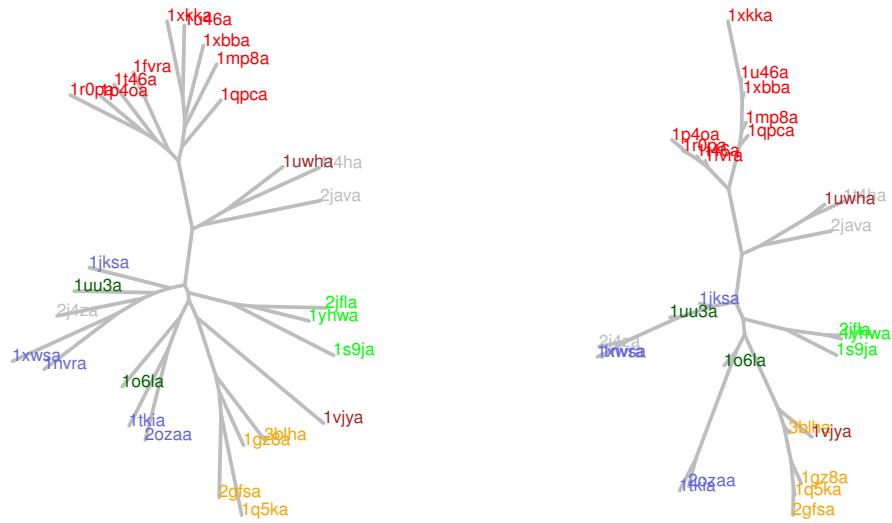


FIGURE 3.11: The consensus tree for the full protein kinase set, with branches scaled according to mean branch length (left), and mean σ_k (right), taken across all trees containing a clade that appears in the consensus (σ_k rather than σ_k^2 used for plotting the tree for ease of visualisation). Taxons are colour-coded according to the scheme in Manning et al. (2002): red = tyrosine kinases, blue = calmodulin-dependent kinases, light green = yeast sterile kinases, dark green = (PKA,PKC,PKG), orange = (CDK,MAPK,GSK3,CLK), brown = tyrosine kinase-like, grey = uncategorised.

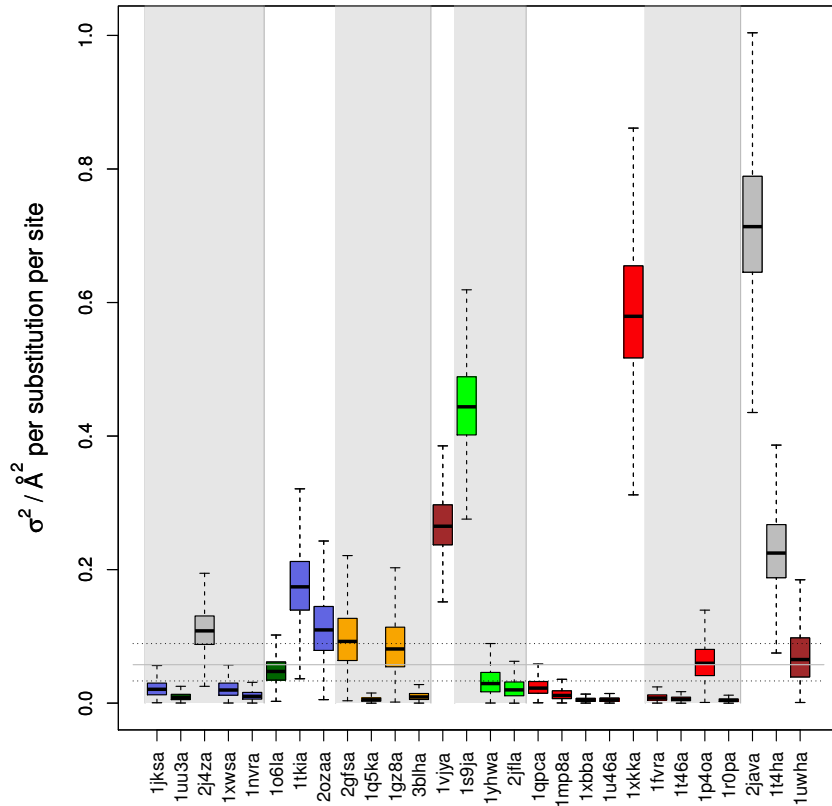


FIGURE 3.12: Summary of distributions for diffusivity coefficients at the leaf branches for the tree given in Figure 3.11. Taxons are colour-coded according to the scheme in Manning et al. (2002): red = tyrosine kinases, blue = calmodulin-dependent kinases, light green = yeast sterile kinases, dark green = (PKA,PKC,PKG), orange = (CDK,MAPK,GSK3,CLK), brown = tyrosine kinase-like, grey = uncategorised. Grey boxes in the background indicate boundaries between clades based on the consensus tree. Median and highest posterior density interval for the global σ_g^2 is shown by the dotted lines running across the boxplot.

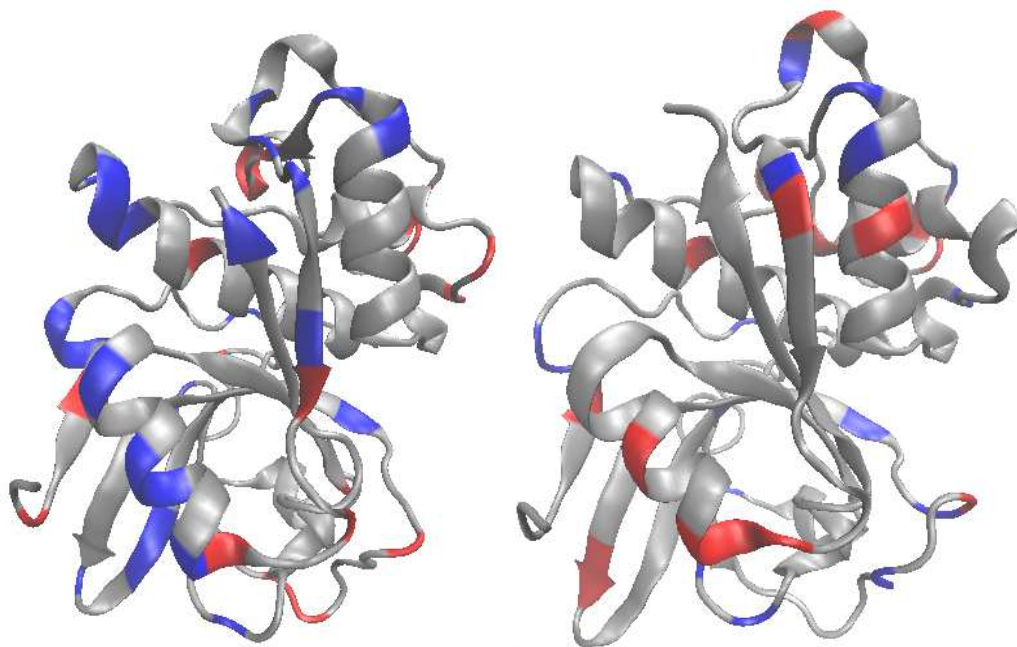


FIGURE 3.13: Structures for 1mem (left) and 8pch (right), with charged residues highlighted in red (positive) and blue (negative), showing a large number of differences between the two proteins.

The Cutoff Phenomenon and Piecewise Priors in Models of Biological Sequence Evolution

4.1 Introduction

Biopolymer sequences are one of the most commonly used sources of information for reconstruction of evolutionary relationships between extant organisms. It is well known, however, that sequences are limited in the amount of information that can be extracted for distant relationships, and there have been some indications of the rapid transition characteristic of the “cutoff phenomenon.” (defined in Section 4.2) In this chapter I draw attention to the existence of a probabilistic cutoff in popular models of biological sequence evolution. I show that current default priors for evolutionary distance and phylogenetic branch lengths are inadequate for inference when a cutoff is present and suggest a class of piecewise priors better suited for inference in this setting.

The “cutoff phenomenon” (Aldous, 1983; Diaconis and Shahshahani, 1981) is used to describe a behavior of Markov chains which remain far from equilibrium for some time before converging quickly to their limiting distribution. (A formal definition is

given in Section 4.2). In terms of evolutionary inference, models behaving in this way retain a large amount of information about evolutionary distances for some period of time, before rapidly losing inferential ability. The cutoff phenomenon has been demonstrated for many Markov chains; a (partial) list of known cutoffs is given in Saloff-Coste (2004).

Although the cutoff phenomenon is not fully understood, a common characteristic of chains exhibiting this behavior is a multiplicity of the second eigenvalue of the Markov transition matrix (Diaconis, 1996). As the kernel of an iid product chain of length n has multiplicity n of the second eigenvalue, these models might be expected to exhibit cutoff behavior. For example, as pointed out by Mossel and Steel, the binary purine/pyrimidine process forms a random walk on the hypercube, which has been shown by Diaconis and Shahshahani (1987) to exhibit a cutoff at time $\frac{1}{4}n\log(n)$. Diaconis et al. (1990) furthered understanding of the binary random walk to show that the same cutoff holds with the analogous (Cavender-Farris-Neyman) continuous-time model (Cavender, 1978; Farris, 1973; Neyman, 1971), and explicitly derived the asymptotic behavior of the total variation distance (defined below) in the cutoff region. I show an analogous asymptotic result for *any* fully symmetric evolution on an alphabet of m characters (this includes, for example, the well-known Jukes-Cantor nucleotide model (Jukes and Cantor, 1969)), that fully characterizes the transition to equilibrium in the cutoff region.

Finally, Ycart (1999) showed that the cutoff applied to iid reversible Markov chains, and Barrera et al. (2006) extended the result to show that the cutoff occurs for a large class of products of independent processes, which includes all commonly used models of sequence evolution. This applies not only to constant-rate evolutionary models such as the Dayhoff and JTT models (Dayhoff et al., 1978; Jones et al., 1992), but also to $+\Gamma$ (Yang, 1994) and other rate heterogeneous varieties.

Rost (1999) noted a transition from the “safe zone” of sequence alignment into

the “twilight zone,” marked by an explosion of false negatives in homology detection. Other authors offer results related to phase transitions in the length of sequences required to recover phylogenetic trees and ancestral sequences (Mossel, 2003, 2011), or offer a bound on maximal posterior inference on binary sequences with exponential priors (Mihaescu and Steel, 2010). However, the presence of a formal cutoff has gone largely unnoticed by the evolutionary biology community, due to the purely probabilistic nature of the result (Barrera et al., 2006; Ycart, 1999). Also, the cutoff is both abstract and asymptotic in nature, making it difficult to see exactly how this phenomenon will impact evolutionary inference in practical settings. I provide guidelines for understanding and locating the cutoff for specific models of sequence evolution and for finite sequence lengths.

The cutoff of a model provides a particular range over which it is effective. In terms of an evolutionary process beginning from a fixed sequence, before the cutoff is reached the distribution over sequences is far from the equilibrium distribution of the process, and so inference may be performed reliably. Once the cutoff region is reached, the process quickly transitions into the equilibrium distribution, after which little information can be gleaned by comparing the two sequences. Understanding the location of the cutoff for a model provides a reference point for prior formulation and identifies regions over which to expect uncertainty to emerge in the posterior. The cutoff is an asymptotic result, technically defined for infinite families of processes. Behavior of these processes begins to approach the asymptotic result as n increases (here, n = sequence length). I show that cutoff behavior is seen for relatively short sequences ($n = 100$), making it applicable to nearly all biological sequences of interest. I describe attributes of the cutoff and how to locate it generally from the rate matrix of an evolutionary model and specifically for a particular sequence under the model (using the Hellinger distance). In general, there is a loss of inferential ability occurring rapidly at the beginning of the cutoff region.

A common theme in statistical inference is the tradeoff between bias and variance, which in the Bayesian framework is controlled by the prior. The existence of the cutoff implies that sequences retain a high degree of information about evolutionary distances over a particular timeframe, after which information is rapidly lost. Recent studies on the influence of branch length priors on Bayesian inference of phylogeny (Brown et al., 2010; Kolaczkowski and Thornton, 2007; Ekman and Blaalid, 2011; Yang and Rannala, 2005) note trees with unreasonably long branches. Here I demonstrate that the underlying cause of these behaviors is the cutoff phenomenon, and provide a more appropriate prior for Bayesian inference with a cutoff. Where there is sufficient information to determine distances, the prior should apply very little shrinkage, but beyond the cutoff more extreme shrinkage is necessary to avoid unreasonably long branch lengths. The default exponential prior results in constant shrinkage over all distances, and so is unable to provide appropriate inference in both regimes. I introduce piecewise priors which allow a changepoint in shrinkage behavior.

4.2 Preliminaries

4.2.1 Sequence Evolution Models

A single-position Markov chain is defined by an instantaneous rate matrix Q , with transitions over time t calculated by: $P^t = e^{tQ}$. I denote by $P_{\mathbf{x}}^t$ the distribution at time t , given initial state \mathbf{x} .

For constant-rate, site-independent models of sequence evolution, the model is fully specified by a single rate matrix Q which operates at every position. The simplest of these models is the Cavender-Farris-Neyman model, which operates symmetrically on two characters, while the Jukes-Cantor model for nucleotide evolution is the fully symmetric process on four characters. The Cavender-Farris-Neyman and Jukes-Cantor models are members of a family of fully symmetric models. I use the

term **standard symmetric evolution** to denote the model with $m \times m$ matrix Q :

$$Q = \begin{pmatrix} -1 & \frac{1}{m-1} & \cdots & \frac{1}{m-1} \\ \frac{1}{m-1} & -1 & \cdots & \frac{1}{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m-1} & \frac{1}{m-1} & \cdots & -1 \end{pmatrix}$$

More recently, large studies of amino acid substitution have been performed to develop more realistic rate matrices for protein evolution (Dayhoff et al., 1978; Jones et al., 1992). These matrices provide a more accurate representation of amino acid substitutions, but retain the assumptions of constant rates and independence. Newer models assume independence between sites but allow the rate of evolution to differ through multiple rate matrices (Yang, 1994). The most popular of these are the $+\Gamma$ models of rate variation, in which a single rate matrix Q is scaled by several values drawn from a gamma distribution, resulting in regions which evolve in the same fashion, but at different rates.

While the cutoff applies to all models of sequence evolution with site independence, including $+\Gamma$ models, etc., for simplicity in the development of ideas only constant-rate evolutionary models will be dealt with here. The majority of ideas and quantitative bounds developed are easily applied to heterogeneous models as well.

We adopt the convention that each model has one expected substitution per site per time unit (at equilibrium). Thus this normalization differs from the formulation of Diaconis et al. (1990) in that I hold the substitution rate per site constant as n increases, rather than the absolute rate. This is more in keeping with conventional use in biological applications, but results in cutoffs that differ by a factor of n from previous work.

4.2.2 Definitions

Total variation distance. The total variation distance between two discrete probability distributions p and q on \mathcal{X} is defined by $\|p - q\|_{TV} := \max_A |p(A) - q(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|$.

Cutoff. Given transition kernels P_n with state spaces \mathcal{X}_n and limiting distributions π_n , the cutoff phenomenon holds (in total variation) for the family (\mathcal{X}_n, P_n) if there exists a sequence (t_n) of positive reals such that

- $\lim_{n \rightarrow \infty} t_n = \infty$;
- For any $\epsilon \in (0, 1)$, $\lim_{n \rightarrow \infty} \|P_n^{(1+\epsilon)t_n} - \pi_n\|_{TV} = 0$;
- For any $\epsilon \in (0, 1)$, $\lim_{n \rightarrow \infty} \|P_n^{(1-\epsilon)t_n} - \pi_n\|_{TV} = 1$.

Note that in the present work only cutoffs in total variation distance are considered, but the cutoff may be defined with respect to other distances as well. Also, as formally defined, the cutoff is a property of an infinite *family* of distributions, and so makes no explicit statement about the behavior of any finite member of this family. I explicitly demonstrate that for common lengths of biological sequences, the asymptotic results are good approximations for actual behavior.

Product spaces. A series of rate matrices Q_1, \dots, Q_n operating on state spaces $\mathcal{M}_1, \dots, \mathcal{M}_n$ with $|\mathcal{M}_i| = m_i$ define an $\prod m_i \times \prod m_i$ matrix $Q^{(n)}$ operating on $\mathcal{M}^n = \mathcal{M}_1 \otimes \dots \otimes \mathcal{M}_n$. $Q^{(n)}$ can be calculated with the recursion

$$Q^{(n)} = I_{m_n} \otimes Q^{(n-1)} + Q_n I_{m^{(n-1)}}$$

with $Q^{(1)} = Q_1$, I_m the m -dimensional identity matrix, and $m^{(n-1)} = \prod_{i=1}^{n-1} m_i$. Generally $Q^{(n)}$ need not be constructed explicitly, as knowledge of the eigenvalues and eigenvectors will suffice and these can be calculated from the eigenvalues and

eigenvectors of the individual Q_i . If Q_i has eigenvalues and eigenvectors λ_{ij} and V_{ij} , the eigenvalues of $Q^{(n)}$ are

$$\lambda_I = \sum_{i=1}^n \lambda_{iI_i}, \quad I \in \{1, \dots, m_1\} \otimes \dots \otimes \{1, \dots, m_n\} \quad (4.1)$$

and $V_I(x_1, \dots, x_n) = V_{1i_1}(x_1) \dots V_{ni_n}(x_n)$ where $I = (i_1, \dots, i_n)$. Also note that if the eigenvectors V_{ij} are normalized according to $\sum_x V_{ij}(x)^2 \pi_i(x) = 1$, then $\sum V_I(\mathbf{x})^2 \phi(\mathbf{x}) = 1$, where $\phi(\mathbf{x}) = \prod \pi_i(x_i)$ is the equilibrium distribution on the product space.

Eigenvalue bound. The total variation distance between a continuous-time Markov chain at time t and its equilibrium distribution can be bounded by the eigenvalues and eigenvectors of Q (Aldous and Fill, 2002). The rate matrix Q has eigenvalues λ_i , $\lambda_1 = 0 > \lambda_2 \geq \lambda_3 \dots \geq \lambda_{|\mathcal{X}|}$, with corresponding eigenvectors V_i . If the eigenvectors are normed such that $\sum V(x)^2 \pi(x) = 1$ (where π is the stationary distribution of Q), then a bound follows from the Cauchy-Schwarz inequality (Diaconis and Stroock, 1991):

$$4\|P_x^t - \pi\|_{TV}^2 \leq \sum_{i=2}^{|\mathcal{X}|} V_i(x)^2 e^{2\lambda_i t} \quad (4.2)$$

Hellinger distance. Another distance defined on distributions, the Hellinger distance is particularly useful as a bound for the total variation distance. The (squared) Hellinger distance is defined by

$$H(p, q)^2 = \sum_{i=1}^{|\mathcal{X}|} (\sqrt{p_i} - \sqrt{q_i})^2.$$

4.3 Symmetric Evolution

Understanding of a prior on evolutionary distances will change depending on the location and behavior of the cutoff for particular models and sequence lengths. With

this in mind, much of this chapter is concerned with relating the asymptotic phenomenon of the cutoff to practical results for finite sequence lengths and determining the nature of the transition to equilibrium in the cutoff region. I begin in this section by extending one of the few results that not only identifies the location of a cutoff, but fully characterizes the transitional behavior of the distribution. Diaconis et al. (1990) derived the asymptotic behavior of the cutoff transition for binary symmetric evolution. I follow their argument and prove that the result holds for fully symmetric evolution on m characters, for all $m \in \mathbb{N}$. For $m = 4, 20$, and 64 , this is the neutral evolutionary process on nucleotides (Jukes-Cantor), amino acids, and codons, respectively.

For models of this type the total variation distance to equilibrium can be calculated explicitly, and thus observe the cutoff behavior exactly. For a single variable X_t following this process:

$$\Pr[X_t = i \mid X_0 = i] = \frac{1}{m} (1 + (m-1)e^{-\frac{mt}{m-1}})$$

$$\Pr[X_t = j : j \neq i \mid X_0 = i] = \frac{1}{m} (1 - e^{-\frac{mt}{m-1}})$$

Thus the probability of a particular sequence \mathbf{x} descending from \mathbf{x}_0 in time t depends only on the Hamming distance $d(\mathbf{x}, \mathbf{x}_0)$ between the two, which allows partitioning of the sequence space into $n + 1$ groups, each of whose elements share the same probability at time t for a given \mathbf{x}_0 :

$$\Pr[\mathbf{X}_t = \mathbf{x} \mid \mathbf{X}_0 = \mathbf{x}_0] = \frac{1}{m^n} \left(1 - e^{-\frac{mt}{m-1}}\right)^{d(\mathbf{x}, \mathbf{x}_0)} \left(1 + (m-1)e^{-\frac{mt}{m-1}}\right)^{n-d(\mathbf{x}, \mathbf{x}_0)}$$

For each distance $0 \leq d \leq n$, there are $\binom{n}{d}(m-1)^d$ sequences. Thus the total

variation distance to uniform U for any initial vector \mathbf{x}_0 is

$$\|P_n^t - U\|_{TV} = \frac{1}{2m^n} \sum_{d=0}^n \binom{n}{d} (m-1)^d \left| (1 - e^{-\frac{mt}{m-1}})^d (1 + (m-1)e^{-\frac{mt}{m-1}})^{n-d} - 1 \right|$$

Theorem 1. *Let $t = \frac{m-1}{2m} \log(n(m-1)) + c$. For P_n^t , the distribution of standard symmetric evolution over sequences in \mathbb{Z}_m^n , then as $n \rightarrow \infty$,*

$$\|P_n^t - U\|_{TV} = \text{Erf}(e^{-\frac{mc}{m-1}}/\sqrt{8}) + o(1)$$

where $\text{Erf}(z) := (2/\sqrt{\pi}) \int_0^z e^{-t^2} dt$ denotes the error function, and $U(\mathbf{x}) = m^{-n}$ is the uniform distribution.

Proof. Note that in P_n^t there is no reference to an initial state \mathbf{x}_0 because under symmetric evolution the total variation distance at time t is the same regardless of initial state. First, rewrite the total variation distance as a sum over all sequences with probability greater than uniform; since the probability of a sequence is monotonically decreasing in $d(\cdot, \mathbf{x}_0)$, there is a threshold d_t^* that defines this set of sequences. Let $\mathcal{X}_d = \{\mathbf{x} : d(\mathbf{x}_0, \mathbf{x}) = d\}$, then $d_t^* = \max\{d : P_n^t(\mathcal{X}_d) \geq U(\mathcal{X}_d)\}$. Then

$$\begin{aligned} \|P_n^t - U\|_{TV} &= \frac{1}{m^n} \sum_{d=0}^{d_t^*} \binom{n}{d} (m-1)^d \left((1 - e^{-\frac{mt}{m-1}})^d (1 + (m-1)e^{-\frac{mt}{m-1}})^{n-d} - 1 \right) \\ &= \sum_{d=0}^{d_t^*} \binom{n}{d} p^d (1-p)^{n-d} - \sum_{d=0}^{d_t^*} \binom{n}{d} \frac{(m-1)^d}{m^n} \\ &= \Pr(\text{Bin}(n, p) \leq d_t^*) - \Pr(\text{Bin}(n, \frac{m-1}{m}) \leq d_t^*) \end{aligned} \quad (4.3)$$

with $\text{Bin}(n, q)$ a binomial random variable and $p = \frac{m-1}{m}(1 - e^{-\frac{mt}{m-1}})$. Then the first binomial variable results in $P_n^t(\mathcal{X}_{d_t^*})$ and the second gives $U(\mathcal{X}_{d_t^*})$.

To solve for d_t^* , note that $P_n^t(\mathbf{x}) \geq U(\mathbf{x})$ when

$$d \leq \frac{n \log(1 + (m-1)e^{-\frac{m}{m-1}t})}{\log(1 + (m-1)e^{-\frac{m}{m-1}t}) - \log(1 - e^{-\frac{m}{m-1}t})}$$

Taking the Taylor expansion of the logs and substituting $t = \frac{m-1}{2m} \log(n(m-1)) + c$ leads to

$$d_t^* = \frac{2(n(m-1))^{3/2} - n(m-1)^2 e^{-\frac{mc}{m-1}}}{2\sqrt{n(m-1)} - m(m-2)e^{-\frac{mc}{m-1}}} + O(1).$$

Finally, using this expression in (4.3) and applying the central limit theorem results in

$$\begin{aligned} & \Phi\left(\frac{1}{2}e^{-\frac{mc}{m-1}}\right) - \Phi\left(-\frac{1}{2}e^{-\frac{mc}{m-1}}\right) + o(1) \\ & 2\Phi\left(\frac{1}{2}e^{-\frac{mc}{m-1}}\right) - 1 + o(1) \end{aligned}$$

□

Corollary 2. *Standard symmetric evolution undergoes a cutoff at $t_n = \frac{m-1}{2m} \log(n(m-1))$.*

Proof. To satisfy the definition of a cutoff, consider the total variation distance for $k_n = (1 + \epsilon)t_n$. With $t_n = \frac{m-1}{2m} \log(n(m-1))$, set $c = \frac{(m-1)\epsilon}{2m} \log(n(m-1))$. Then $k_n = t_n + c$, and application of Theorem 1 results in $2\Phi\left(\frac{1}{2}(n(m-1))^{-\epsilon/2}\right) - 1$, whose limit as $n \rightarrow \infty$ is 0 for ϵ positive and 1 for ϵ negative. □

Diaconis et al. (1990) show that there is good agreement with asymptotic results for n as small as 100. The same holds for larger m , although there is a slight tendency to require larger n as m increases. This gives an initial indication that the asymptotic cutoff result does indeed have implications for realistic lengths of biological sequences.

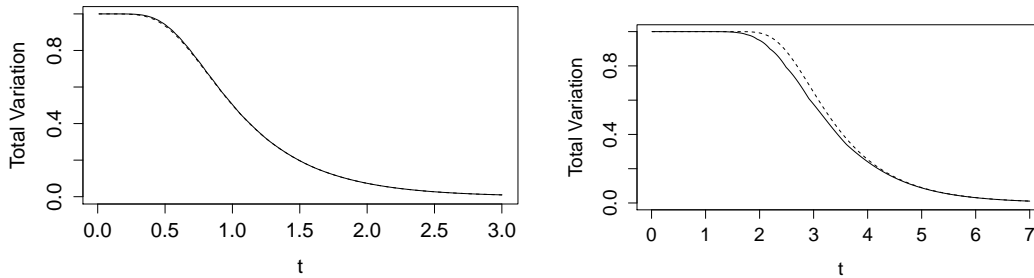


FIGURE 4.1: Actual total variation (solid lines) vs asymptotic approximation (dashed lines) for standard symmetric evolution on sequences of length $n = 100$ with $m = 2$ (left) and $m = 20$ (right). The approximation matches closely in both cases, but not as well for $m = 20$ at this sequence length.

4.3.1 Connection to Distance Estimation

Theorem 1 specifies the way in which the process of standard symmetric evolution approaches equilibrium, measured by total variation distance. The information loss in the approach to equilibrium implies that variance and/or bias in estimators will increase sharply in these region. In this section I make the connection between evolutionary distance estimation more explicit by showing that the cutoff is also the exact boundary beyond which the MLE of t may not be defined.

For standard symmetric evolution, the asymptotic distribution of the MLE \hat{t} of the evolutionary distance can be derived. The likelihood for a single position of the sequence is

$$\left(\frac{1 + (m-1)e^{-\frac{mt}{m-1}}}{m} \right)^x \left(\frac{1 - e^{-\frac{mt}{m-1}}}{m} \right)^{1-x}$$

where x is 1 for a match and 0 otherwise. We can rewrite this as

$$p^x \left(\frac{1-p}{m-1} \right)^{1-x}$$

which, although slightly different from a Bernoulli distribution, yields the same Fisher information, $\frac{1}{p(1-p)}$. Transforming the Fisher information of p into that of t and taking the inverse leads to the asymptotic distribution of the MLE:

$$\sqrt{n}(\hat{t} - t) \rightarrow N \left(0, \frac{e^{\frac{2mt}{m-1}} - (m-1)^2}{m^2} \right)$$

Of course, this asymptotic result will only hold when n is large enough relative to t . This case exhibits the particular problem that the MLE does not exist for any pair of sequences with n/m matches or fewer. Instead, the likelihood is monotonically increasing as $t \rightarrow \infty$. For any fixed t , the probability of this region approaches 0 as n increases. However, if t grows, consider the relationship that must exist between n and t for the MLE to be defined. I show that t must be growing slower than the cutoff $t_n = \frac{m-1}{2m} \log(n(m-1))$. Letting the random variable X be the number of matches in the sequences, the probability of the region where the MLE does not exist can be represented as

$$P \left[X \leq \frac{n}{m} \right] = \sum_{x=0}^{\lfloor \frac{n}{m} \rfloor} \left(\frac{1 + (m-1)e^{-\frac{mt}{m-1}}}{m} \right)^x \left(\frac{(m-1)(1 - e^{-\frac{mt}{m-1}})}{m} \right)^{n-x}$$

Note that this is a slight modification of the likelihood to a binomial distribution by summing over all possible ways a mismatch can occur. We then set $t = t_n$ to obtain a binomial distribution with parameters (n, q) where

$$q = \frac{1 + n^{-1/2} \sqrt{m-1}}{m}$$

Applying the central limit theorem to this binomial distribution results in

$$P \left[X \leq \frac{n}{m} \right] = P \left[Z \leq \frac{-\sqrt{n(m-1)}}{\sqrt{n(m-1)} + O(\sqrt{n})} \right]$$

where Z is a standard normal variable. As $n \rightarrow \infty$, this approaches $P[Z \leq -1]$, so there is always positive probability that the MLE does not exist. Thus the asymptotics do not apply at the cutoff, and n never grows large enough to accurately estimate t_n . However, with $t = (1 - \epsilon)t_n$ (just as in the definition of the cutoff), then

$$P\left[X \leq \frac{n}{m}\right] = P\left[Z \leq \frac{-(n(m-1))^{\frac{1+\epsilon}{2}}}{\sqrt{n(m-1) + O(n^{\frac{1+\epsilon}{2}})}}\right]$$

which is $P[Z < -\infty]$ in the limit, so the MLE is defined with probability 1 as $n \rightarrow \infty$. This provides a further interpretation of the cutoff by explicitly linking it with properties of the MLE. Also note that the variance of the MLE increases exponentially for t less than the cutoff, resulting in the greatest increase in variance as t nears the cutoff.

Asymptotically, the posterior distribution also depends upon the MLE (when it is defined), regardless of the specified prior (Walker, 1969). Thus Bayesian inference for large n will also experience exponential growth in variance as t increases before transitioning past the cutoff. At this point, the likelihood is monotonically increasing as $t \rightarrow \infty$, so the posterior places very little probability on small values of t and inherits the tail behavior of the prior.

4.4 The Cutoff in General Sequence Models

With an understanding of the behavior of the cutoff, I now seek to identify it in the general setting of sequence evolution. The cutoff is formally proven in Theorem 3 of Barrera et al. (2006), which applies to a broad class of exponentially converging Markov chains on product spaces, including all commonly used models of sequence evolution. Essentially, the results says that a sequence of independent processes $(X^{(n)}) = (X_i)_{i \leq n}$ that are exponentially convergent with rates ρ_i , given in increasing

order as $\rho_{(i,n)}$, has a cutoff in total variation at $\tau_n = \max\{\frac{i}{2\rho_{(i,n)}}\}$. There are additional conditions upon the ρ_i , but these are necessary to deal with the possibility of a different process (X_i) for every i . In the case where each X_i is a member of some finite family of processes, the conditions on the ρ_i are satisfied. Note that with common models of sequence evolution there is a finite number of independent processes operating, dictated by the initial character of each process and the number of heterogeneous evolution rate categories. In Appendix C, I give an alternative theorem and proof with sequence models specifically in mind which some readers may find instructive.

As is often the case, most of the difficulty with the cutoff proof is due to special cases. In general, the exponential rate of convergence of an irreducible, continuous-time Markov chain is $|\lambda_2|$, where the eigenvalues of Q are $0 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_k$, and so for most constant-rate models the total variation cutoff time is $\tau_n = \frac{1}{2|\lambda_2|} \log(n)$.

4.4.1 Asymmetric Evolution

While the general rule for the cutoff is dictated by the second eigenvalue, when symmetry is broken cases arise where the timing of the cutoff depends upon the initial state of the chain, even in the iid case. This phenomenon is illustrated with a simple example.

Consider first the spectral decomposition of P^t , the transition matrix generated from rate matrix Q at time t . We have

$$\begin{aligned} P_{ij}^t &= \sum_{k=1}^m V_{ik}(V)_{kj}^{-1} e^{\lambda_k t} \\ &= \pi_j + \sum_{k=2}^m V_{ik}(V)_{kj}^{-1} e^{\lambda_k t} \end{aligned} \tag{4.4}$$

where V is the matrix of eigenvectors and the λ_k are the eigenvalues of Q . Thus every element of the transition matrix experiences exponential decay toward equilibrium,

governed by the eigenvalues of Q . This decay will of course be dominated by λ_2 as long as the coefficient for λ_2 is nonzero. These coefficients depend upon particular states, and so some states will move to equilibrium more quickly than others. A particularly simple case of a zero coefficient occurs when $V_{ik} = 0$ for all k belonging to a particular unique eigenvalue.

Consider the standardized Jukes-Cantor model altered so that the evolution rate of one nucleotide is multiplied by a scalar $\delta \neq 1$. Then Q is

$$Q = \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{\delta}{3} & \frac{\delta}{3} & \frac{\delta}{3} & -\delta \end{pmatrix}$$

The eigenvalues of this matrix are $(0, -4/3, -4/3, -\delta - 1/3)$. It is easy to see that $(1, -1, 0, 0)$ and $(1, 0, -1, 0)$ form a basis for the eigenspace of $-4/3$. From (4.4), the zeroes in the fourth entry of both eigenvectors imply that $-4/3$ does not affect the convergence of state 4. Thus if $\delta > 1$, $-4/3$ is the second eigenvalue, but has no effect on state 4. Note that when $\delta < 1$, the second eigenvalue is $-\delta - 1/3$, and convergence from all initial positions depends upon this value.

4.4.2 Locating the Cutoff for Finite Sequence Lengths

Even among states whose convergence is eventually dominated by λ_2 , for finite times convergence rates will differ. With the JTT and Dayhoff matrices, for example, the eigenvalue coefficients can be easily checked to see that the cutoff of an iid chain does occur at the second eigenvalue, regardless of the initial sequence. The cutoff is an asymptotic result, however, so there is no guarantee that convergence to equilibrium will happen in the region of $\tau_n = \frac{1}{2|\lambda_2|} \log(n)$ for any particular sequence. In particular, while as $t \rightarrow \infty$ eventually only the second eigenvalue is important, at finite times t convergence will be influenced by other eigenvalues to different degrees, depending

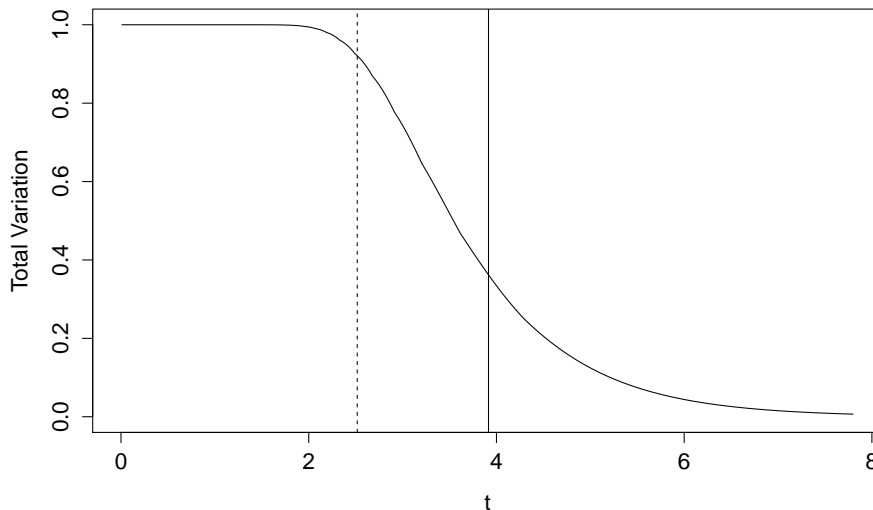


FIGURE 4.2: Location of τ_n (dashed) and $\frac{m-1}{2m}\log(n(m-1))$ (solid) against total variation for standard symmetric evolution with $n = 200$ and $m = 20$. While τ_n is still a valid cutoff for this family, it does not identify the cutoff region as accurately as the more specific expression for symmetric evolution.

on initial conditions.

We note that two principal results in this work appear to contain a contradiction. Theorem 3 guarantees a cutoff for the product family at τ_n for all aperiodic irreducible processes, while Theorem 1 implies that the cutoff for the standard symmetric model as $\frac{m-1}{2m}\log(n(m-1)) = \tau_n + \frac{m-1}{2m}\log(m-1)$. In fact, the cutoff time for a family need not be unique, and both of these are valid cutoff times for standard symmetric evolution. However, the inclusion of the $\log(m-1)$ term results in a closer match to the actual decay in total variation distance for finite sequence lengths (see Figure 4.2). This example highlights the asymptotic nature of the definition of the cutoff.

There are two principal factors that will cause the actual information loss in a sequence to differ from the general cutoff τ_n . The first is the multiplicity of λ_2 , as illustrated in the example above. The additional $\log(m-1)$ factor is explained by the $m-1$ multiplicity of λ_2 . The other major factor (for asymmetric models)

is the initial sequence \mathbf{x}_0 . Diaconis (1996) gives examples of some Markov chains for which there is a cutoff for some starting positions but not for others. While for evolutionary sequence models the cutoff always exists, it can still vary widely according to starting position. For this reason it is useful to examine the behavior of the total variation for specific models and sequences. With the standard symmetric model the total variation can be calculated explicitly, but with the general model this requires summation over m^n terms. I therefore consider some bounds on total variation to more closely examine behavior for finite lengths.

Eigenvalue Bounds on Total Variation

The full eigenvalue bound requires a summation over all elements of \mathcal{X} , just as calculation of the total variation distance directly. With the eigenvalue bound, however, the first few terms will often dominate, and we introduce a useful approximation later on. With symmetric evolution, we can calculate this bound explicitly, and observe the relationship to the actual total variation distance. The eigenvalues for the symmetric case are $\lambda = \{0, \frac{-m}{m-1}, \dots, \frac{-m}{m-1}\}$. The eigenvalues of $Q^{(n)}$, the implied $m^n \times m^n$ rate matrix on sequences, are all possible sums over n choices from λ (taken with replacement). The distinct eigenvalues of $Q^{(n)}$ are then $\{\frac{-mi}{m-1}\}_{i=0}^n$, where the i th value has multiplicity $\binom{n}{i}(m-1)^i$.

With only $n+1$ distinct eigenvalues of $Q(n)$, it remains to calculate the coefficient of each in (4.2). We first make a few observations that will aid in the calculation of this coefficient. For any symmetric Q , we can form an orthonormal matrix of eigenvectors U , so $\sum_x U_i(x)^2 = 1 \ \forall i$ (where U_i is an eigenvector, or column of U). With U orthogonal, we must also have that $\sum_i U_i(x)^2 = 1 \ \forall x$. In other words, the rows of U are also normalized. Again due to symmetry of Q , π is uniform, so we have $V = m^{1/2}U$ in order to achieve the normalization $\sum V_i(x)^2 \pi(x) = 1$. Thus both the rows and columns of V have squared magnitude m . Also note that $V_0 = \mathbf{1}$, so

$$\sum_{j=1}^{m-1} V_j(x)^2 = m - 1 \quad \text{for any } x.$$

Recall that the eigenvectors of $Q(n)$ are formed from products of elements of the vectors of Q . Let λ be the unique nonzero eigenvalue of Q . The third eigenvalue of $Q^{(n)}$ is then 2λ , which results when all elements of I in (4.1) are 1 except for 2. For the coefficient we sum over all ways I can be chosen to satisfy this constraint:

$$\begin{aligned} \sum_{I:\lambda_I=2\lambda} V_I(x)^2 &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} (V_i(x_a)V_j(x_b))^2 \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \sum_{i=1}^{m-1} V_i(x_a)^2 \sum_{j=1}^{m-1} V_j(x_b)^2 \\ &= \sum_{a=1}^{n-1} \sum_{b=a+1}^n \sum_{i=1}^{m-1} V_i(x_a)^2 (m-1) \\ &= \binom{n}{2} (m-1)^2 \end{aligned}$$

which is the multiplicity of the eigenvalue. Clearly the same telescoping can be applied for arbitrary layers of sums, and so the coefficient of each eigenvalue in the bound is exactly equal to its multiplicity. We can then rewrite the eigenvalue bound as a sum over n terms.

$$4\|P^t - \pi\|_{TV}^2 \leq \sum_{j=1}^n \binom{n}{j} (m-1)^j e^{2j\lambda t} \tag{4.5}$$

$$= [(m-1)e^{2\lambda t} + 1]^n - 1 \tag{4.6}$$

from the binomial expansion, and where $\lambda = -\frac{m}{m-1}$.

The total variation distance and eigenvalue bound for standard symmetric evolution with $m = 20$ are given for several values of n in Figure 4.3. The bound matches the actual distance better as n and t increase. For all n , the approach to equilibrium begins more quickly than strictly demanded by the bound.

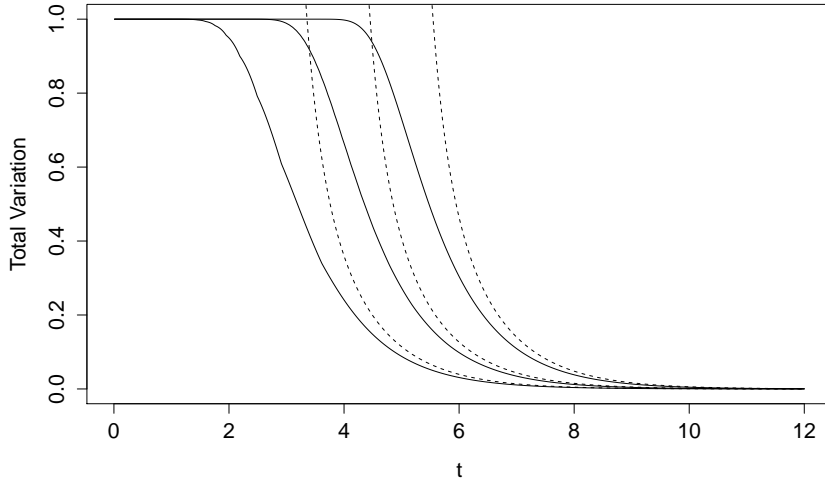


FIGURE 4.3: Total variation distance and eigenvalue bounds for symmetric random walk with $m = 20$ for $n = 100, 1,000, \text{ and } 10,000$. In all cases, the eigenvalue bound (dashed line) approximates the total variation distance closely toward the end of the cutoff region. The cutoff begins significantly sooner than demanded by the eigenvalue bound, although the cutoff grows sharper as n increases.

Approximate Eigenvalue Bound for Asymmetry

Calculation of the eigenvalue bound for a general process is more difficult than the symmetric case because there are more distinct eigenvalues and the eigenvectors do not normalize as nicely. We can still calculate an approximate bound under the assumption that the eigenvector coefficient for each eigenvalue is simply the multiplicity of the eigenvalue. The approximation will hold more closely for near-symmetric Q , which is the case for many models of biological sequence evolution. Then the approximate bound on (four times the squared) total variation distance is

$$\sum_{k_1+k_2+\dots+k_m} \binom{n}{k_1, k_2, \dots, k_m} \exp \left\{ 2t \sum_{i=1}^m k_i \lambda_i \right\} - 1$$

which is the multinomial expansion of

$$\left(\sum_{i=1}^m e^{2t\lambda_i} \right)^n - 1$$

Hellinger Distance

The connection between a multiplicity of the second eigenvalue and the cutoff has led to eigendecompositions taking a central role in finding bounds on the total variation distance (Diaconis, 1996). For product spaces, however, while approximate eigenvalue bounds can still give an estimate of total variation, the bounds are difficult to calculate exactly. In addition, the eigenvalue bounds (discussed in detail in Appendix 4.4.2) only provide an upper bound on total variation. For the application to evolutionary models a lower bound is of particular interest in order to identify the region over which uncertainty will begin to increase.

The Hellinger distance is particularly useful for this situation because it is easily calculated over product distributions and can provide both upper and lower bounds on total variation. The first property says that the Hellinger distance between two product distributions can be calculated from the Hellinger distances between their components:

$$H(P_1 \otimes \cdots \otimes P_n, Q_1 \otimes \cdots \otimes Q_n)^2 = 1 - \prod_{i=1}^n (1 - H(P_i, Q_i)^2)$$

The total variation distance does not have this property, making the Hellinger distance much easier to compute than total variation for product distributions. Further, the Hellinger distance provide both upper and lower bounds on total variation:

$$H(P, Q)^2 \leq \|P - Q\|_{TV} \leq \sqrt{H(P, Q)^2(2 - H(P, Q)^2)}$$

We can use these bounds to explore the effect of \mathbf{x}_0 on convergence behavior for

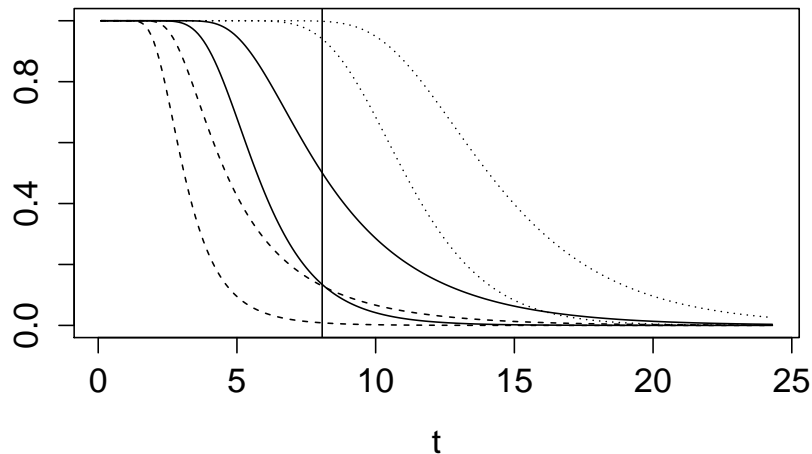


FIGURE 4.4: Hellinger distance bounds on total variation for different initial sequences \mathbf{x}_0 of length $n = 150$ with the adjusted JTT model. The dashed bounds correspond to an initial sequence composed entirely of serine, one of the fastest evolving amino acids under the model, while the dotted bounds result from an initial sequence of tryptophan. The solid bounds were computed for a sequence drawn from the equilibrium distribution. The vertical line is τ_n and falls well within the solid bounds.

realistic models of sequence evolution. For $\mathbf{x}_0 \sim \pi$, the equilibrium distribution, τ_n does a reasonable job identifying the region where the process begins to move quickly toward equilibrium, but for many models choices of \mathbf{x}_0 exist which evolve much more slowly or quickly. Figure 4.4 illustrates this for the adjusted JTT model. For an initial sequence composed entirely of serine, the cutoff occurs much more quickly than for an initial sequence of tryptophan, and neither is well predicted by τ_n . In both cases, however, the Hellinger distance is simple to calculate and provides useful bounds on the total variation.

4.5 Simulations

We have detailed asymptotic results for the cutoff in symmetric models and shown how to use the Hellinger distance to find the region over which total variation distance to equilibrium begins to decay for general models and sequences. In this section the connection between the cutoff in total variation distance and a “cutoff” in evolutionary distance estimation is explored by comparing simulations to the theoretical results derived.

For distance estimation between a pair of proteins with the adjusted JTT matrix, a swift change in the posterior is observed between 4 and 5 time units apart. I simulated 100 descendants of the alpha chain of the human globin 2DN2 at multiple time intervals, and estimated the evolutionary distance for each. Figure 4.5 gives the mean of the 2.5%, 50% and 97.5% percentiles of the posteriors at each simulated time point. For each trajectory and each time point, the posterior was numerically integrated, and for the prior exponentials with rates 1 and 0.1 were used. For the more diffuse prior, the width of the intervals begins to increase at about $t = 3$ and is very large by $t = 5$. The $\text{Exp}(1)$ prior has tighter intervals but underestimates divergence times throughout the trajectory.

The behavior of the posterior as information from the likelihood disappears strongly depends upon the prior. For the large-mean prior used, 97.5% percentiles are large (the 97.5% percentile of the prior is about 37), so even a single posterior with a large interval causes an increase in the mean interval length. The effect of the particular prior can be controlled somewhat by examining the proportion of simulations for which the prior begins to dominate. This could be measured by the proportion of simulations with 97.5% interval above some threshold. In order to avoid bias from any particular threshold, Figure 4.6a gives results for thresholds 10 and 25, along side the Hellinger distance bounds. Regardless of the threshold used,

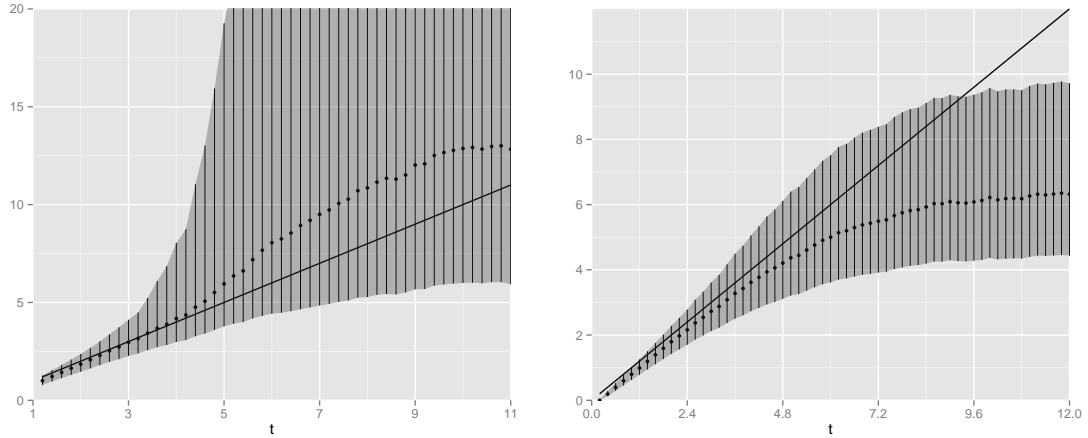


FIGURE 4.5: Mean of 2.5%, 50%, and 97.5% percentiles for posteriors of 100 trajectories simulated at several time points, with exponential prior with mean 10 (left) and mean 1 (right). The solid line ($y = x$) gives the true simulated evolutionary distance. With the mean 10 prior, posterior intervals begin to widen at $t = 3$ and by $t = 5$ are very broad, while the mean 1 prior consistently underestimates the true distance.

nearly all simulations make the transition between 3 and 6 time units, with the majority between 4 and 5.

The JTT cutoff in evolutionary distance estimation occurs earlier than the predicted cutoff in total variation. This occurs because of the shape of the likelihood, which asymptotically approaches $\pi(\mathbf{x})$ as $t \rightarrow \infty$, and thus does not approach 0. The binary case (CFN model) is particularly simple but is representative of this behavior, with the likelihood approaching 2^{-n} as $t \rightarrow \infty$. The general shape of the likelihood is shown in Figure 4.8. The choice of prior will determine exactly when the flat “tail” of the likelihood comes to dominate the lower mode; for noninformative priors this begins to happen especially quickly. Also note that, as preliminarily observed in Chapter 2, as the average credible interval widens on the left side of Figure 4.5, not all of the intervals are so wide. However, those that are not tend to be the simulations in the ‘tail’ of the distribution of sequences at each t , and so t is underestimated. Figure 4.7 gives those simulations which maintain a relatively small

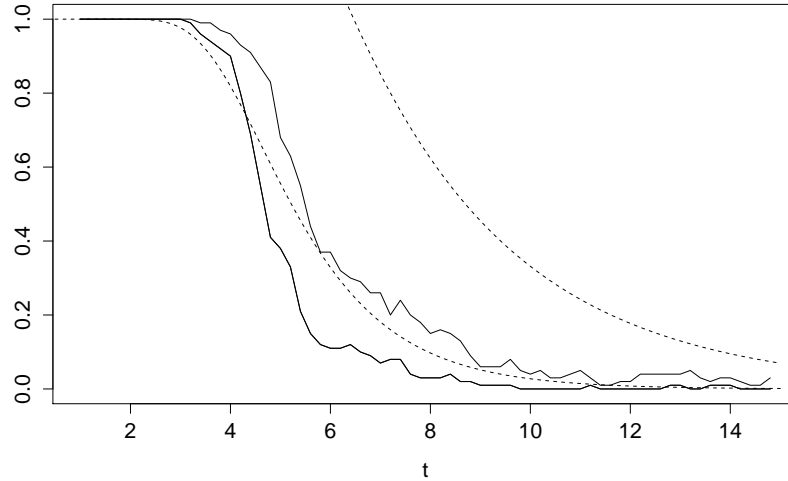


FIGURE 4.6: Solid lines: proportion of simulations at each time point with 95% interval length less than 10 (lower) and 25 (upper). Dashed lines: Hellinger distance bounds on total variation distance. The proportion of simulations without broad intervals follows the lower bound much more closely than the upper, and for intervals of length 10 falls considerably below even the lower bound on total variation distance, indicating that the lower Hellinger bound should be used conservatively.

credible interval longer than other simulations. It is clear to see that that these are the simulations with fewer than expected substitutions from the beginning of the simulation trajectory.

Uncertainty in evolutionary distance can occur while the total variation distance is still relatively high because significant probability mass can be placed on distant sequences long before all sequences receive equilibrium probability. For example, in the binary case of length n with the Hamming distance $d(\mathbf{x}_0, \mathbf{x})$, $\min d(\mathbf{x}_0, \mathbf{x}) = 0$ and $\max d(\mathbf{x}_0, \mathbf{x}) = n$. However, for all sequences with $d(\mathbf{x}_0, \mathbf{x}) \geq \frac{n}{2}$ the likelihood is monotonically increasing as $t \rightarrow \infty$. Thus there is no maximum likelihood estimate and Bayesian inference is principally determined by the prior. However, there can be significant mass on this set while the distribution is still far from equilibrium, and considerable uncertainty begins even before this set is reached.

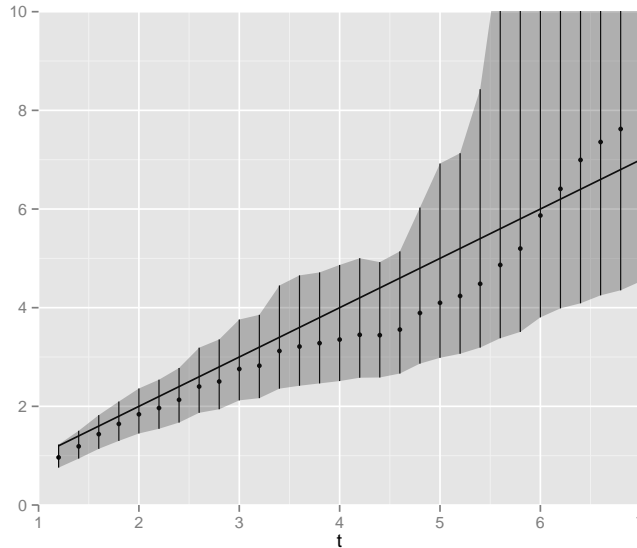


FIGURE 4.7: Simulations with mean 10 exponential prior with width of credible interval less than 3 at $t = 4.4$. This subset maintains tighter intervals only because they appear closer to the initial sequence than they actually are, resulting in underestimation of t from the beginning of the trajectory. This set also makes the transition to wide intervals soon afterward.

4.6 Priors

Well-chosen priors help to manage the trade-off between variance and bias and allow posterior distributions to accurately reflect uncertainty. The previous section gives an initial indication that exponential priors may not be suitable for evolutionary distance estimation, as they are prone to either high variance or high bias. Under the exponential prior the relative weight given to two branch lengths depends solely on the difference between them

$$t_2 - t_1 = t_4 - t_3 \iff \frac{\pi(t_2)}{\pi(t_1)} = \frac{\pi(t_4)}{\pi(t_3)}$$

This can result in long tails which extend into the cutoff region; too little shrinkage is performed in this area. In fact, it seems impossible to choose an exponential prior which will apply appropriate shrinkage beyond the cutoff region without overly

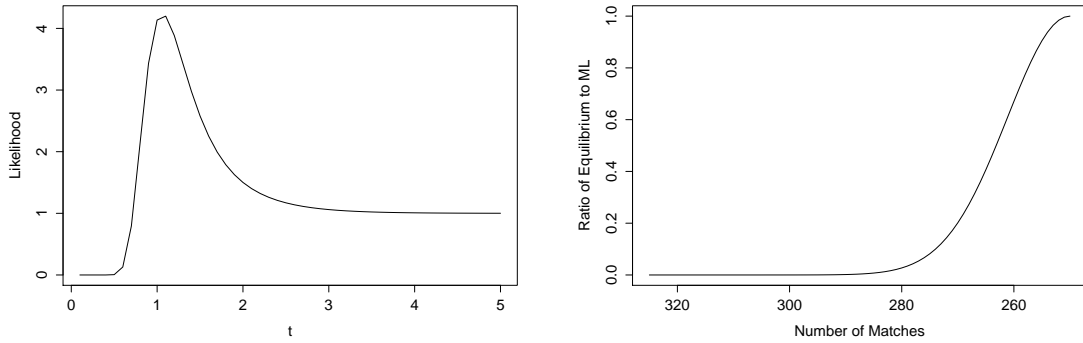


FIGURE 4.8: Left: The shape of the likelihood for distance estimation of binary sequences. The maximum occurs at $t = -\frac{1}{2}\log(\frac{2k}{n} - 1)$ for $2k > n$ where k is the number of identical positions, after which the likelihood decays asymptotically toward 2^{-n} (curve shown is proportional). The likelihood plateaus as $t \rightarrow \infty$ causes the increase in posterior variance as sequence identity decreases. Right: The ratio of equilibrium likelihood to maximum likelihood as the number of observed substitutions increases for a binary sequence of length 500. The ratio is very small while the number of matches is > 280 , but transitions quickly toward 1 beyond this point. The likelihood for more sophisticated models does not depend solely on the number of matches, but exhibits similar behavior.

shrinking smaller estimates. Ideally, a prior that would be essentially flat over a range of biologically values and decay quickly thereafter.

The exponential distribution is not versatile enough to provide this behavior. For that matter, parametric distributions in general are not well-suited for this behavior. I propose a piecewise prior with the first segment diffuse and second with light tails. Although this leaves many possible choices, here I explore use of a uniform-normal prior, where the normal distribution is truncated at the mean and continuity is enforced between the uniform and normal densities. The prior then requires specification of only two parameters: c , the point of transition between the two distributions, and σ^2 , the variance of the (truncated) normal distribution. The other parameters

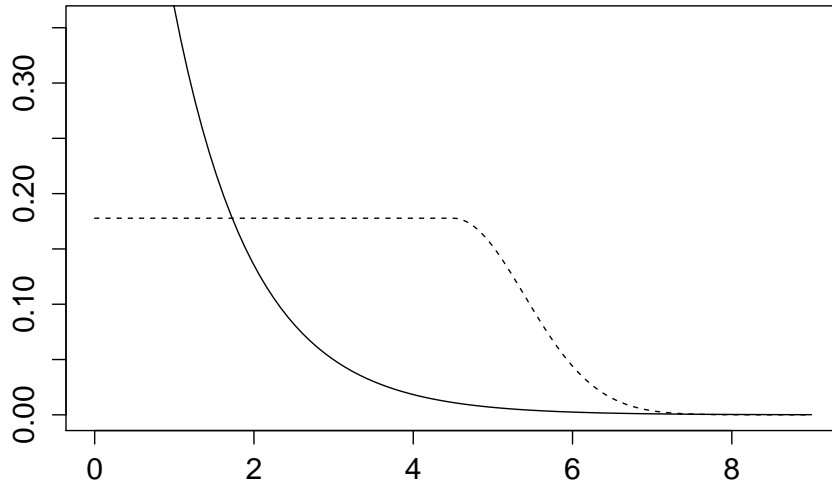


FIGURE 4.9: Mean-one exponential prior (solid line) vs uniform-normal prior with $p = .8$ and $c = 4.5$. The uniform-normal prior results in no shrinkage until c , then drops off more quickly than the exponential prior.

of the piecewise distribution are then determined:

$$\mu = c$$

$$p = \frac{2c}{2c + \sqrt{2\pi\sigma^2}}$$

where μ is the mean of the normal distribution and p is the probability mass of the uniform component. Figure 4.9 gives an example of this prior with $c = 4.5$ and $p = .8$, compared to an $\text{Exponential}(1)$ prior. The choice of c and σ^2 allows many possible behaviors of this distribution. In particular, as σ^2 decreases, the transition from uniform to normal becomes sharper, while for larger σ^2 the normal distribution flattens and the transition is smoother.

We give an example of inference on a pair of proteins to illustrate the potential difference in posteriors with these priors. Assume for the sake of the example that prior knowledge is available which places a strict upper bound on the evolutionary

distance between two proteins at θ expected substitutions per site. This situation is not uncommon, as previous studies of the rate of evolution using fossil records can give estimates of chronological time. Parameters can be chosen for the uniform-normal prior such that the probability of exceeding θ is extremely low, and yet no shrinkage is performed until a critical value c . In the example take $\theta = 7$ and $c = 4.5$, and compare to an Exponential(1) prior. I simulated two sequences, one at distance 4 from its ancestor, the other at distance 6. The exponential prior shrinks the smaller distance more than the larger, relative to the uniform-normal prior, as shown in Figure 4.10. Although the exponential prior places less probability beyond the threshold of 7 (0.00091 vs 0.0026 for the uniform-normal prior), the posterior resulting from the exponential prior has much more mass in this region, due to the tail behavior of the distributions.

The example is meant to illustrate that tail behavior of priors is particularly important because of the cutoff phenomenon, and the prior probability placed on the region beyond the cutoff can be less important than the behavior of the prior in this region. Exponential priors in particular have heavy tails, and are unable to provide appropriate decay at extreme distances without overly shrinking smaller distances.

4.6.1 *Tree Priors*

The most common approach to tree priors has been to assume that branch lengths are independent and identically distributed *a priori*, with the length of each branch following an exponential distribution. Software such as MrBayes and StatAlign use this default assumption. There have been several recent studies related to the impact of branch-length priors on phylogenetic inference, which indicate that this default assumption is inadequate. Brown et al. (2010) note that many recent Bayesian phylogenetic studies contain branch lengths that are much larger than maximum

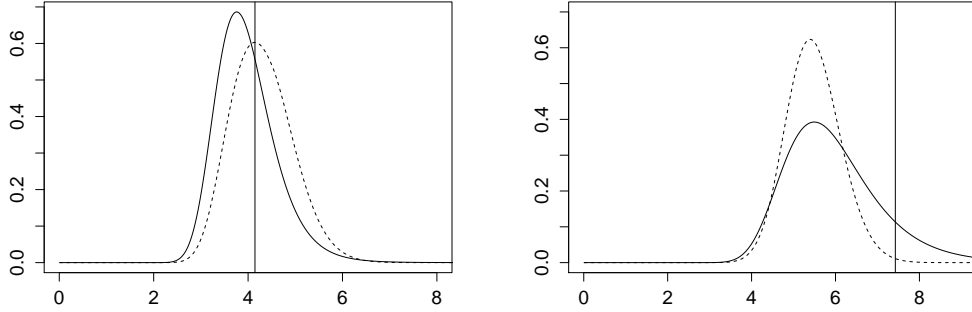


FIGURE 4.10: Posterior distributions of evolutionary distance between two protein sequences with mean-one exponential (solid lines) and uniform-normal prior with $p = .8$ and $c = 4.5$. The maximum likelihood estimate is given in each plot as a vertical line. The exponential prior shrinks intermediate distances more than the uniform-normal prior, but applies less shrinkage to extreme values. The normal tail of the uniform-normal prior forces more appropriate decay beyond the plausible region.

likelihood estimates. The authors provide an excellent exploration of three general hypotheses which may explain this observation, the first of which (involving multimodality) is rejected in the paper. The second and third hypotheses both deal with a large flat region of posterior space outside of a peak in the neighborhood of the maximum likelihood estimate. The difference between the second and third hypotheses is in the posterior mass contained in this flat region. In the first of these, the large flat region has very little mass but Markov chains spend a large amount of time there because of the flatness of the posterior. In the second, the mass of the flat region is large, and so Markov chains are performing correct inference given the likelihood and prior. The authors find support for both of these hypotheses. Indeed, with knowledge of the cutoff behavior in the sequence likelihood, these two situations are precisely what is expected, with a rapid transition between the two as evolutionary distances increase. The authors attribute the cause of Bayesian overestimation to the large volume of the posterior space with long branch lengths.

While this increase in volume as dimension increases exacerbates the problem, the underlying cause of this behavior is the cutoff, which is a fundamental characteristic of sequence evolution models.

Brown et al. (2010) provide a formula for a common parameter for independent branch lengths, which involves a preliminary estimation of the tree through maximum likelihood or some approximate method. From this the average branch length b is calculated, and λ is taken as $\frac{\log(0.5)}{b}$. While the authors are correct to draw attention to the problem of default branch length priors, there are multiple problems with this approach. As I have discussed, the exponential prior is inadequate as a branch length prior and will either shrink small branches too much, or not shrink long branches enough. Further, the prior is dictated by the data with the goal of bringing Bayesian estimates in line with maximum likelihood estimates. This will result in improved inference in many settings, but can easily result in shrinkage and inflated confidence when true branch lengths are approaching the cutoff region.

Much of the difficulty in specification of branch length priors is due to the complexity of the space of phylogenetic trees. It is not easy to see how the choice of branch-length prior affects the implied prior on phylogenies. Brown et al. (2010) deduce that when the tree prior is iid $\text{Exponential}(\lambda)$ on branch lengths, the implied prior on total tree length is $\text{Erlang}(b, \lambda)$, so the prior expectation of tree length grows linearly with the number of branches. Rannala et al. (2012) introduced the compound Dirichlet prior as a joint prior on branch lengths, which places an explicit prior on the tree length rather than allowing this to be determined by the independent branch lengths. However, it is too difficult to quantify prior beliefs related to the total length of a tree, as the relationship between the number of branches and the total tree length is unclear. In general tree length is expected to increase with the number of taxa, but not in the linear way that results from iid branch length priors. Instead, as the number of taxa increases, so does the likelihood that additional

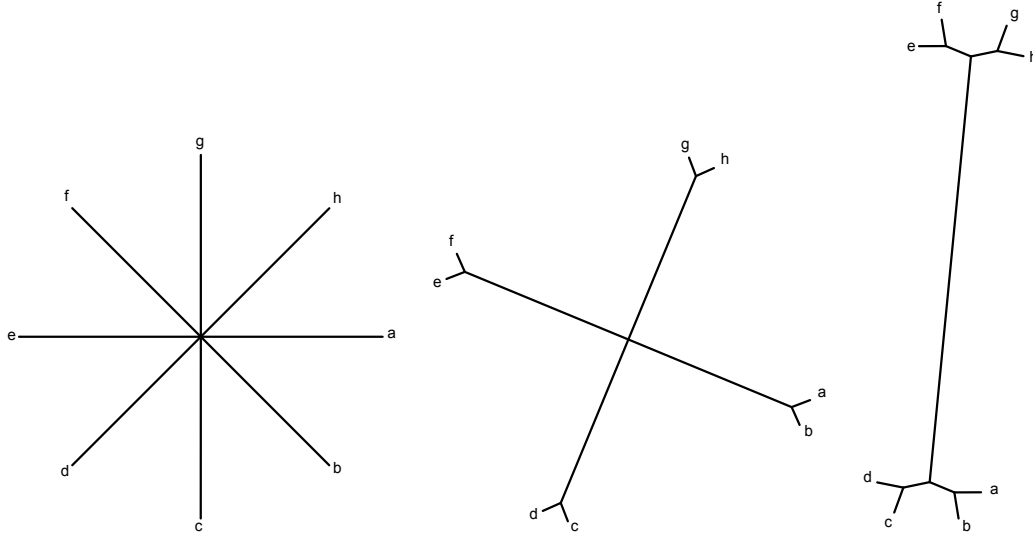


FIGURE 4.11: Three trees with the same number of taxa and same diameter but widely varying tree lengths, illustrating the difficulty of specifying a prior on tree length.

taxa are closely related to those already in the tree, and so the total branch length increases only slightly. I suggest instead the tree diameter as a simpler quantity for prior formulation. Researchers in general will have greater ability to anticipate the maximal distance on the tree than the total tree length, and the two quantities need not have a strong relationship. To illustrate, three trees sharing diameter and number of taxa but with different tree lengths are given in Figure 4.11.

Let Υ be a tree in Θ , the space of all unrooted binary trees with l leaves. Take the prior on Υ to be proportional to the uniform-normal prior on tree diameter

$$\pi(\Upsilon) \propto f(d(\Upsilon))$$

where $f(d)$ is the piecewise uniform-normal prior, and $d(\Upsilon)$ is the diameter of Υ . It can be shown that $\pi(\Upsilon)$ is proper. This requires

$$\int_0^\infty \int_{\Theta_d} \pi(\Upsilon) d\Upsilon dd < \infty \quad (4.7)$$

where Θ_d is the set of all trees with diameter d . The inner integral is proportional to the volume of Θ_d , since $\pi(\Upsilon)$ is uniform on this space. Θ_d can be further subdivided into spaces $\Theta_d^{(i)}$ where i indexes the distinct topologies in Θ . Letting $b = 2l - 3$ be the number of branches of every tree in Θ , note that each $\Theta_d^{(i)}$ is a union of $b - 1$ -dimensional convex polytopes (examples given in Figure 4.13). The i th polytope has volume $c_i d^{b-1}$, where c_i is a constant that depends upon the topology. The volume of Θ is then cd^{b-1} , where c is a constant that does not depend on d . The intuition behind this is that the shape of the space does not change as d varies; only the size does. In the case of three branches, the constrained tree space is the union of three 2-dimensional triangles in \mathbb{R}^3 . The volume of this space is the sum of the areas of the triangles, or $\frac{3d^2}{2\sqrt{2}}$. In the general case Equation (4.7) can be written as

$$\int d^{b-1} f(d) dd < \infty$$

The integrand is again a piecewise function $g(d)$, defined by

$$g(d) \propto \begin{cases} d^{b-1} & d \leq c \\ d^{b-1} e^{-\frac{(d-c)^2}{2\sigma^2}} & d > c \end{cases}$$

The first portion of the function is a beta distribution on the interval $(0, c)$. The second is more complicated, but can be represented as a mixture of generalized gamma distributions on the positive quantity $d - c$. This is sufficient to see that the induced prior on trees is proper. However, the increasing volume of the constrained space of trees with diameter d means that the marginal prior on tree diameters $\pi_d(d)$ does not follow the shape given in Figure 4.9. Alternatively, a new prior can be

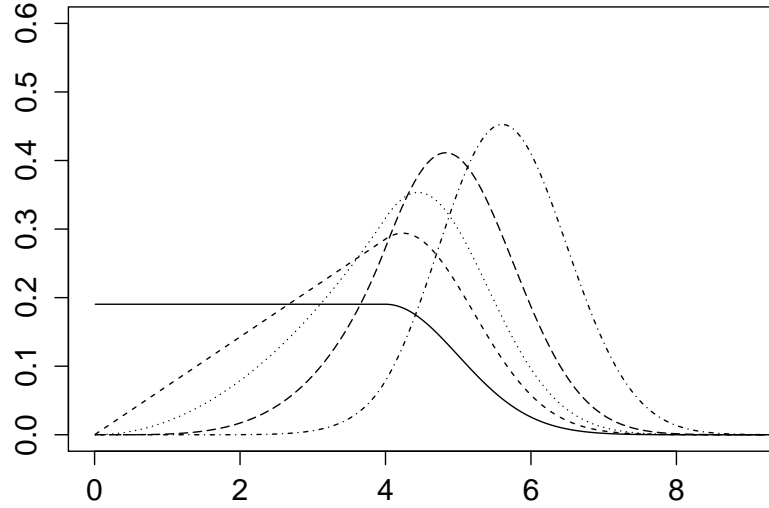


FIGURE 4.12: Marginal prior on tree diameter for trees with 1, 2, 3, 5, and 10 branches when tree prior is chosen such that $\pi(\Upsilon) \propto f(d(\Upsilon))$. The increasing volume of longer diameters causes the distribution to shift to the right as the number of branches increases. The additional shrinkage term of d^{-b+1} results in a marginal piecewise-uniform marginal for any number of branches (solid line).

employed, $\phi(\Upsilon)$, created by adjusting $\pi(\Upsilon)$ so that $\phi(\Upsilon)$ results in the uniform-normal piecewise prior on diameters, $f(d)$. This only a matter of dividing by the proportional volume for each tree:

$$\phi(\Upsilon) \propto d^{-b(\Upsilon)+1} f(d(\Upsilon)) \tag{4.8}$$

Normalization then results in $\phi_d(d) = f(d)$. The choice of exponent on d in $\phi(\Upsilon)$ also offers an additional degree of flexibility in specification of the prior on trees. While the exponent $-b(\Upsilon) + 1$ results in the uniform-normal prior, all exponents larger than this result in varying beta-generalized gamma distributions on the diameter. Examples of the possible changes to the prior by branch lengths and adjustments are given in Figure 4.6.1.

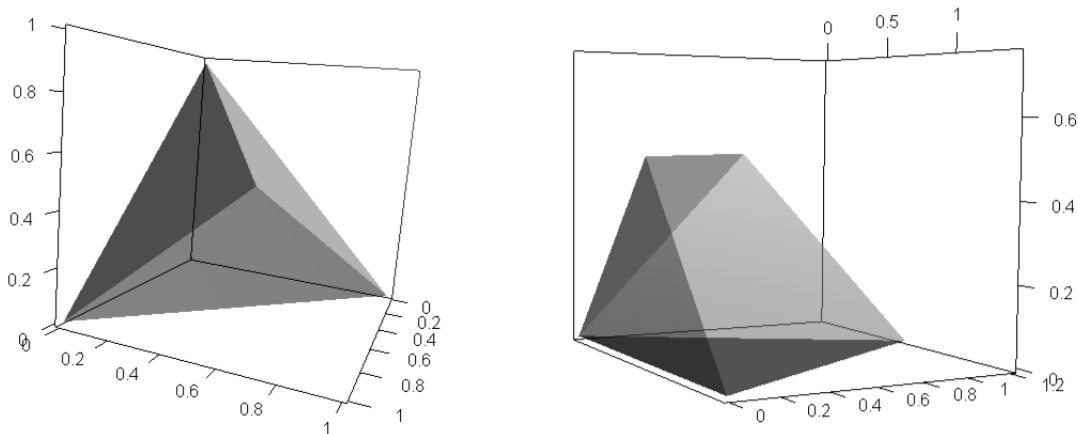


FIGURE 4.13: Examples of the constrained space of trees for a fixed diameter $d = 1$. Left: a full depiction of the constrained 3-branch tree space, where each axis is a branch of the tree. The space consists of three congruent triangles joined at edges. Right: In four dimensions the full space cannot be depicted, but it consists of the union of 3-dimensional polyhedra residing in \mathbb{R}^4 . The figure depicts one of these polyhedra, the 3-dimensional volume bounded by four triangles and a trapezoid.

4.6.2 Simulations

In order to test the diameter prior, I simulated proteins on a known tree and compared inference under exponential branch length priors and the diameter prior. I used the Dayhoff matrix on a symmetric tree of twelve leaves under four different scales, with total tree length ranging from 8.25 to 33 (see Figure 4.14). For each tree, I simulated 20 data sets and analyzed the resulting sequences in MrBayes under the default exponential prior (mean 0.1), as well as mean 1 and mean 10. To arrive at the samples under the diameter prior, I reweighted the MrBayes samples according to the ratio between priors. For each simulation, I also calculated the maximum likelihood tree using the program PhyML (Guindon et al., 2010). Only the MrBayes samples with the mean 1 prior are given in Figure 4.15, as the other exponential priors led to far too much and far too little shrinkage.

The simulations bring to light both the possible over-shrinkage of exponential

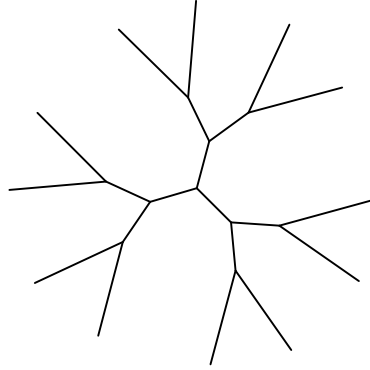


FIGURE 4.14: Tree topology used for simulations. Inner branches are half the length of outer branches. Simulations were performed with the outer branches scaled to 0.5, 1, 1.5, and 2.

priors and the effect of expanding volume of tree space as diameter increases. The default exponential prior (mean 0.1), overly shrinks the branch lengths of these trees. In general, the 0.1 default is somewhat better suited for DNA, where the cutoff (in terms of substitutions per site) occurs earlier than in protein evolution models. The mean 1 exponential prior is more appropriate for this protein setting, but here this prior actually inflates estimates of tree length even for the shorter trees, because the shrinkage from the prior is not enough to counteract the increase in volume of the tree space.

We use the diameter prior with the additional penalty on diameter length given in (4.8), resulting in a piecewise uniform-normal marginal distribution on the tree diameter. Without the additional diameter penalty to counteract the increasing volume of the tree space, the diameter prior also results in overestimated tree length. As seen in Figure 4.15, the diameter prior results in inference centered around maximum likelihood estimates for the shorter trees, because the shrinkage in the prior over this range is matched to the increase in tree volume, with no additional shrinkage to branch lengths performed. As the tree grows longer the MLE of tree length becomes unstable, and the Gaussian tail of the diameter prior is required to prevent inflated

inference.

In this example, the difference between priors grows pronounced as the tree lengthens and approaches the cutoff region. Sequence models with partitions or variable rate categories can experience this problem even on small trees with a high degree of sequence similarity. As noted by Brown et al. (2010) and Marshall (2010), Bayesian trees can become much longer than maximum likelihood trees in these circumstances. This occurs when there is low sequence similarity in one of the rate categories, allowing very long times to be estimated for this category. In the other categories, the rate of evolution can adjust to compensate for the long branch lengths. Exponential branch length priors do not penalize these long trees sharply enough, and do not incorporate prior expectations of tree length or diameter.

4.7 Discussion

We have proposed a new prior for Bayesian inference of phylogenies. The distribution better represents prior information and leads to more appropriate inference over all time scales than independent exponential branch priors. I have also discussed the use of various bounds to identify the limits of sequence information in specific settings. We recommend use of the Hellinger distance lower bound on total variation distance in particular to identify regions beyond which inference is impractical, and light-tailed priors chosen to reflect regions of biological plausibility to correctly manage uncertainty. If distances may plausibly lie in the cutoff region, there may be fundamental uncertainty in the model, and further restricting inference through tighter priors will yield inflated confidence. Rather than modifying priors in an attempt to limit uncertainty, in these circumstances researchers will need to seek out additional information or more informative models. However, even in these settings the use of an appropriate prior will lead to a meaningful posterior distribution, even if it has large variance.

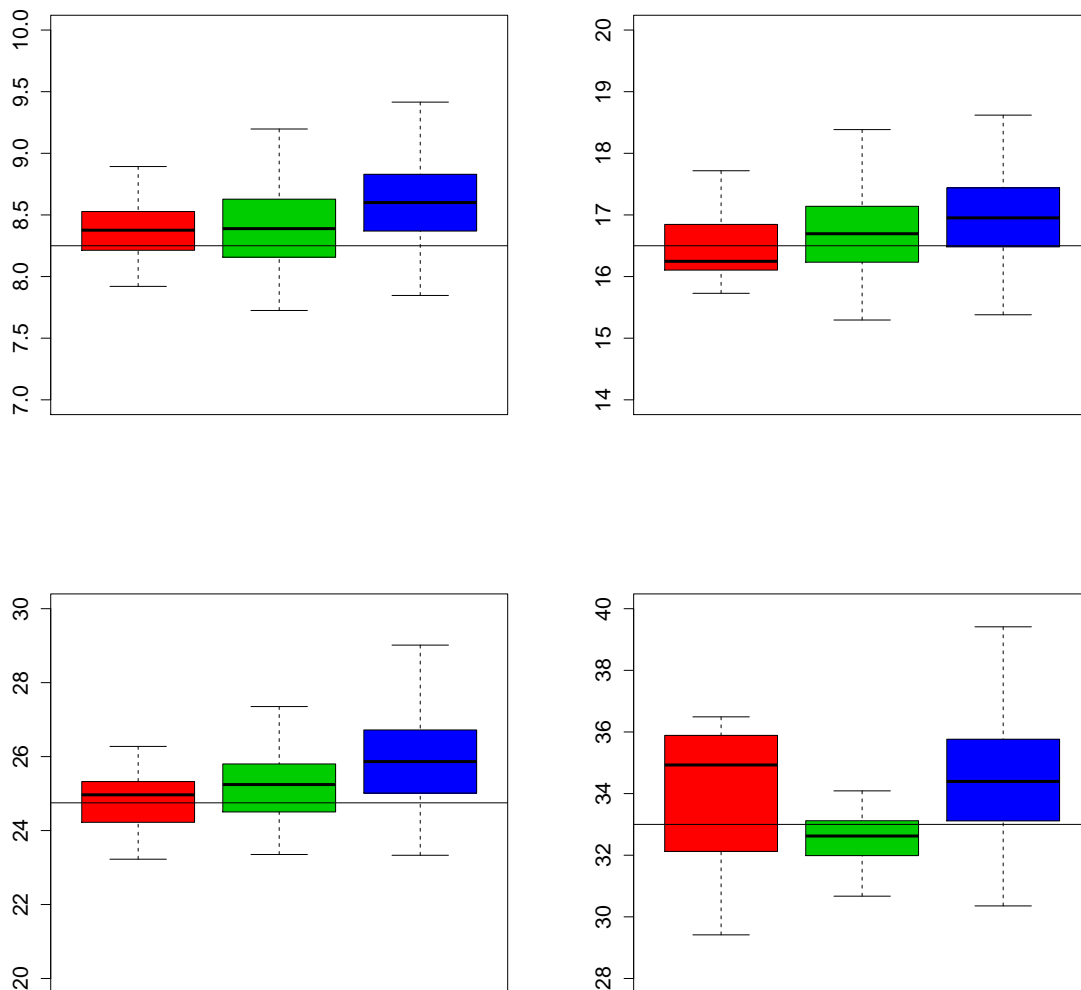


FIGURE 4.15: The box plots in each figure employ the same methods, from left to right. The y -axis gives the total tree length. Left: MLE of tree length over 20 simulations, as calculated by PhyML. Center: average quantiles of posterior distributions with diameter prior. Right: average quantiles of posterior distributions with mean 1 exponential prior. Horizontal line: true tree length used in simulations. The diameter prior allows inference to closely match the MLE for trees of reasonable length, and appropriately scales back estimates for the longest trees. With the exponential prior, tree length is overestimated even for the smallest trees, and this overestimation becomes more pronounced as the tree grows longer.

There are a few other important considerations in the treatment of cutoffs and prior distributions for sequence evolution. I do not attempt to fully address them here, but instead present initial ideas and recommend further work in these areas.

4.7.1 Insertions and Deletions

The presence of an insertion and deletion model in addition to the substitution process causes information to be lost more quickly, even if the alignment is known. This is due in part to the fact that reversible evolutionary models require that inserted characters follow the stationary distribution.

The notion of a cutoff applies only to families of distributions which can be indexed by n operating on state spaces of increasing size. An unrestricted indel model allows sequence lengths to vary and places positive probability on positive lengths, and so even sequences with very different initial lengths are part of the same state space.

A change in the cutoff can be formally shown, however, for a modified indel process where insertions and deletions occur as pairs at the same time instant, thus preserving sequence length. Consider a sequence of length n evolving according to a substitution rate matrix Q with stationary distribution π . In addition to Q , the following process operates on the sequence at rate $n\lambda$: choose a position at uniform and delete it from the sequence. Next, insert a character according to π at a new uniformly chosen location. Considering only sequence space (ie, disregarding alignment space), the stationary distribution of this sequence is the same as the sequence without the indel process present, $\pi^{(n)}$, but convergence to stationarity occurs faster due to the additional changes to the sequence.

The distance to stationarity of this process is identical to a slightly simpler process which (at rate $n\lambda$) chooses a position at random and replaces the character there with a character drawn from equilibrium. A new substitution matrix Q' can be written

which describes this process:

$$Q'_{ij} = Q_{ij} + \lambda\pi_j \quad i \neq j$$
$$Q'_{ii} = Q_{ii} - \lambda(1 - \pi_i)$$

Clearly the eigenvalues of Q' are strictly less than those of Q , causing the cutoff to occur faster in the presence of indels.

Beyond the sequence itself reaching equilibrium more quickly, however, information is also lost with an indel process because of alignment uncertainty.

4.7.2 *Structural Cutoff*

As illustrated by simulations in Chapter 2, alignment uncertainty causes the ‘cutoff’ in estimation ability of evolutionary distance to occur much sooner. Much of this alignment uncertainty can be eliminated through use of a structural diffusion model such as the one developed in Chapter 2. The Ornstein-Uhlenbeck process employed also falls in the category of exponentially convergent processes described by Barrera et al. (2006), and so also exhibits a cutoff, which can be shown to occur at $\frac{1}{2\theta}\log(n)$, where θ is the mean reversion coefficient in the OU process. As shown in Chapter 2, the estimated value of θ is typically quite small; it falls in the neighborhood of 0.005 for the globin analysis conducted in that chapter. For a sequence of length 150, this implies a cutoff time of 501, which is well beyond the maximum biologically plausible distance between sequences.

At the very least, then, inclusion of a structural model allows the cutoff to occur closer to where it would in a sequence model with no alignment uncertainty. Also, if a structural diffusion model is chosen that can well represent structural changes through time, structural information can directly inform evolutionary distances and prolong the cutoff even further. The local- σ model of Chapter 3 is the first model

that allows structure to play a role in determining evolutionary distance without overwhelming sequence information.

5

Summary

The principal focus of this dissertation was to introduce methods of utilizing structural information in protein-based phylogenetics. Although a vast literature on protein alignment and phylogenetics exists, there was previously no way to incorporate the additional source of structural information into phylogenetic analyses.

Chapter 2 outlines the general requirements for a reversible structural model that operates analogously to sequence models, and provides a computationally convenient instance which is shown to yield significant stabilization of evolutionary distance estimation and parameter inference. The particular gains resulting from the inclusion of structural information are in estimation of distant relationships and resolution of splits deep in a phylogeny. The limits of the model to pairwise analysis and the relatively simple form of the diffusion leave opportunity for further development.

Chapter 3 fully addresses the first of these concerns and partially handles the second as well. The model is extended from a pairwise setting to full-tree inference. In addition, the introduction of rate heterogeneity by branch allows for more complex and flexible relationships between sequence and structure divergence. It is shown that while the effect of using structural information is similar to including additional

sequences, circumstances exist in which structure allows resolution of topology not possible with sequence alone. There is also ample evidence of highly variable rates of structural diffusion, which give rise to particular patterns of structural divergence suggestive of biological mechanisms. A software tool for analysis under the model is also provided.

Chapter 4 formalizes the need for structural phylogenetic inference by placing theoretical limits on the information contained in sequences as evolutionary distances grow. An examination of the cutoff phenomenon of Markov chains reveals that sequence evolutionary processes transition swiftly to equilibrium beyond some modest distance, making inference of distance and other parameters impossible. This problem is only exacerbated when the alignment is also unknown. The most commonly used phylogenetic priors are poorly equipped to handle this phenomenon, so a new class of tree prior is introduced, utilizing a piecewise parametric approach more appropriately apply (or not apply) shrinkage in branch length estimation.

Combined, these chapters offer a significant contribution to the field of phylogenetics and offer researchers a new source of information for robust protein phylogenetics. It is the hope of the author that the present work will provide both an immediately useful approach as well as the basis for further work in structural phylogenetics.

Appendix A

TKF91 Transition Matrix

The transition matrix for the Pair HMM used to compute the marginal likelihood across all alignments. Parameters λ and μ and functions $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ are given in Section 2.2.1.

$$\begin{array}{c}
 \text{Start} \\
 \text{Match} \\
 \text{Delete} \\
 \text{Insert} \\
 \text{End}
 \end{array}
 \begin{pmatrix}
 \text{Start} & \text{Match} & \text{Delete} & \text{Insert} & \text{End} \\
 0 & \frac{\lambda}{\mu}(1 - \beta(t))\alpha(t) & \frac{\lambda}{\mu}(1 - \beta(t))(1 - \alpha(t)) & \beta(t) & (1 - \frac{\lambda}{\mu})(1 - \beta(t)) \\
 0 & \frac{\lambda}{\mu}(1 - \beta(t))\alpha(t) & \frac{\lambda}{\mu}(1 - \beta(t))(1 - \alpha(t)) & \beta(t) & (1 - \frac{\lambda}{\mu})(1 - \beta(t)) \\
 0 & \frac{\lambda}{\mu}(1 - \gamma(t))\alpha(t) & \frac{\lambda}{\mu}(1 - \gamma(t))(1 - \alpha(t)) & \gamma(t) & (1 - \frac{\lambda}{\mu})(1 - \gamma(t)) \\
 0 & \frac{\lambda}{\mu}(1 - \beta(t))\alpha(t) & \frac{\lambda}{\mu}(1 - \beta(t))(1 - \alpha(t)) & \beta(t) & (1 - \frac{\lambda}{\mu})(1 - \beta(t)) \\
 0 & 0 & 0 & 0 & 1
 \end{pmatrix}
 \tag{A.1}$$

Appendix B

Data

Several protein datasets are analyzed throughout this work. Details for each set of proteins and the associated structural coordinates are given here.

B.1 Globins from Chapter 2

Chapter 2 provides several examples involving globins from Tables B.1 and B.2. The tables provide the PDB ID from which structural coordinates were obtained and the organism of the protein.

B.2 Simulated data

To begin with, a series of simulations were performed in order to evaluate the ability of the MCMC framework to recover known parameters, alignments, and branch lengths. The data were simulated according to the structural drift model, with $\sigma^2 = 0.7$, $\lambda = 0.03$, $\mu = 0.0305$, $r = 0.67$, and all B -factors equal to 1 for simplicity, using three different tree topologies, with 6, 8, and 10 leaves respectively. The structure at the root was set to be equal to the haemoglobin 2DN2, and model parameters were

Table B.1: PDB entries and corresponding species from Figures 2.1, 2.3, 2.4, and 2.5.

PDB ID	Species	Common Name
1ASH	<i>Ascaris suum</i>	Nematode
1B0B	<i>Lucina pectinata</i>	Lucine clam
1CG5	<i>Dasyatis akajei</i>	Stingray
1GCV	<i>Mustelus griseus</i>	Houndshark
1HBH	<i>Pagothenia bernacchii</i>	Emerald rockcod
1HLB	<i>Caudina arenicola</i>	Sea cucumber
1HV4	<i>Anser indicus</i>	Bar-head goose
1IDR	<i>Mycobacterium tuberculosis</i>	Tuberculosis
1OUT	<i>Oncorhynchus mykiss</i>	Rainbow trout
1X3K	<i>Tokunagayusurika akamusi</i>	Midge larva
1XQ5	<i>Perca flavescens</i>	Perch
2BK9	<i>Drosophila melanogaster</i>	Fruit fly
2C0K	<i>Gasterophilus intestinalis</i>	Botfly
2DHB	<i>Equus caballus</i>	Horse
2DN2	<i>Homo sapiens</i>	Human
2LHB	<i>Petromyzon marinus</i>	Lamprey
2RAO	<i>Oryctolagus cuniculus</i>	Rabbit
2XKI	<i>Cerebratulus lacteus</i>	Milky ribbon worm
2ZFB	<i>Psittacula krameri</i>	Parrot
3A59	<i>Struthio camelus</i>	Ostrich
3A5B	<i>Propsilocerus akamusi</i>	Midge larva
3AT5	<i>Podocnemis unifilis</i>	Side-necked turtle
3BCQ	<i>Brycon cephalus</i>	Red-tailed brycon
3K8B	<i>Meleagiris gallopavo</i>	Turkey
3MKB	<i>Isurus oxyrinchus</i>	Shortfin mako

chosen based upon typical values observed on test runs on small globin datasets. For each topology, branch lengths were multiplied by three different scale factors (1.0, 1.5 and 2.0) in order to yield varying levels of divergence. Since the identifiability of σ^2 and ϵ is of particular interest, we performed each scenario with different values of ϵ . Inference was carried out on each dataset under the structural drift model and the sequence-only model, to see how parameter inference is affected by model choice. Each parameter combination was simulated ten independent times, and results averaged over the ten repetitions.

Table B.2: PDB entries and corresponding species from Figure 2.7.

PDB ID	Species	Common Name
1ASH	<i>Ascaris suum</i>	Nematode
1B0B	<i>Lucina pectinata</i>	Clam
1H97	<i>Paramphistomum epiclitum</i>	Fluke
1HLB	<i>Caudina arenicola</i>	Sea cucumber
1IT2	<i>Eptatretus burgeri</i>	Inshore hagfish
1ITH	<i>Urechis caupo</i>	Innkeeper worm
1MBA	<i>Aplysia limacina</i>	Slug sea hare
1NGK	<i>Mycobacterium tuberculosis</i>	Tuberculosis
1OR6	<i>Bacillus subtilis</i>	Bacillus subtilis
1VHB	<i>Vitreoscilla stercoraria</i>	Vitreoscilla stercoraria
1X9F	<i>Lumbricus terrestris</i>	Earthworm
2D2M	<i>Oligobranchia mashikoi</i>	Gutless beard worm
2DN2	<i>Homo sapiens</i>	Human
2HBG	<i>Glycera dibranchiata</i>	Bloodworm
2XKI	<i>Cerebratulus lacteus</i>	Milky ribbon worm
3A5B	<i>Propiloscerus akamusi</i>	Midge larva

In each case, the structural parameters are recovered to a high degree of accuracy, lying within the 95% highest posterior density interval in all cases, with the posterior median usually very close to the true value. Importantly, we are able to resolve the different contributions from ϵ and σ even without repeated observations at the leaves, illustrating that these separate types of variability are fully identifiable. When the true ϵ is actually equal to zero, it is usually estimated slightly higher than this, partially due to the effect of the prior mass pulling the posterior away from zero, such that σ is also slightly underestimated. However, in almost all other cases parameters are well recovered, even for high evolutionary distances (*see Figures B.3 to B.5*).

B.3 5-globin dataset

As a second example, we consider a set of five globins, which although highly structurally similar (average pairwise RMSD around 1Å), have a very low average sequence identity of about 20%. We take the alignment in the HOMSTRAD database

(Mizuguchi et al., 1998) as the reference alignment for this set.

Table B.3: The 5-globin dataset.

Structure	Protein	Organism
1hlb	haemoglobin	Caudina arenicola (sea cucumber)
1myt	myoglobin	Thunnus albacares (tuna)
2lhb	haemoglobin	Petromyzon marinus (lamprey)
1lh1	leghaemoglobin	Lupinus luteus (lupin bean)
2hbg	haemoglobin	Glycera dibranchiata (bloodworm)

B.4 8- and 12-globin datasets

In addition to the 5-globin dataset, we also consider a larger group of globins that spans across the whole family. Certain regions of this phylogeny corresponding to ancient divergence events are poorly resolved when using only sequence information, and it is therefore of great interest to examine the effect of including structure.

Table B.4: The 8- and 12-globin dataset, grouped according to grouped according to observed clades. Sequences marked with a ‡ are present in both datasets. NsGb = non-symbiotic plant globin; Lhb = leghaemoglobin; Ngb = neuroglobin; HGbI = bacterial Hell’s gate globin I; Cygb = cytoglobin; CycHb = cyclostome haemoglobin; Hb = haemoglobin; Mb = myoglobin. * - length shown for the portion present in the PDB file.

Structure	Protein	Organism	Resolution	<i>R</i> -value	Length*
2oif	NsGb	<i>H. vulgare</i> (barley)	1.80	20.2	153
1bin	Lhb	<i>G. max</i> (soybean)	2.20	19.8	143
1lh1	Lhb	<i>L. luteus</i> (lupin bean)	2.00	27.3	153
1oj6 ‡	Ngb	<i>H. sapiens</i> (human)	1.95	17.8	147
3s1j	HGbI	<i>M. inferorum</i> (thermophile)	1.80	21.0	131
1urv ‡	Cygb	<i>H. sapiens</i> (human)	2.00	22.2	154
2lhb ‡	CycHb	<i>P. marinus</i> (lamprey)	2.00	14.2	149
1myt ‡	Mb	<i>T. albacares</i> (tuna)	1.74	17.7	146
2mm1 ‡	Mb	<i>H. sapiens</i> (human)	2.80	15.8	153
1spga ‡	α -Hb	<i>L. xanthurus</i> (spot croaker)	1.95	19.1	143
2hhba ‡	α -Hb	<i>H. sapiens</i> (human)	1.74	16.0	141
2hhbb ‡	β -Hb	<i>H. sapiens</i> (human)	1.74	16.0	146

B.5 Cysteine proteinase and human protein kinases

Chapter 3 also analyzes a set of cysteine proteinases, given in Table B.5, and human protein kinases, given in Table B.6.

Table B.5: The cysteine proteinase dataset. Average pairwise identity using the HOMSTRAD alignment is 42%. * - length shown for the portion present in the PDB file.

Structure	Protein	Organism	Resolution	<i>R</i> -value	Length*
1aim	Cruzain	<i>T. cruzi</i> (trypanosome)	2.00	18.8	216
8pcha	Cathepsin H	<i>S. scrofa</i> (wild boar)	2.10	NA	221
1mema	Cathepsin K	<i>H. sapiens</i> (human)	1.80	18.3	216
2acta	Actinidin	<i>A. chinensis</i> (kiwi fruit)	1.70	16.5	219
1cqda	Proteinase II	<i>Z. officinale</i> (ginger)	2.10	21.3	217
1yal	Chymopapain	<i>C. papaya</i>	1.70	19.2	217
1ppn	Monoclinic papain	<i>C. papaya</i>	1.60	16.0	213
1gece	Glycyl endopeptidase	<i>C. papaya</i>	2.10	19.6	217
1ppo	Protease omega	<i>C. papaya</i>	1.80	15.5	217

Table B.6: The human protein kinase dataset.

Structure	Protein
1gz8a	Cell division protein kinase 2
2gfsa	Mitogen-activated protein kinase 14
1q5ka	Glycogen synthase kinase-3 beta
1o61a	Aminotransferase
1uu3a	3-phosphoinositide dependent protein kinase 1
1tkia	Titin
1jksa	Death-associated protein kinase
2ozaa	MAP kinase-activated protein kinase 2
1yhwa	Serine/threonine-protein kinase PAK-1
2j4za	Serine/threonine-protein kinase 6
1qpca	LCK kinase
1mp8a	Focal adhesion kinase 1
1t46a	Tyrosine kinase
1p4oa	Insulin-like growth factor I receptor
1r0pa	Hepatocyte growth factor receptor
1t4ha	Serine/threonine-protein kinase WNK1
1u46a	Activated CDC42 kinase 1
1xbba	Tyrosine-protein kinase SYK
1xwsa	Serine/threonine-protein kinase PIM-1
2java	Serine/threonine-protein kinase NEK-2
3blha	Cell division protein kinase 9
2jfla	STE20-like serine/threonine-protein kinase
1vjya	TGF-beta receptor type I
1fvra	Tyrosine protein kinase TIE-2
1nvra	Serine/threonine-protein kinase CHK-2
1uwha	B-RAF Serine/threonine-protein kinase
1xkka	Epidermal growth factor receptor
1s9ja	Mitogen-activated protein kinase 1

B.6 Additional figures

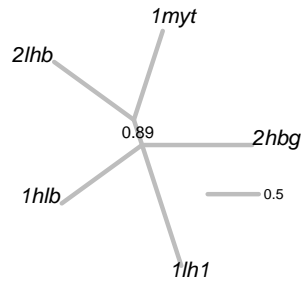


FIGURE B.1: Consensus tree for the 5-globin dataset, derived using BALi-Phy with default settings, running until convergence (10,000 iterations, roughly 30 minutes' runtime on a 2.13Ghz Intel core, with burn-in set to 365 as recommended by the `statreport` utility).

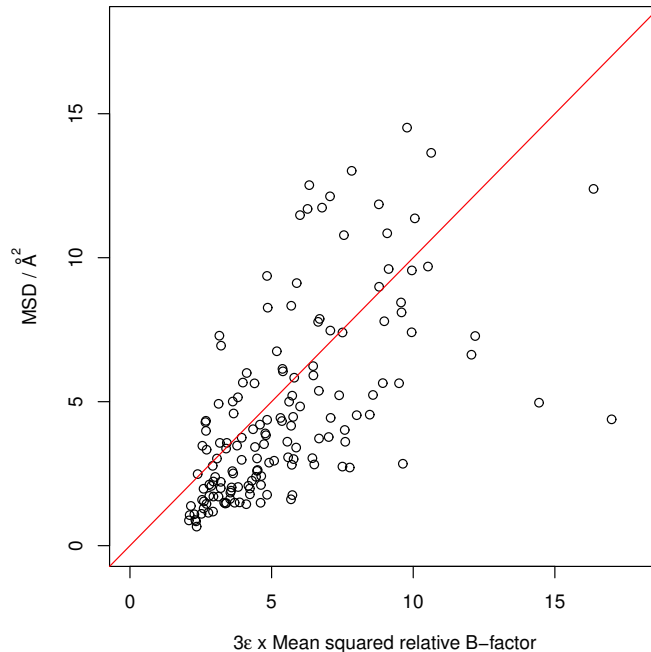


FIGURE B.2: Average pairwise mean squared deviation (MSD) for each column plotted against $3\epsilon_i$ [cf. equation (3.12) in the main text] for the maximum likelihood MCMC sample for the 12-globin set under the drift model, showing that for most columns the B -factor-derived information is a good predictor of the MSD (variance), which supports the use of B -factors as a measure of baseline variability. The multiplication by 3 is necessary because MSD contains a contribution from x, y and z . The surplus variability beyond the baseline is modelled by the diffusion component of the drift model.

B.7 Supplementary methods

B.7.1 Alignment accuracy and uncertainty

In order to measure the accuracy of a multiple sequence alignment (MSA), it is necessary to have a measure of distance between alignments. We utilise two measures here, namely the BALiBASE sum-of-pairs score Thompson et al. (1999), which is a measure of correct pairwise homology statements, and the *column score*, which is widely used in benchmarking sequence alignment algorithms.

The sum-of-pairs score is defined for any two alignments T (true) and P (pre-

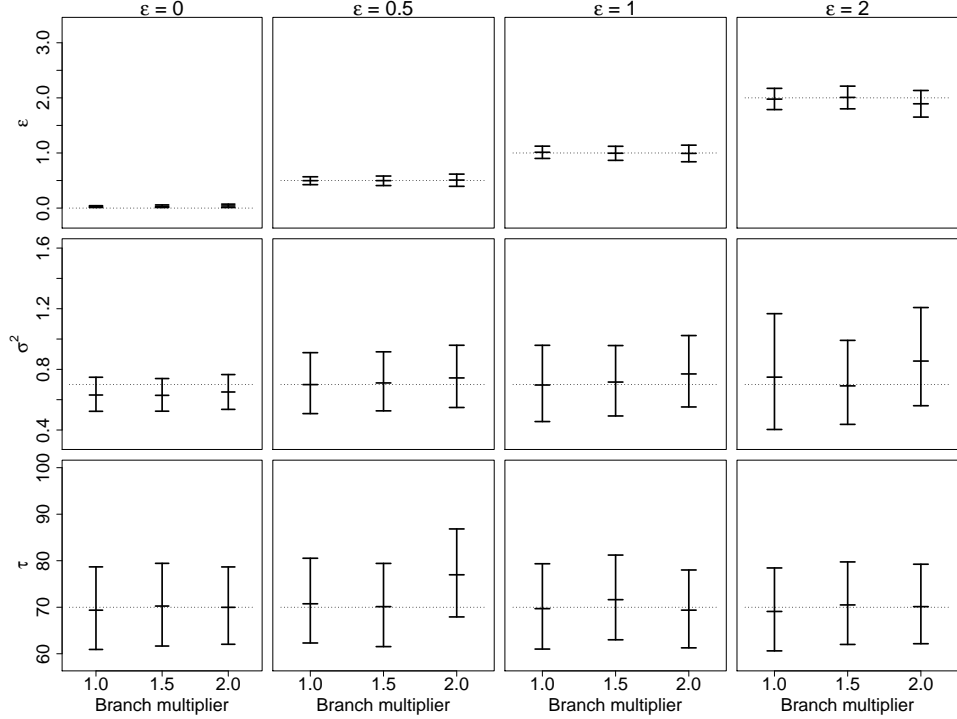


FIGURE B.3: 95% highest posterior density intervals for structural model parameters estimated on simulated data, on a 4-leaf tree.

dicted) as

$$\text{score}_{\text{pair}} = \frac{|h_H(T) \cap h_H(P)|}{L_T} \quad (\text{B.1})$$

where L_T is the length of the true alignment T , and $h_H(X)$ is the set of pairwise homology statements implied by alignment X . This score is effectively a measure of recall.

The column score provides a more global measure of agreement, and is defined in terms of the proportion of columns containing the same characters in both alignments

$$\text{score}_{\text{col}} = \frac{1}{L_T} \sum_{c \in T} \mathbb{1}(c \in P) \quad (\text{B.2})$$

where c denotes an alignment column, specified as an ordered n -tuple containing the indices of the characters from each sequence that are aligned to the column, such that n is equal to the number of non-gap characters in the column.

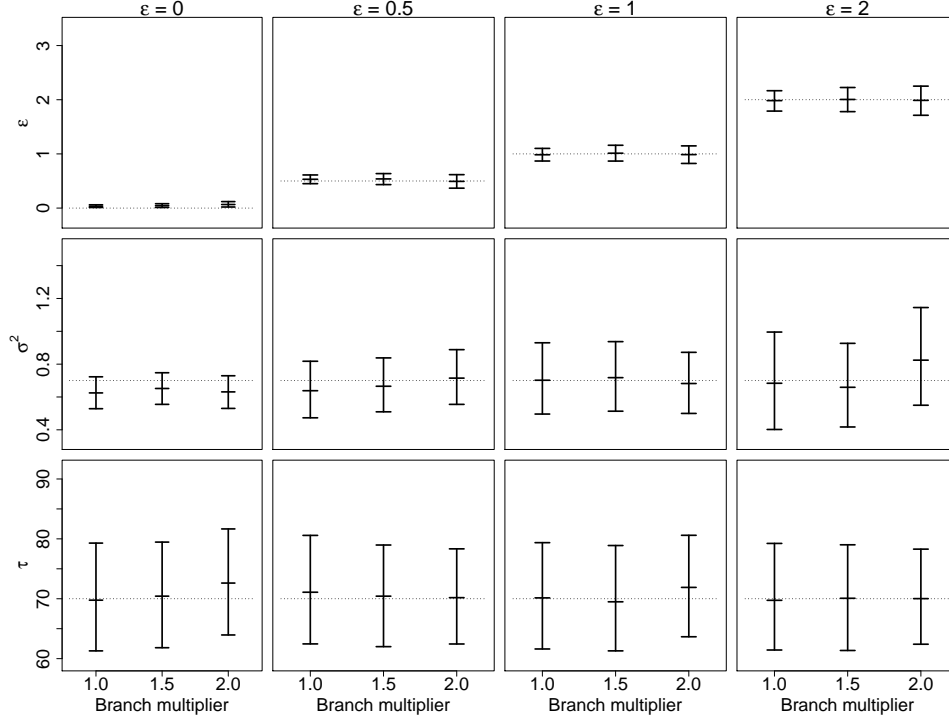


FIGURE B.4: 95% highest posterior density intervals for structural model parameters estimated on simulated data, on a 8-leaf tree.

B.7.2 Linear relationship between structure and branch length in global- σ model

For a single σ^2 parameter over the whole tree, the expected mean-square-deviation (MSD) is

$$\begin{aligned}
 \frac{1}{n} \sum_{ij} \mathbb{E}[(C_{ij}^{(t)} - C_{ij}^{(0)})^2 \mid C_{ij}^{(0)}] &= \frac{1}{n} \sum_{ij} \left((1 - e^{-\theta t}) C_{ij}^{(0)} \right)^2 + \frac{\sigma^2}{2\theta} (1 - e^{-2\theta t}) \\
 &\approx \frac{1}{n} \sum_{ij} (\theta t C_{ij}^{(0)})^2 + \sigma^2 \left(t - \frac{\theta t^2}{2} \right) \\
 &\approx \theta t^2 \sigma^2 + \sigma^2 t \\
 &\approx \sigma^2 t
 \end{aligned}$$

where the first approximation results from $1 - e^{-\theta t} \approx \theta t$, the second follows the relationships $\frac{1}{3n} \sum (C_{ij}^{(0)})^2 \approx \tau^2$ and $\tau^2 = \sigma^2/2\theta$, and the third from $\theta \ll \sigma^2$.

It should be noted that this expected linear relationship between MSD and branch

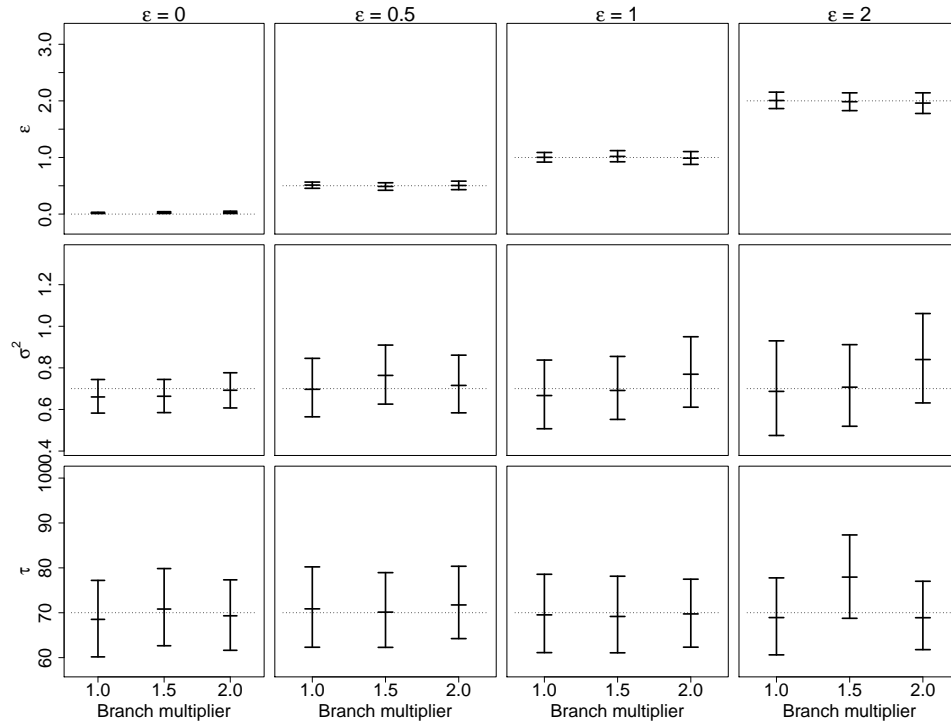


FIGURE B.5: 95% highest posterior density intervals for structural model parameters estimated on simulated data, on a 10-leaf tree.

length holds in a structure-only model; when combined with the sequence model, different relationships may be observed, since sequence information will also affect the estimation of the branch lengths.

Appendix C

Proof of Cutoff

We give an alternative proof (to Barrera et al. (2006)) of the existence of a cutoff for a class of irreducible independent product chains in continuous and discrete time. Proofs of the cutoff phenomenon have been closely linked with representation theory of finite groups in their development (Diaconis, 1988). The present work does not take a group theoretic approach, but still seeks to bound the total variation distance to equilibrium through eigenvalue decomposition, Hellinger distance, or direct calculation when possible. There are several definitions and conditions to establish. We consider an arbitrary sequence of matrices Q_1, \dots, Q_n , each of which is a rate matrix for a continuous-time Markov chain operating on finite space \mathcal{M}_i . The majority of previous cutoff treatments are for symmetric cases where all initial vectors are equivalent with respect to convergence to stationarity (Saloff-Coste, 2004). In the current setting of general Markov chains, it is necessary to take more care with initial conditions.

We make the following definitions with respect to a particular matrix Q . These definitions lead to identification of λ_i^* , the crucial eigenvalue component for state

$i \in \mathcal{M}$, the state space of Q . The matrix Q has eigenvalues $\Lambda = \{\lambda_l = a_l + b_l i\}$ and eigenmatrix V containing corresponding eigenvectors V_l . Partition the indices into ordered subsets Λ_k such that $i, j \in \Lambda_k \Leftrightarrow a_i = a_j$ and $i \in \Lambda_k, j \in \Lambda_l, k < l \Rightarrow a_i > a_j$. This groups and orders the eigenvalues according to their real components. Let

$$c_{ijk} = \sum_{l \in \Lambda_k} V_{il}(V^{-1})_{lj}(\cos(b_l) + i \sin(b_l))$$

This is the coefficient of $e^{a_k t}$ in P_{ij}^t . The convergence of state i to equilibrium is affected by each entry in the row $P_{i\cdot}^t$, so we define

$$k_i = \min\{k > 1 : \exists j \text{ such that } c_{ijk} \neq 0\}$$

which is the index of the largest eigenvalue important to state i . Finally, set

$$\lambda_i^* = -a_{k_i}$$

Informally, λ_i^* corresponds to the largest eigenvalue that affects convergence of the process dictated by Q and beginning in state $i \in \mathcal{M}$. We take the negative in the definition of λ_i^* to arrive at a positive value. We can now easily generalize the definition of λ_i^* to the case with many rate matrices Q_1, \dots, Q_n by writing $\lambda_{x_j}^*$ where $x_j \in \mathcal{M}_1 \cup \mathcal{M}_2 \cup \dots \cup \mathcal{M}_n$. We define the following functions which relate to the “effective multiplicity” of eigenvalues in the product chain given a particular initial sequence \mathbf{x}_0 .

$$f_\lambda(n) = \#\{j : \lambda_{x_j}^* = \lambda, j \leq n\}$$

$$f_{\lambda-}(n) = \#\{j : \lambda_{x_j}^* < \lambda, j \leq n\}$$

$$g(n) = \min\{\lambda_{x_j}^* : j \leq n\}$$

Here $f_\lambda(n)$ is the number of times that λ is the crucial eigenvalue of the first n positions, while $f_{\lambda-}(n)$ is the number of times that the crucial eigenvalue is larger than λ . The function $g(n)$ tracks the largest dominant eigenvalue of any of the first n positions.

Theorem 3. *Given a sequence of irreducible, continuous-time Markov processes on finite spaces \mathcal{M}_i with transition rate matrices Q_i , define $\mathcal{M}^n = \mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \cdots \otimes \mathcal{M}_n$ and $P_{\mathbf{x}_0}^n$ the distribution arising from application of Q_i to individual positions of $\mathbf{x}_0 \in \mathcal{M}^n$. If there exists λ^* such that $\forall \epsilon > 0$*

$$\lim_{n \rightarrow \infty} \frac{f_{\lambda^*}(n)}{n^{1-\epsilon}} = \infty \quad (\text{C.1})$$

$$\lim_{n \rightarrow \infty} \frac{f_{\lambda^*}(n)}{n^{(1+\epsilon)g(n)/\lambda^*}} = 0, \quad (\text{C.2})$$

the family $(\mathcal{M}^n, P_{\mathbf{x}_0}^n)$ has a total variation cutoff at $\tau_n = \frac{1}{2\lambda^*} \log(n)$.

Proof. We work at first with a single matrix Q in order to simplify notation, and later generalize to an arbitrary sequence of matrices. Using the spectral decomposition of Q , the elements of the transition matrix P^t can be expressed as linear combinations of the exponentiated eigenvalues of tQ

$$P_{ij}^t = \pi_j + \sum_{k=2}^m V_{ik}(V^{-1})_{kj} e^{t\lambda_k}$$

Every element of the matrix experiences exponential decay toward the corresponding entry of π . This sum can be rewritten in terms of the $m_u \leq m - 1$ unique real components of the λ_i .

$$P_{ij}^t = \pi_j + \sum_{k=1}^{m_u} c_{ijk} e^{ta_k} \quad (\text{C.3})$$

where we have grouped the eigenvector coefficients and imaginary components of e^{λ_k} into new constants c_{ijk} . Note that the c_{ijk} are the same as those used in the definition of λ^* , and that these values are real because the λ_k occur in conjugate pairs. The Hellinger distance can be used to bound the total variation distance

via the inequalities $H(p, q)^2 \leq \|p - q\|_{TV} \leq 2H(p, q)$. This bound is particularly useful for working with product spaces, because $H(p_1 \otimes \cdots \otimes p_n, q_1 \otimes \cdots \otimes q_n)^2 = 1 - \prod(1 - \frac{1}{2}H(p_i, q_i)^2)$. We bound the Hellinger distance on the product space by considering the “worst case” distance for a single position. The squared Hellinger distance for a single position beginning in state i is

$$H_i^2 \triangleq H^2(P_i^t, \pi) = \sum_j \left(\sqrt{P_{ij}^t} - \sqrt{\pi_j} \right)^2$$

Recall that λ_i^* is the largest real component of an eigenvalue with nonzero coefficient in (C.3) and that k_i is the index of the unique eigenvalue. Let $c_{ij} = c_{ijk_i}$ for readability. We find a lower bound on H_i^2 with the observation that for some time T , $\frac{3|c_{ij}|}{2}e^{\lambda_i^*t} \geq |P_{ij}^t - \pi_j| \geq \frac{|c_{ij}|}{2}e^{\lambda_i^*t}$, $\forall t > T$. (The largest eigenvalue begins to dominate and the effect of the others can be at most $\frac{|c_{ij}|}{2}e^{\lambda_i^*t}$). Then we have the bound

$$H_i^2 \geq \sum_j \left(\sqrt{\pi_j + \frac{|c_{ij}|}{2}e^{\lambda_i^*t}} - \sqrt{\pi_j} \right)^2$$

where we add the $|c_{ij}|$ term because $(\sqrt{a+b} - \sqrt{a})^2 \leq (\sqrt{a-b} - \sqrt{a})^2$ when $a > b > 0$. We then substitute $t = \tau_n^- \triangleq (1 - \epsilon)\tau_n$ and take a Taylor expansion of the square root:

$$= \sum_j \frac{c_{ij}^2 n^{-(1-\epsilon)\lambda_i^*/\lambda^*}}{16\pi_j} \pm O(n^{-3(1-\epsilon)\lambda_i^*/2\lambda^*}) = O(n^{-(1-\epsilon)\lambda_i^*/\lambda^*})$$

This gives us a lower bound on the Hellinger distance for each state i with respect to a specific matrix Q . We can now again generalize from a state i in Q to a state x_j in the initial sequence \mathbf{x}_0 composed of many state spaces. With a lower bound on the Hellinger distance from each state x_j , the Hellinger distance on the product space can then be bounded by

$$H^2(P_{\mathbf{x}_0}^{\tau_n^-}, \pi^{(n)}) \geq 1 - \prod_{j=1}^n \left(1 - O(n^{-(1-\epsilon)\lambda_{x_j}^*/\lambda^*})\right)$$

Clearly any sub-product falls in the interval $[0,1]$ in the limit so we need only consider the terms where $\lambda_{x_j}^* = \lambda^*$.

$$\begin{aligned} & \lim_{n \rightarrow \infty} \prod_{j:\lambda_{x_j}^*=\lambda^*} \left(1 - O(n^{-(1-\epsilon)\lambda_{x_j}^*/\lambda^*})\right) \\ &= \lim_{n \rightarrow \infty} \left(1 - O(n^{-(1-\epsilon)})\right)^{f_{\lambda^*}(n)} \\ &= \exp \left\{ - \lim_{n \rightarrow \infty} \frac{f_{\lambda^*}}{O(n^{1-\epsilon})} \right\} = 0 \end{aligned}$$

The limit of the entire product is then 0, and so we have

$$\lim_{n \rightarrow \infty} \|P_{\mathbf{x}_0}^{\tau_n^-} - \pi^{(n)}\|_{TV} \geq H^2(P_{\mathbf{x}_0}^{\tau_n^-}, \pi^{(n)}) \geq 1.$$

This completes the argument for the ‘backward’ side of the proof. The argument for the ‘forward’ side works similarly, using instead $\frac{3|c_{ij}|}{2}e^{\lambda_i^* t}$ to find an upper bound.

With $t = \tau_n^+ \triangleq (1 + \epsilon)\tau_n$, we then arrive at

$$H_i^2 \leq O(n^{-(1+\epsilon)\lambda_i^*/\lambda^*})$$

Again generalizing to an arbitrary sequence of matrices, the bound on the product space is

$$H^2(P_{\mathbf{x}_0}^{\tau_n^-}, \pi^{(n)}) \leq 1 - \prod_{j=1}^n \left(1 - O(n^{-(1+\epsilon)\lambda_{x_j}^*/\lambda^*})\right).$$

In this case we split the product into

$$\lim_{n \rightarrow \infty} \prod_{j:\lambda_{x_j}^* < \lambda^*} \left(1 - O(n^{-(1+\epsilon)\lambda_{x_j}^*/\lambda^*})\right) \prod_{j:\lambda_{x_j}^* \geq \lambda^*} \left(1 - O(n^{-(1+\epsilon)\lambda_{x_j}^*/\lambda^*})\right)$$

and require both products to converge to 1 in order to force H^2 to 0. The product on the right is always 1 unless the number of terms in the product grows faster than n , which cannot be the case. For the left product we create a bound with $g(n)$ (defined previously as the minimal $\lambda_{x_j}^*$ of the first n) to obtain:

$$\begin{aligned} \lim_{n \rightarrow \infty} \prod_{i: \lambda_{x_i}^* < \lambda^*} \left(1 - O(n^{-(1+\epsilon)\lambda_{x_i}^*/\lambda^*})\right) &\geq \lim_{n \rightarrow \infty} \left(1 - O(n^{-(1+\epsilon)g(n)/\lambda^*})\right)^{f_{\lambda_-^*}(n)} \\ &= \exp \left\{ - \lim_{n \rightarrow \infty} \frac{f_{\lambda_-^*}(n)}{O(n^{-(1+\epsilon)g(n)/\lambda^*})} \right\} \\ &= 1 \end{aligned}$$

The bound holds by replacing all $\lambda_{x_i}^* < \lambda^*$ with the minimum, $g(n)$, and the limit within the exponent is 0 by hypothesis. Thus we have shown that

$$\lim_{n \rightarrow \infty} \|P_{\mathbf{x}_0}^{\tau_n^+} - \pi^{(n)}\|_{TV} \leq 2H(P_{\mathbf{x}_0}^{\tau_n^+}, \pi^{(n)}) \leq 0$$

□

We provide a few examples to give a better sense of the meaning of the conditions required by the theorem.

Example 1. For sequences Q_1, \dots, Q_n which do not repeat eigenvalues, we cannot prove the existence of a cutoff. For example if Q_m is standard symmetric evolution on $m + 1$ characters, no $\lambda_{x_j}^*$ is repeated so (C.1) of Theorem 3 cannot be satisfied.

Example 2. When the matrices Q_i are chosen from some finite set of matrices \mathcal{Q} we have a finite set of eigenvalues, thus there must always be at least one satisfying (C.1), so it only remains to verify whether the largest of these eigenvalues also satisfies (C.2), which will depend upon the initial sequence \mathbf{x}_0 . We consider the especially simple case where all Q_i are the modified Jukes-Cantor model introduced previously, with $\delta > 1$. Then $\lambda_i^* = 4/3$ for $i = 1, 2, 3$ and $\lambda_4^* = \delta + 1/3$. We can then design an

initial sequence consisting only of states 1 and 4 where state 1 does not occur often enough to ensure a cutoff for the value $4/3$, but is present enough that we cannot guarantee a cutoff for $\delta + 1/3$. For example, if state 1 occurs with frequency $\lfloor n^{\frac{\delta+5/3}{2(\delta+1/3)}} \rfloor$ and all other positions are state 4, $4/3$ does not satisfy (C.1) while $\delta + 1/3$ does, but does not satisfy (C.2).

C.0.3 Discrete Time

The discrete-time case has a few additional complications so we require a few changes to conditions and definitions in order to follow the same proof. Firstly, we require discrete-time transition matrices to be aperiodic. The other issue arises with zero eigenvalues. In the continuous case only the first eigenvalue is zero and all others must have negative real component, but in discrete-time all eigenvalues but the first may be zero. In most cases a zero eigenvalue is no problem because a different eigenvalue will dominate the rate of convergence. However, when a row P_i is exactly equal to π , $c_{ijk}a_k = 0 \forall j, k$. This requires the following changes in the definition of λ_i^* :

$$c_{ijk} = \sum_{l \in \Lambda_k} V_{il}(V^{-1})_{lj}(\cos(b_l) + i\sin(b_l))$$

$$k_i = \min\{k : \exists j \text{ such that } c_{ijk} \neq 0 \text{ and } a_k \neq 0\}$$

In the definition of k_i we omit the requirement that $k > 1$ because all eigenvalues beyond the first may be zero. The definition of λ_i^* is also altered to include this case:

$$\lambda_i^* = \begin{cases} -\log(a_{k_i}) & k_i > 1 \\ \infty & \text{otherwise} \end{cases}$$

Finally, we restrict $\lambda^* < \infty$. With these definitions, the argument carries forward in exactly the same manner and under the same conditions, proving a cutoff at $\frac{1}{2\lambda^*} \log(n)$.

Bibliography

- Aldous, D. (1983), “Random walks on finite groups and rapidly mixing Markov chains,” in *Séminaire de Probabilités, (Strasbourg)*, vol. 17, pp. 243 – 297, Springer, Berlin.
- Aldous, D. and Fill, J. (2002), *Reversible Markov chains and random walks on groups (in preparation)*, Online version available at <http://www.stat.berkeley.edu/~aldous/RWG/book.html>.
- Aris-Brosou, S. and Yang, Z. (2002), “Effects of Models of Rate Evolution on Estimation of Divergence Dates with Special Reference to the Metazoan 18S Ribosomal RNA Phylogeny,” *Systematic Biology*, 51, 703–714.
- Barrera, J., Lachud, B., and Ycart, B. (2006), “Cut-off for n-tuples of exponentially converging processes,” *Stochastic Processes and their Applications*, 116, 1433–1446.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009), “GenBank,” *Nucleic Acids Res.*, 37, D26–31.
- Bishop, M. J. and Thompson, E. A. (1986), “Maximum likelihood alignment of DNA sequences,” *J. Mol. Biol.*, 190, 159–65.
- Blackburne, B. P. and Whelan, S. (2013), “Class of Multiple Sequence Alignment Algorithm Affects Genomic Analysis,” *Molecular Biology and Evolution*, 30, 642–653.
- Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. (2006), “Structural Determinants of the Rate of Protein Evolution in Yeast,” *Molecular Biology and Evolution*, 23, 1751–1761.
- Bouchard-Côté, A. and Jordan, M. I. (2013), “Evolutionary inference via the Poisson Indel Process,” *Proceedings of the National Academy of Sciences*, 110, 1160–1166.
- Brown, J. M., Hedtke, S. M., Lemmon, A. R., and Lemmon, E. M. (2010), “When Trees Grow Too Long: Investigating the Causes of Highly Inaccurate Bayesian Branch-Length Estimates,” *Systematic Biology*, 59, 145–161.

- Bujnicki, J. M. (2000), “Phylogeny of the Restriction Endonuclease-Like Superfamily Inferred from Comparison of Protein Structures,” *Journal of Molecular Evolution*, 50, 39–44.
- Burmester, T., Weich, B., Reinhardt, S., and Hankeln, T. (2000), “A vertebrate globin expressed in the brain,” *Nature*, 407, 520–523.
- Burmester, T., Ebner, B., Weich, B., and Hankeln, T. (2002), “Cytoglobin: A Novel Globin Type Ubiquitously Expressed in Vertebrate Tissues,” *Molecular Biology and Evolution*, 19, 416–421.
- Cavender, J. A. (1978), “Taxonomy with confidence,” *Math. Biosci.*, 40, 271–280.
- Challis, C. J. and Schmidler, S. C. (2012), “A Stochastic Evolutionary Model for Protein Structure Alignment and Phylogeny,” *Molecular Biology and Evolution*, 29, 3575 – 3587.
- Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007), “Quantifying the Impact of Protein Tertiary Structure on Molecular Evolution,” *Molecular Biology and Evolution*, 24, 1769–1782.
- Chothia, C. and Lesk, A. M. (1986), “The relationship between the divergence of sequence and structure in proteins,” *EMBO J.*, 5, 823–826.
- Cruickshank, D. W. J. (1960), “The required precision of intensity measurements for single-crystal analysis,” *Acta Crystallographica*, 13, 774–777.
- Cruickshank, D. W. J. (1999), “Remarks about protein structure precision,” *Acta Crystallographica Section D*, 55, 583–601.
- Daskalakis, C., Mossel, E., and Roch, S. (2011), “Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel’s conjecture,” *Probability Theory and Related Fields*, 149, 149–189.
- Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978), “A model of evolutionary change in proteins,” *Atlas of protein sequence and structure*, 5, 345–351.
- DePristo, M. A., de Bakker, P. I., and Blundell, T. L. (2004), “Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography,” *Structure*, 12, 831 – 838.
- Dessimoz, C. and Gil, M. (2010), “Phylogenetic assessment of alignments reveals neglected tree signal in gaps,” *Genome Biology*, 11, R37.
- Diaconis, P. (1988), *Group representations in probability and statistics*, Institute of Mathematical Statistics, Hayward, California.

- Diaconis, P. (1996), “The cutoff phenomenon in finite Markov chains,” *Proc. Natl. Acad. Sci.*, 93, 1659 – 1664.
- Diaconis, P. and Shahshahani, M. (1981), “Generating a random permutation with random transpositions,” *Z. Wahrsch. Verw. Geb.*, 57, 159–179.
- Diaconis, P. and Shahshahani, M. (1987), “Time to reach stationarity in the Bernoulli-Laplace diffusion model,” *SIAM J. Math Anal.*, 18, 208 – 218.
- Diaconis, P. and Stroock, D. (1991), “Geometric bounds for eigenvalues of Markov chains,” *Ann. Appl. Probab.*, 1, 36–61.
- Diaconis, P., Graham, R. L., and Morrison, J. A. (1990), “Asymptotic Analysis of a Random Walk on a Hypercube with Many Dimensions,” *Random Structures and Algorithms*, 1, 51 – 72.
- Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2013), “Bayesian phylogenetics with BEAUti and the BEAST 1.7,” *Mol. Biol. Evol.*, In press.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998), *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, Cambridge.
- Dutheil, J. Y., Galtier, N., Romiguier, J., Douzery, E. J., Ranwez, V., and Boussau, B. (2012), “Efficient Selection of Branch-Specific Models of Sequence Evolution,” *Molecular Biology and Evolution*, 29, 1861–1874.
- Ebner, B., Panopoulou, G., Vinogradov, S., Kiger, L., Marden, M., Burmester, T., and Hankeln, T. (2010), “The globin gene family of the cephalochordate amphioxus: implications for chordate globin evolution,” *BMC Evolutionary Biology*, 10, 370.
- Eidhammer, I., Jonassen, I., and Taylor, W. (2000), “Structure comparison and structure patterns,” *J. Comp. Biol.*, 7, 685–716.
- Ekman, S. and Blaalid, R. (2011), “The Devil in the Details: Interactions between the Branch-Length Prior and Likelihood Model Affect Node Support and Branch Lengths in the Phylogeny of the Psoraceae,” *Systematic Biology*, 60, 541–561.
- England, J. L. and Shakhnovich, E. I. (2003), “Structural Determinant of Protein Designability,” *Phys. Rev. Lett.*, 90, 218101.
- Farris, J. S. (1973), “A probability model for inferring evolutionary trees,” *Syst. Zool.*, 22, 250–256.
- Felsenstein, J. (1981), “Evolutionary trees from DNA sequences: A maximum likelihood approach,” *Journal of Molecular Evolution*, 17, 368–376.

- Felsenstein, J. (2003), *Inferring phylogenies*, Sinauer Associates.
- Garau, G., Di Guilmi, A. M., and Hall, B. G. (2005), “Structure-Based Phylogeny of the Metallo-Lactamases,” *Antimicrobial Agents and Chemotherapy*, 49, 2778–2784.
- Garrocho-Villegas, V., Gopalasubramaniam, S. K., and Arredondo-Peter, R. (2007), “Plant hemoglobins: What we know six decades after their discovery,” *Gene*, 398, 78 – 85, [Devoted to the {XIV} International Conference on Dioxygen Binding and Sensing Proteins](#).
- Gelman, A. and Rubin, D. B. (1992), “Inference from iterative simulation using multiple sequences,” *Statist. Sci.*, 7, 457–511.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis, Second Edition*, Chapman & Hall/CRC.
- Goodall, C. R. and Mardia, K. V. (1993), “Multivariate Aspects of Shape Theory,” *The Annals of Statistics*, 21, pp. 848–866.
- Gopalasubramaniam, S. K., Kovacs, F., Violante-Mota, F., Twigg, P., Arredondo-Peter, R., and Sarath, G. (2008), “Cloning and characterization of a caesalpinoid (*Chamaecrista fasciculata*) hemoglobin: The structural transition from a nonsymbiotic hemoglobin to a leghemoglobin,” *Proteins: Structure, Function, and Bioinformatics*, 72, 252–260.
- Green, P. J. and Mardia, K. V. (2006), “Bayesian alignment using hierarchical models, with applications in protein bioinformatics,” *Biometrika*, 93, 235–254.
- Green, P. J., Mardia, K. V., Nyirongo, V. B., and Ruffieux, Y. (2010), *Bayesian modelling for matching and alignment of biomolecules*, pp. 27–50, The Oxford Handbook of Applied Bayesian Analysis, Oxford University Press, Oxford.
- Griffiths, A., Gelbart, W., and Miller, J. (1999), *Modern Genetic Analysis*, W. H. Freeman, New York.
- Grishin, N. V. (1997), “Estimation of evolutionary distances from protein spatial structures,” *J. Mol. Evol.*, 45, 359–369.
- Groussin, M., Boussau, B., and Gouy, M. (2013), “A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences,” *Systematic Biology*, 62, 523–538.
- Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., and O., G. (2010), “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.” *Systematic Biology*, 59, 307–21.

- Gunçar, G., Podobnik, M., e Punger ar, J., trukelj, B., Turk, V., and an Turk, D. (1998), “Crystal structure of porcine cathepsin H determined at 2.1 resolution: location of the mini-chain C-terminal carboxyl group defines cathepsin H aminopeptidase function,” *Structure*, 6, 51 – 61.
- Gutin, A. M. and Badretdinov, A. Y. (1994), “Evolution of Protein 3D Structures as Diffusion in Multidimensional Conformational Space,” *J. Mol. Evol.*, 39, 206–209.
- Hansen, T. F. and Martins, E. P. (1996), “Translating Between Microevolutionary Process and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data,” *Evolution*, 50, 1404–1417.
- Hasegawa, H. and Holm, L. (2009), “Advances and pitfalls of protein structural alignment,” *Curr. Opin. Struct. Biol.*, 19, 341–8.
- Hastings, W. K. (1970), “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, 57, 97–109.
- Hein, J., Wifu, C., Knudsen, B., Møller, M. B., and Wibling, G. (2000), “Statistical alignment: Computational Properties, Homology Testing and Goodness-of-Fit,” *J. Mol. Biol.*, 302, 265–279.
- Herman, J. L., Novák, A., Lyngsø, R., Miklós, I., and Hein, J. (2013), “Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs,” *Submitted*.
- Hoffmann, F. G., Opazo, J. C., and Storz, J. F. (2010), “Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates,” *Proceedings of the National Academy of Sciences*, 107, 14274–14279.
- Hoffmann, F. G., Opazo, J. C., Hoogewijs, D., Hankeln, T., Ebner, B., Vinogradov, S. N., Bailly, X., and Storz, J. F. (2012a), “Evolution of the Globin Gene Family in Deuterostomes: Lineage-Specific Patterns of Diversification and Attrition,” *Molecular Biology and Evolution*, 29, 1735–1745.
- Hoffmann, F. G., Opazo, J. C., and Storz, J. F. (2012b), “Whole-Genome Duplications Spurred the Functional Diversification of the Globin Gene Superfamily in Vertebrates,” *Molecular Biology and Evolution*, 29, 303–312.
- Holder, M. T., Lewis, P. O., Swofford, D. L., and Larget, B. (2005), “Hastings Ratio of the LOCAL Proposal Used in Bayesian Phylogenetics,” *Systematic Biology*, 54, 961–965.
- Holmes, I. and Bruno, W. J. (2001), “Evolutionary HMMs: a Bayesian approach to multiple alignment,” *Bioinformatics*, 17, 803–820.

- Howe, K., Bateman, A., and Durbin, R. (2002), “QuickTree: building huge neighbor-joining trees of protein sequences,” *Bioinformatics*, 18, 1546–1547.
- Hoy, J. A., Robinson, H., III, J. T. T., Kakar, S., Smagghe, B. J., and Hargrove, M. S. (2007), “Plant Hemoglobins: A Molecular Fossil Record for the Evolution of Oxygen Transport,” *Journal of Molecular Biology*, 371, 168 – 179.
- Huelsenbeck, J. P. and Ronquist, F. (2001), “MrBayes: Bayesian inference in phylogenetic trees,” *Bioinformatics*, 17, 754–55.
- Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002), “Potential applications and pitfalls of Bayesian inference of phylogeny,” *Syst. Biol.*, 51, 673–688.
- Hughes, A. L. (1994), “The Evolution of Functionally Novel Proteins after Gene Duplication,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 256, 119–124.
- Illergård, K., Ardell, D. H., and Elofsson, A. (2009), “Structure is three to ten times more conserved than sequence: A study of structural response in protein cores,” *Proteins: Structure, Function, and Bioinformatics*, 77, 499–508.
- Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J. M., Gaucher, E. A., Sanchez-Ruiz, J. M., and Gavira, J. A. (2013), “Conservation of Protein Structure over Four Billion Years,” *Structure*, 21, 1690 – 1697.
- Johnson, M. S., Sali, A., and Blundell, T. L. (1990), “[42] Phylogenetic relationships from three-dimensional protein structures,” vol. 183 of *Methods in Enzymology*, pp. 670 – 690, Academic Press.
- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992), “The rapid generation of mutation data matrices from protein sequences,” *CABIOS*, 8, 275–282.
- Jukes, T. H. and Cantor, C. R. (1969), “Evolution of protein molecules,” in *Mammalian protein metabolism, III*, pp. 21–132, Academic Press, New York.
- Karlin, S. and Taylor, H. M. (1981), *A second course in stochastic processes*, Academic Press, San Diego.
- Katoh, K. and Standley, D. M. (2013), “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability,” *Molecular Biology and Evolution*, 30, 772–780.
- Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005), “MAFFT version 5: improvement in accuracy of multiple sequence alignment,” *Nucleic Acids Res.*, 33, 511–518.

- Kimura, M. (1980), “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences,” *Journal of Molecular Evolution*, 16, 111–120.
- Kleinman, C. L., Rodrigue, N., Lartillot, N., and Philippe, H. (2010), “Statistical Potentials for Improved Structurally Constrained Evolutionary Models,” *Molecular Biology and Evolution*, 27, 1546–1560.
- Knudsen, B. and Miyamoto, M. M. (2003), “Sequence alignments and pair hidden Markov models using evolutionary history,” *J. Mol. Biol.*, 333, 453–460.
- Kolaczkowski, B. and Thornton, J. W. (2007), “Effects of Branch Length Uncertainty on Bayesian Posterior Probabilities for Phylogenetic Hypotheses,” *Molecular Biology and Evolution*, 24, 2108–2118.
- Kosiol, C. and Goldman, N. (2005), “Different Versions of the Dayhoff Rate Matrix,” *Mol. Biol. Evol.*, 22, 193–199.
- Krissinel, E. (2007), “On the relationship between sequence and structure similarities in proteomics,” *Bioinformatics*, 23, 717–723.
- Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., and Tamura, K. (2012), “Statistics and Truth in Phylogenomics,” *Molecular Biology and Evolution*, 29, 457–472.
- Lake, J. A. (1991), “The order of sequence alignment can bias the selection of tree topology,” *Mol. Biol. Evol.*, 8, 378–385.
- Lamy, J. N., Green, B. N., Toulmond, A., Wall, J. S., Weber, R. E., and Vinogradov, S. N. (1996), “Giant Hexagonal Bilayer Hemoglobins,” *Chemical Reviews*, 96, 3113–3124.
- Landsmann, J., Dennis, E. S., Higgins, T. J. V., Appleby, C. A., Kortt, A. A., and Peacock, W. J. (1986), “Common evolutionary origin of legume and non-legume plant haemoglobins,” *Nature*, 324, 166–168.
- Larget, B. and Simon, D. (1999), “Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees,” *Molecular Biology and Evolution*, 16, 750.
- Levy, E. D., Boeri Erba, E., Robinson, C. V., and Teichmann, S. A. (2008), “Assembly reflects evolution of protein complexes.” *Nature*, 453, 1262–1265.
- Löytynoja, A. and Goldman, N. (2005), “An algorithm for progressive multiple alignment of sequences with insertions,” *Proceedings of the National Academy of Sciences of the United States of America*, 102, 10557–10562.

- Lukatsky, D., Shakhnovich, B., Mintseris, J., and Shakhnovich, E. (2007), “Structural Similarity Enhances Interaction Propensity of Proteins,” *Journal of Molecular Biology*, 365, 1596 – 1606.
- Lundin, D., Poole, A. M., Sjöberg, B.-M., and Hgbom, M. (2012), “Use of Structural Phylogenetic Networks for Classification of the Ferritin-like Superfamily,” *Journal of Biological Chemistry*, 287, 20565–20575.
- Lunter, G., Miklós, I., Drummond, A., Jensen, J., and Hein, J. (2003), “Bayesian Phylogenetic Inference under a Statistical Insertion-Deletion Model,” in *Algorithms in Bioinformatics*, eds. G. Benson and R. Page, vol. 2812 of *Lecture Notes in Computer Science*, pp. 228–244, Springer Berlin Heidelberg.
- Lunter, G., Drummond, A., Mikls, I., and Hein, J. (2005a), “Statistical Alignment: Recent Progress, New Applications, and Challenges,” in *Statistical Methods in Molecular Evolution*, Statistics for Biology and Health, pp. 375–405, Springer New York.
- Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., and Hein, J. (2008), “Uncertainty in homology inferences: Assessing and improving genomic sequence alignment,” *Genome Res.*, 18, 298–309.
- Lunter, G. A., Miklós, I., Drummond, A., Jensen, H. L., and Hein, J. L. (2005b), “Bayesian coestimation of phylogeny and sequence alignment,” *BMC Bioinformatics*, 6.
- Lunter, G. A., Miklós, I., Drummond, A., Jensen, H. L., and Hein, J. L. (2005c), “Bayesian coestimation of phylogeny and sequence alignment,” *BMC Bioinformatics*, 6.
- Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002), “The Protein Kinase Complement of the Human Genome,” *Science*, 298, 1912–1934.
- Mardia, K. V. and Jupp, P. E. (2000), *Directional Statistics*, Wiley.
- Marshall, D. C. (2010), “Cryptic Failure of Partitioned Bayesian Phylogenetic Analyses: Lost in the Land of Long Trees,” *Systematic Biology*, 59, 108–117.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), “Equation of State Calculations by Fast Computing Machines,” *J. Chem. Phys.*, 21, 1087–1092.
- Mihaescu, R. and Steel, M. (2010), “Logarithmic bounds on the posterior divergence time of two sequences,” *Applied Mathematics Letters*, 23, 975–979.

- Miklós, I., Lunter, G. A., and Holmes, I. (2004), “A “long indel” model for evolutionary sequence alignment,” *Mol. Biol. Evol.*, 21, 529–540.
- Miklós, I., Novák, A., Dombai, B., and Hein, J. (2008), “How reliably can we predict the reliability of protein structure predictions?” *BMC Bioinformatics*, 9.
- Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998), “HOMSTRAD: a database of protein structure alignments for homologous families,” *Prot. Sci.*, 7, 2469–2471.
- Morrison, D. A. and Ellis, J. T. (1997), “Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa,” *Mol. Biol. Evol.*, 14, 428–441.
- Mossel, E. (2003), “On the impossibility of reconstructing ancestral data and phylogenies,” *J. Comp. Biol.*, 10, 669–676.
- Mossel, E. (2004), “Phase transitions in phylogeny,” *Trans. Amer. Math. Soc.*, 356, 2379–2404.
- Mossel, E. (2011), “On the inference of large phylogenies with long branches: How long is too long?” *Bull. Math. Biol.*, 73, 1627–1644.
- Neyman, J. (1971), “Molecular studies of evolution: a source of novel statistical problems,” in *Statistical decision theory and related topics*, pp. 1–27, Academic Press, New York.
- Novák, A., Miklós, I., Lyngsø, R., and Hein, J. (2008), “StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees,” *Bioinformatics (Oxford, England)*, 24, 2403–2404, LR: 20091104; GR: BB/C509566/1/Biotechnology and Biological Sciences Research Council/United Kingdom; JID: 9808944; 2008/08/27 [aheadofprint]; ppublish.
- Page, R. D. M. (1996), “TREEVIEW: An application to display phylogenetic trees on personal computers,” *Comp. App. Biosci.*, 12, 357–358.
- Panchenko, A. R., Wolf, Y. I., Panchenko, L. A., and Madej, T. (2005), “Evolutionary plasticity of protein families: Coupling between sequence and structure variation,” *Proteins*, 61, 535–544.
- Rannala, B., Zhu, T., and Yang, Z. (2012), “Tail Paradox, Partial Identifiability, and Influential Priors in Bayesian Branch Length Inference,” *Molecular Biology and Evolution*, 29, 325–335.
- Rastogi, S. and Liberles, D. (2005), “Subfunctionalization of duplicated genes as a transition state to neofunctionalization,” *BMC Evolutionary Biology*, 5, 28.

- Redelings, B. D. and Suchard, M. A. (2005), “Joint Bayesian Estimation of Alignment and Phylogeny,” *Syst. Biol.*, 54, 401–418.
- Robinson, D. and Foulds, L. (1981), “Comparison of phylogenetic trees,” *Mathematical Biosciences*, 53, 131 – 147.
- Robinson, D., Jones, D., Kishino, H., Goldman, N., and Thorne, J. (2003), “Protein evolution with dependence among codons due to tertiary structure,” *Mol. Biol. Evol.*, 20, 1692–1704.
- Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. (2005), “Site interdependence attributed to tertiary structure in amino acid sequence evolution,” *Gene*, 347, 207 – 217.
- Rodrigue, N., Kleinman, C. L., Philippe, H., and Lartillot, N. (2009), “Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons,” *Mol. Biol. Evol.*, 26, 1663 – 76.
- Rodriguez, A. and Schmidler, S. C. (2013), “Bayesian Protein Structure Alignment,” (*submitted*).
- Ronquist, F. and Huelsenbeck, J. P. (2003), “MrBayes 3: Bayesian phylogenetic inference under mixed models,” *Bioinformatics*, 19, 1572–1574.
- Rost, B. (1999), “Twilight zone of protein sequence alignments,” *Protein Eng.*, 12, 85–94.
- Rueda, M., Ferrer-Costa, C., Meyer, T., Prez, A., Camps, J., Hospital, A., Gelp, J. L., and Orozco, M. (2007), “A consensus view of protein dynamics,” *Proceedings of the National Academy of Sciences*, 104, 796–801.
- Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A., and Sternberg, M. J. (1997), “Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation,” *Journal of Molecular Biology*, 269, 423 – 439.
- Saitou, N. and Nei, M. (1987), “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” *Mol. Biol. Evol.*, 4, 406–425.
- Saloff-Coste, L. (2004), “Random walks on finite groups,” in *Probability on Discrete Structures, Encyclopaedia of Mathematical Sciences, Vol. 110*, pp. 263 – 346, Springer, Berlin.
- Satija, R., Novak, A., Miklos, I., Lyngso, R., and Hein, J. (2009), “BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC,” *BMC Evolutionary Biology*, 9, 217, M3: 10.1186/1471-2148-9-217.

- Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrach, I. I., Ostel, I. J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009), “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, 37, D5–15.
- Schmidler, S. C. (2006), “Fast Bayesian shape matching using geometric algorithms (with discussion),” in *Bayesian Statistics 8*, eds. J. M. Bernardo, S. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 471–490, Oxford University Press, Oxford.
- Schneider, T. R. (2000), “Objective comparison of protein structures: error-scaled difference distance matrices,” *Acta Crystallographica Section D*, 56, 714–721.
- Smith, T. and Waterman, M. (1981), “Identification of common molecular subsequences,” *Journal of Molecular Biology*, 147, 195 – 197.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), “Bayesian measures of model complexity and fit,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.
- Storz, J. F., Opazo, J. C., and Hoffmann, F. G. (2013), “Gene duplication, genome duplication, and the functional diversification of vertebrate globins,” *Mol. Phylogenet. Evol.*, 66, 469–478.
- Suchard, M. A. and Redelings, B. D. (2006), “Bali-Phy: simultaneous Bayesian inference of alignment and phylogeny,” *Bioinformatics*, 22, 2047–2048.
- Teh, A.-H., Saito, J. A., Baharuddin, A., Tuckerman, J. R., Newhouse, J. S., Kanbe, M., Newhouse, E. I., Rahim, R. A., Favier, F., Didierjean, C., Sousa, E. H., Stott, M. B., Dunfield, P. F., Gonzalez, G., Gilles-Gonzalez, M.-A., Najimudin, N., and Alam, M. (2011), “Hells Gate globin I: An acid and thermostable bacterial hemoglobin resembling mammalian neuroglobin,” *{FEBS} Letters*, 585, 3250 – 3258.
- Terwilliger, N. (1992), “Molecular Structure of the Extracellular Heme Proteins,” in *Blood and Tissue Oxygen Carriers*, ed. C. Mangum, vol. 13 of *Advances in Comparative and Environmental Physiology*, pp. 193–229, Springer Berlin Heidelberg.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994), “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” *Nucleic Acids Research*, 22, 4673–4680.

- Thompson, J. D., Plewniak, F., and Poch, O. (1999), “BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs.” *Bioinformatics*, 15, 87–88.
- Thorne, J., Kishino, H., and Felsenstein, J. (1991), “An evolutionary model for maximum likelihood alignment of DNA sequences,” *J. Mol. Evol.*, 33, 114–124.
- Thorne, J., Kishino, H., and Felsenstein, J. (1992), “Inching toward reality: an improved likelihood model of sequence evolution,” *J. Mol. Evol.*, 34, 3–16.
- Thorne, J. L., Kishino, H., and Painter, I. S. (1998), “Estimating the rate of evolution of the rate of molecular evolution.” *Molecular Biology and Evolution*, 15, 1647–1657.
- Tiana, G., Shakhnovich, B. E., Dokholyan, N. V., and Shakhnovich, E. I. (2004), “Imprint of evolution on protein structures,” *Proceedings of the National Academy of Sciences of the United States of America*, 101, 2846–2851.
- Uhlenbeck, G. E. and Ornstein, L. S. (1930), “On the Theory of the Brownian Motion,” *Phys. Rev.*, 36, 823–841.
- Vázquez-Limón, C., Hoogewijs, D., Vinogradov, S. N., and Arredondo-Peter, R. (2012), “The evolution of land plant hemoglobins,” *Plant Science*, 191192, 71 – 81.
- Vinogradov, S. N., Hoogewijs, D., Bailly, X., Arredondo-Peter, R., Guertin, M., Gough, J., Dewilde, S., Moens, L., and Vanfleteren, J. R. (2005), “Three globin lineages belonging to two structural classes in genomes from the three kingdoms of life,” *Proceedings of the National Academy of Sciences of the United States of America*, 102, 11385–11389.
- Walker, A. M. (1969), “On the Asymptotic Behaviour of Posterior Distributions,” *Journal of the Royal Statistical Society*, 31, 80–88.
- Wang, R. and Schmidler, S. C. (2013), “Bayesian Multiple Protein Structure Alignment and Analysis of Protein Families,” (*submitted*).
- Westesson, O., Lunter, G., Paten, B., and Holmes, I. (2012), “Accurate Reconstruction of Insertion-Deletion Histories by Statistical Phylogenetics,” *PLoS ONE*, 7, e34572.
- Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008), “Alignment uncertainty and genomic analysis,” *Science*, 319, 473–76.
- Wood, A. (1994), “Simulation of the von Mises Fisher distribution,” *Comm. Statist. Sim. Comp.*, 23, 157–164.

- Wood, T. C. and Pearson, W. R. (1999), “Evolution of protein sequences and structures,” *Journal of Molecular Biology*, 291, 977 – 995.
- Yang, Z. (1994), “Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods,” *J. Mol. Evol.*, 39, 306–314.
- Yang, Z. and Rannala, B. (2005), “Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny,” *Systematic Biology*, 54, 455–470.
- Ycart, B. (1999), “Cutoff for samples of Markov chains,” *ESAIM: Probability and Statistics*, 3, 89–106.
- Zhang, Z., Wang, Y., Wang, L., and Gao, P. (2010), “The combined effects of amino acid substitutions and indels on the evolution of structure within protein families,” *PLoS One*, 5, e14316.

Biography

Christopher John Challis was born in Olathe, KS on April 30, 1982. He graduated with a Bachelor of Science in Mathematics and Statistics from Brigham Young University in April, 2006. The current work is fulfillment of requirements of a Master of Science and Doctor of Philosophy in Statistics at Duke University.