Estimating the history of mutations on a phylogeny

Jonathan P. Bollback, Paul P. Gardner, Rasmus Nielsen

Center for Bioinformatics and Institute of Biology
University of Copenhagen, Universitetsparken 15
2100 Copenhagen Ø, Denmark
bollback@binf.ku.dk, pgardner@binf.ku.dk, rasmus@binf.ku.dk

1 Introduction

Evolution is a historical process that has left its signature in the molecules and morphology of living organisms. Attempts to better understand the specific and general features of evolution involve making inferences about the past from these tell tale signs (Brooks and McLennan, 1991; Felsenstein, 1985; Harvey and Pagel, 1991; Pagel, 1999). Ancestral state reconstruction is a powerful tool in this endeavor as exemplified by its application to a wide array of questions (e.g., Bishop et al., 2000; Gillespie, 1991; Langley and Fitch, 1974; Messier and Stewart, 1997; Templeton, 1996). Traditionally, ancestral reconstruction has relied on well understood approaches such as parsimony. However, the last decade has seen an excited flurry of research into statistical approaches as exemplified by the contents of this volume and the primary literature.

In a general sense, most methods for ancestral reconstruction have focused on reconstructing the ancestral states at the internal nodes of a phylogeny. Often, we are not interested in particular nodes of the phylogeny but the whole history of a character. In this chapter we focus on a Bayesian method for estimating these histories on phylogenies (we refer to a complete description of character's history as its 'mutational path'). Mutational paths differ most notably from other approaches in its ability to estimate, not only, the ancestral states at the internal nodes of a phylogeny but also the order and timing of mutational changes on the phylogeny. Our goal here is to provide a concise introduction to the statistical tools necessary for estimating mutational histories, making inferences from these histories, and to provide some examples of the power of this recent approach.

2 Likelihood and Bayesian methods

Estimation of ancestral characters states using maximum likelihood (ML) proceeds using a straightforward extension of the usual algorithm for calculation of the likelihood function in phylogenetics. Let $f_{ij}(k)$ be the fractional of nucleotide k in site i, node j, of the phylogeny. That is, $f_{ij}(k)$ is the probability of all the data in site i below node j given that the ancestral state in node j is k. The ML estimate of ancestral state in node j is

then obtained by placing the root in node k and maximizing

$$L(k) = f_{ij}(k) \tag{1}$$

with respect to k. Other parameters of the evolutionary model (branch lengths, parameters of the mutational model, etc) have typically been estimated prior to the analysis and are assumed to be fixed. The method can also be extended to find the joint set of ancestral states for all nodes that maximizes the likelihood (Koshi and Goldstein, 1996; Pupko et al., 2000; Yang et al., 1995). For more details on maximum likelihood estimation, please see other relevant chapters of this volume. One important thing to notice is that under a uniform prior for all possible ancestral states, the ML estimate is also a Bayesian Maximum *A posteriori* Probability (MAP) estimate. However, from a Bayesian perspective it arguably makes little sense to first estimate all parameters of the model (except ancestral states) using ML, and then to estimate ancestral states. Instead, it is preferable to estimate ancestral sequences jointly for all sites at the same time while integrating over all other parameters. The advantage is that the phylogenetic uncertainty, and uncertainty regarding the parameters of the evolutionary model are taken into account in the estimation of ancestral states. This is achievable using Markov Chain Monte Carlo (MCMC) methods described in the following. One of the advantages of this method is that it directly provides an estimate of an entire evolutionary history, i.e. of ancestral states, not only at the nodes of the phylogeny, but also at any point in time along the branches (edges) of the phylogeny.

3 MCMC

The basic idea in the MCMC algorithm is to represent mutations directly on the phylogeny. A history of mutations along one or more branches of the phylogeny is called a 'mutational path'. The concept of a mutational path is illustrated in Figure 1. Bayesian ancestral reconstruction using MCMC exist in two flavors: (1) a 'two-step' approach where a sample of genealogies and parameter values is first sampled using MCMC, and mutational paths are subsequently simulated given particular sampled phylogenies (Bollback, 2006; Nielsen, 2002), and (2) 'direct methods' where the mutations are represented on the phylogeny while simulating the genealogy (Nielsen, 2001). Both approaches achieve the same goal, but they differ in computational efficiency and in which models they can accommodate. In general, we will assume the model of evolution can be described by a Markov chain on a finite, discrete state space with known generator, such as the set of all possible amino acids, all possible codons, or all possible nucleotides. We will also initially assume that this Markov process is independent among sites, although as we will show, one of the advantages of these methods is that they can easily be extended to models of correlated evolution among sites.

3.1 Sampling mutational paths

In the following we will describe how algorithms of type (1) proceed. Consider first the case of a fixed phylogeny with known branch lengths and evolutionary model. We will first

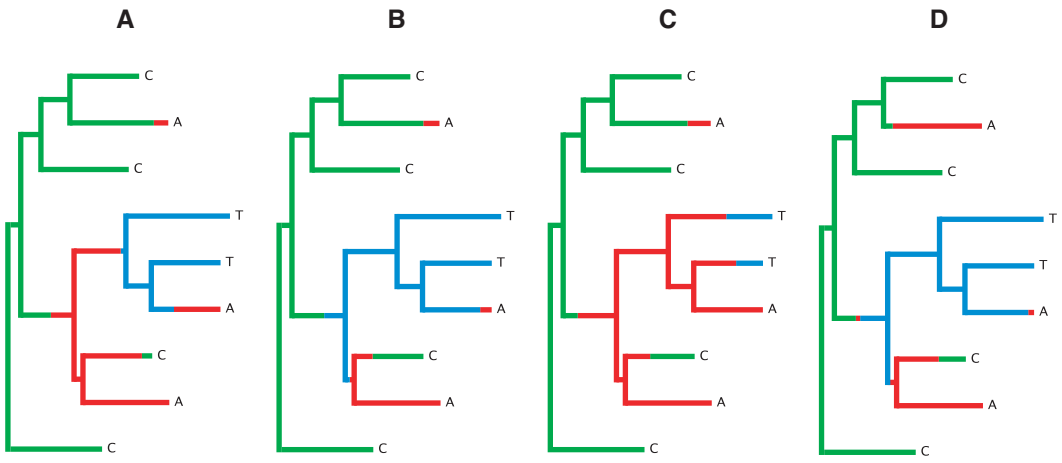

Figure 1: Examples of four possible mutational paths for a single nucleotide site. Histories and images derived from the SIMMAP software package (Bollback, 2006),

describe how to sample a mutational path stochastically under a particular evolutionary model. From Equation (1) we see that

$$Pr(y_{ij} = k|x_i) = \frac{f_{ij}(k)\pi_k}{\sum_b f_{ij}(b)\pi_b}, \tag{2}$$

where π_b is the prior probability of state b (usually assumed to be the stationary frequency of b under the model), x_i is the observed data in site i and y_{ij} is the unknown ancestral state in site i, node j. The ancestral state of in the root of the tree can then be sampled according to these probabilities. At a child node of node j, say node h, the ancestral state, y_{ih}, can then be sampled according to the probabilities

$$Pr(y_{ih} = k|x_i, y_{ij} = v) = \frac{f_{ih}(k)p_{vk}(h)}{\sum_b f_{ih}(b)p_{vb}(h)}, \tag{3}$$

This sampling procedure can then be repeated recursively along the branches in the tree until ancestral sequences have been sampled for all nodes. In a sample of n sequences the resulting vector of nucleotides, $y_i = (y_{i1}, y_{i2}, ..., y_{i2n-3})$ represents a sample from $Pr(y_i|x_i)$. For a time-reversible model of substitution, the distribution of y does not depend on where in the tree the root has been placed. An entire mutational path can then be obtained by sampling paths conditional on the ancestral sequences at the nodes. If we let z_{ih} be the mutational path leading to node h from node j in site i, with sampled ancestral states $y_{ih} = k$ and $y_{ij} = v$, a sample from the density $p(z_{ih}|y_{ih} = k, y_{ij} = v)$ can be obtained using standard methods for simulating Markov chains starting at state v. The conditioning can be achieved by simply eliminating paths which do not end in state k, and can be speeded up in various ways. Repeating this scheme for all branches of the tree provides a full sample of $z_i|y_i$. The simulation procedure is completed by applying this procedure to all sites providing a full sample of $z = (z_1, z_2, ...)$ and $y = (y_1, y_2, ...)$.

3.2 Incorporating phylogenetic uncertainty

The preceding description of the simulation procedure assumed that the phylogenetic tree and the parameters where known, i.e., it produces samples from the density $p(z, y|x, \theta)$, where θ is a vector of all the nuisance parameters, including the mutational model and the phylogenetic tree with branch lengths. However, usually θ will not be known. In such cases, we wish to be able to obtain samples from

$$p(z, y|x) = \int p(z, y|x, \theta)p(\theta|x)d\theta, \tag{4}$$

where the integral is over all supported values of θ. We can think of this integral as a sum over all topologies of the tree and a multiple integral over all possible branch lengths and parameters of the mutational process. The representation in Equation (3) suggest the following method for obtaining samples from $p(z, y|x)$:

1. Sample $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(n)}$ from $p(\theta|x)$.
2. Sample $z^{(s)}, y^{(s)}$ from $p(z, y|x, \theta^{(s)})$ for $s = 1, 2, ..., n$.

The samples from $p(\theta|x)$ can be obtained using any of the well known MCMC procedures, for example using MrBayes (Huelsenbeck and Ronquist, 2001). This method provides a method for obtaining samples from $p(z, y|x)$ which takes advantage of existing computational methods for Bayesian inference in phylogenetics and is comparatively easy to implement.

3.3 Direct methods

In the direct methods (Nielsen, 2001; Robinson et al., 2003), the MCMC procedure is applied directly to a phylogeny with mutational paths, i.e. the state space of the Markov chain is the set of supported values of (y, z, θ). A Markov chain with stationary density $p(y, z, \theta|x)$ can be simulated, for example, by iterating updates of (x, y):
For $i = 1, 2, ..., B$, where B is the number of sites:

1. Simulate a new value of (z_i, y_i), (z_i^*, y_i^*), from a proposal density $q(y_i^*, z_i^*|x_i, \theta)$
2. Accept z_i^*, y_i^* with probability $\frac{p(y^*,z^*,x|\theta)q(y_i,z_i|x_i,\theta)}{q(y_i^*,z_i^*|x_i,\theta)p(y,z,x|\theta)}$

and updates of θ:

1. Simulate a new value of θ, θ^*, from a proposal density $q(\theta^*|x, y, z)$.
2. Accept θ^* with probability $\frac{p(y,z,x|\theta^*)p(\theta^*)q(\theta|x,y,z)}{q(\theta^*|x,y,z)p(y,z,x|\theta)p(\theta)}$

In the above notation, the current state before an update is simply denoted by (x, y, θ) to simplify the notation and (z^*, y^*) is (z, y) with (z_i^*, y_i^*) replacing (z_i, y_i). Notice that any update kernel $q(...)$ can be used as long as it ensure that all values of (θ, z, y) supported by $p(\theta, z, y|x)$ eventually can be reached. This simulation procedure takes advantage of the fact that $p(c, x|\theta)$ easily can be calculated directly from the generator as the sampling

path of a continuous time Markov chain without the need to calculate time-dependent transition probabilities. However, this MCMC procedure can be slow to converge because of the correlation between (y, z) and θ. One of the advantages of this procedure is that it allows the use of models with correlated evolution among sites (e.g., Robinson et al., 2003; Yu and Thorne, 2006). In models of protein evolution involving tertiary structure, or models involving CpG hyper-mutations, the state space of the Markov model is the set of 4^B possible sequences of length B. However, the likelihood function can no longer be written as the product of the likelihood in multiple independent sites. This means than the conventional statistical methods for inference are inapplicable. Numerical calculation of the time-dependent transition probabilities of the process are hard or impossible to calculate and methods based on summing over all possible states (e.g. as part of the likelihood calculation) are intractable. However, while it is not possible to calculate $p(x|\theta)$ in these models, calculations of $p(x, y, z|\theta)$ are straightforward. This means that MCMC procedures with state space on (y, z, θ) can be implemented relatively easily.

3.4 Statistical inference using sampled mutational paths

After obtaining samples from the posterior distribution of (y, z, θ) using either of the two methods, inference in a Bayesian framework proceeds in a straightforward fashion. The posterior expectation (or summary statistic) of any function of the mutational path can easily be calculated. For example, Nielsen (2001) used this method to calculate the posterior expectation of the ages of non-synonymous and synonymous mutations occurring on a phylogeny. The framework provides a computationally tractable framework for making rigorous statistical inferences based on mutational paths. Bollback (2002) and Nielsen and Huelsenbeck (2001) showed how statistical measures of uncertainty can be obtained in this framework based on the posterior predictive distribution.

Briefly, the posterior predictive distribution measures uncertainty by estimating the 'null' or predictive distribution of mutational paths, given the information obtained from the posterior distribution. In this way, a summary statistic can be compared with its predictive distribution to test hypotheses and measures of statistical uncertainty. This approach is appealing because it can be applied and tailored to a wide variety of questions involving only the ability to specify a summary statistic and simulate posterior predictive mutational histories. A typical summary statistic would be calculated by summing over simulated mutational mappings, i.e. $h(x) = \sum_i h(x, y^{(i)}, z^{(i)})$. Using the indirect approach for obtaining samples from the posterior distribution, the estimation of the posterior predictive distribution can be accomplished in the following way:

1. Generate n samples of θ, $(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(n)})$ from the posterior distribution using established MCMC methods.

2. For $j = 1, 2, ..., n$

   Simulate a new set of aligned sequence, $x^{(j)}$, under $\theta^{(j)}$.

   Simulate $k \leq n$ mutational paths, $(z^{(j,i)}, y^{(j,i)})$, $i = 1, 2, ..., k$, based on $x^{(j)}$ and $\theta^{(i)}$.

3. The distribution of $h(x^{(j)}) = \sum_i h(x, y^{(j,i)}, z^{(j,i)})$ then approximates the posterior predictive distribution of $h(x)$.

It should be noticed that for each of the simulated predictive data sets the posterior predictive expectation of a site's mutational history is averaged over all of the samples from the posterior. In this way we can effectively integrate out uncertainty in the posterior distribution using a single MCMC run. For additional details on predictive distributions the reader is referred to the literature (e.g., Bollback, 2002; Nielsen and Huelsenbeck, 2001; Suchard et al., 2003)

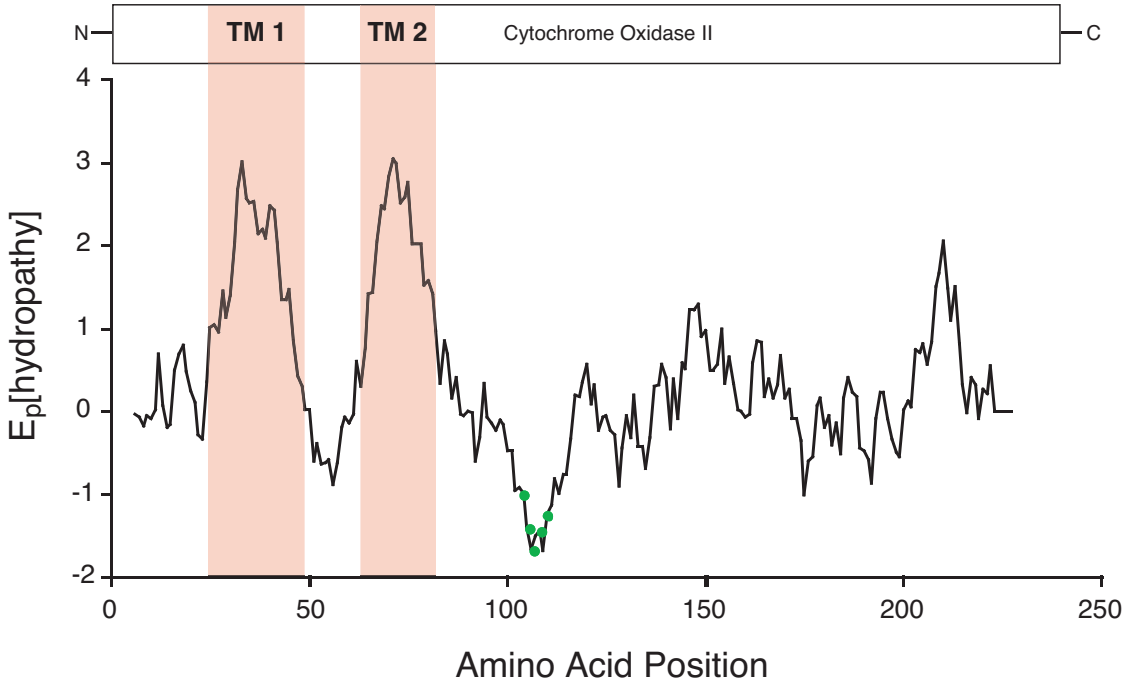

Figure 2: A comparitive Kyte-Doolittle hydropathy plot of the hydopathic character of the primate cytochrome oxidase II protein using mutational mapping to summarize multiple sequences. Transmembrane regions are shown in red and the green dots show the predicted amino acids involved in electron transport. The location of the transmembrane domains are shown at the top of the graph.

4 Examples

The method of mutational mapping has already found wide use in addressing questions about the age of mutations (Nielsen, 2001), positive selection (Nielsen and Huelsenbeck, 2001), morphological evolution (Huelsenbeck et al., 2003), and modeling non-independence among residues (Dimmic et al., 2005; Yu and Thorne, 2006), among others.

In this section we will illustrate two different uses of mutational mapping. Each of the analyses were performed using the SIMMAP software package. (A brief introduction to this package is provided in the next section.) We hope that this section will demonstrate

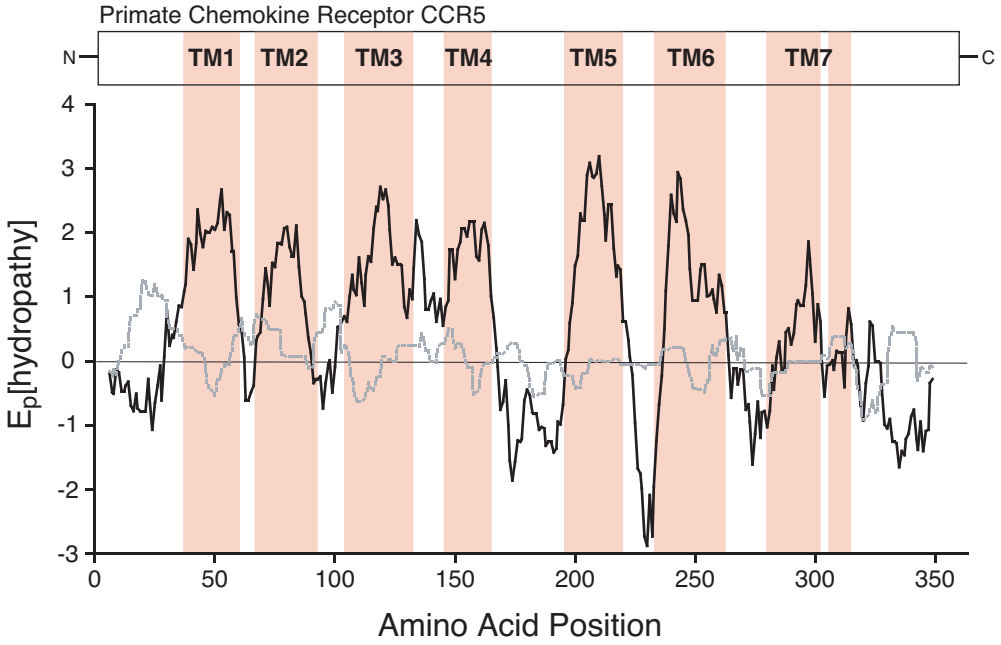

Figure 3: A comparative Kyte-Doolittle hydropathy plot of the hydropathic features of the primate CCR5 protein using mutational mapping to summarize multiple sequences. The posterior expectation of hydropathy along the gene is shown by the black line while the dashed grey line is the posterior expectation of the direction and magnitude of change in hydropathy (see Figure 4 for a plot of the hydropathic magnitude and direction of changes on a residue by residue basis). The plots were generated using a window size of 11 residues. Transmembrane regions are shown in red and the location of the transmembrane domains are shown at the top of the graph.

the types of questions that can be addressed using mutational mapping and motivate its use and further development.

4.1 Characterizing transmembrane regions of proteins

Transmembrane proteins span cellular membranes with intra- and extra-membrane regions. One feature of these proteins is the highly hydrophobic (water-fearing) nature of the helices spanning the membrane and the hydrophilic (water-loving) nature of the protein domains in the cytoplasmic and periplasmic spaces. An early single sequence approach was evaluating the hydropathy of the primary structure of a protein (Kyte and Doolittle, 1982). This approach plots hydropathy, using a sliding window spanning n residues, as a function of sequence position (residue) and have been called Kyte-Doolittle hydropathy plots.

The Kyte-Doolittle method considers only a single sequence. To demonstrate the use of mutational maps we present an approach that produces a comparative Kyte-Doolittle plot that accommodates multiple sequences, uncertainty in the phylogeny relating the sequences, and the evolutionary model describing changes in those sequences. While

more advanced methods exist for discovering transmembrane domains (e.g., Krogh et al., 2001) the simple approach described here offers the ability to summarize data across multiple sequences to initially identify likely transmembrane regions. In addition, other methods may benefit from the inclusion of this type of multi-sequence approach.

Two data sets are analyzed, using the indirect mutational mapping approach, are a primate cytochrome oxidase II data set (Yoder and Yang, 2004) consisting of 52 species of lemur, and the second is a primate chemokine receptor CCR5 data set (a subset of sequences analyzed in Mummidi et al., 2000) consisting of the mouse sequence and 11 primate sequences.

Cytochrome oxidase II (COII) is one of a number of proteins involved in the electron transport chain in the mitochondria. COII has two transmembrane domains (TM1 and TM2; see Figure 2) and a highly aromatic (hydrophilic) region of amino acid residues involved directly in electron transport (Adkins and Honeycutt, 1994). Using the mapping approach the hydropathy of each site, averaged across the probable phylogenies, is evaluated for the primate COII. Sampling estimates of the phylogeny and model were first obtained using MrBayes (Huelsenbeck and Ronquist, 2001). SIMMAP was used to map mutational paths for each codon and from these the history of the hydropathy was summarized. This approach weights a site's hydropathy by taking the branch length weighted sum of the site's amino acid mutational history. The tree weighted posterior expectation of hydropathy for each residue was evaluated across 105 trees and was plotted in a Kyte-Doolittle plot with a sliding window of size 11 (see Figure 2). The method clearly identifies the two transmembrane regions of COII protein and identifies the conserved hydrophilic region with amino acid residues involved directly in electron transport.

The second data set consisted of 11 primate sequences and the mouse sequence (Mummidi et al., 2000) which were retrieved and aligned from GenBank; mouse, human, chimpanzee, gorilla, green monkey, spider monkey, squirrel monkey, golden lion tamarin, golden-rumped tamarin, a marmoset, and two species of lemur. The chemokine receptor (CCR5) protein is a coreceptor target of the HIV and SIV, and possibly other related viruses (Mummidi et al., 2000; Paterlini, 2002). CCR5 has 7 transmembrane domains labeled TM1-7 (Paterlini, 2002). Using mutational histories we analyzed CCR5 for the hydropathic signature of the transmembrane regions (Figure 3) and the association between hydropathic change and positive selection (Figure 4). In addition, for each site we calculated the posterior expectation of the change in hydropathy.

Of the 7 transmembrane domains in CCR5 TM1-6 are clearly identified has being highly hydrophobic while TM7 does not show a large deviation in the hydropathy. The cytoplasmic/periplasmic regions show a considerable bias towards being hydrophilic. Inspection of the mean change in hydropathy indicates that there is a general pattern of transmembrane regions showing values close to zero or slightly negative while hydrophilic regions exhibit larger changes with a tendency towards hydrophobicity (Figure 4). This latter observation at first seems surprising but may reflect tertiary packing in these regions in which evolution is occurring at mostly buried residues or at residues that are under selection for interaction with other proteins; sites that show evidence for positive selection (Figure 4) occur, predominantly, in cytoplasmic/periplasmic domains (74% of sites). In addition, sites in transmembrane regions that show evidence for positive selection show a large tendency for change to more hydrophobic residues. Never the less

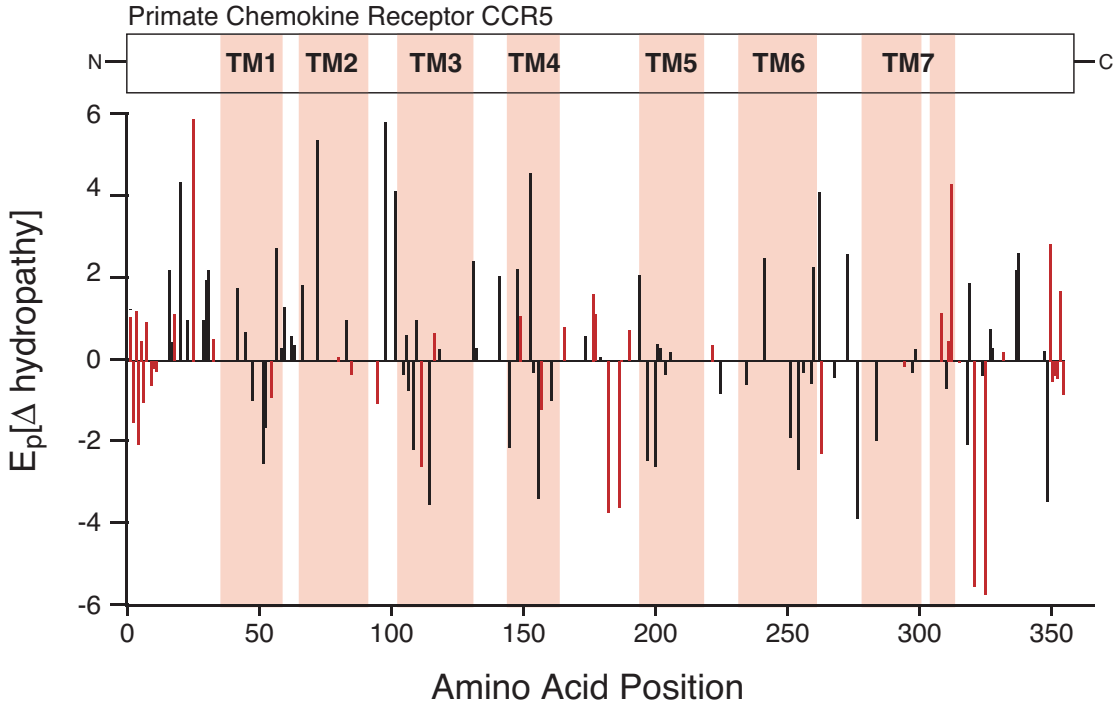

Figure 4: Patterns of change in hydropathy and positive selection. The magnitude and direction of change in hydropathy are plotted along the sequence. Sites that show at least two substitutions and with a non-synonymous/synonymous ratio greater than one are shown in red.

the approach provides comparative information indicating domains that are likely to be highly hydrophobic and spanning the cellular membrane.

4.2 Nucleotide covariation in mitochondrial tRNAs

One powerful use of the mutational mapping approach is in detecting non-independence, or covariation, among nucleotide residues. Mutational histories have been successfully applied to detecting compensatory changes in amino acid residues (Dimmic et al., 2005) and correlation between morphological characters (Huelsenbeck et al., 2003).

To illustrate the use of mutational histories for detecting nucleotide covariation we analyzed all 21 mitochondrial tRNAs for 106 species with the hope of detecting positive evidence for covariation among base pairing residues. As above, we adopted the indirect approach using MrBayes (Huelsenbeck and Ronquist, 2001) to provide a sampling approximation of the phylogeny and substitution model parameters. All 21 tRNA genes were concatenated and the GTR + Gamma substitution model was used (Lanavé et al., 1984; Yang, 1994) to obtain samples from the posterior distribution of the phylogeny and substitution model parameters.

But how can we measure covariation among nucleotides using mutational histories? The answer is straightforward. For each site sample a mutational path and then com-

pare the coviariation (coincidence of states) between each site's history. Any number of statistics can be used. To evaluate the degree of covariation we calculated the association between different states along each branch of a phylogeny using the following statistic,

$$m_{ij} = f_{ij} \log_2 \frac{f_{ij}}{f_i f_j}, \tag{5}$$

where f_{ij} is the fraction of time state i is associated with j in a character history, and f_i is the fraction time in a particular state independent of associations (i.e., the sum of time spent in a particular state on the phylogeny). We refer to this statistic as the *mutual historical information content*, MHIC, because of its relationship in form to the classical mutual information content statistic (Chiu and Kolodziejczak, 1991; Gutell et al., 1992). By averaging over all state associations we get the following statistic for the overall character correlation,

$$M = \sum_{i=1}^{x} \sum_{j=1}^{y} m_{ij}. \tag{6}$$

In addressing whether we can detect covariation in RNA molecules, that is the result of base pairing constraints, we perform the summation over only Watson-Crick (A·U and G·C) and wobble pairs (G·U). In this way we focus on covariation that is due to base pairing. To accommodate uncertainty in the phylogeny and substitution model parameters we calculated the posterior expectation of the MHIC statistic for each pair and then compare the know pairs with the unknown pairs for each tRNA. A comparison of the MHIC reveals that paired sites show a strong signature of covariation relative to unpaired sites (Figure 5) in 20 of the 21 tRNAs; in the single case of the tRNA_{Met} the difference between paired and unpaired is indistinguishable. This is likely the result of very little variation in the tRNA_{Met} at paired sites. These results clearly indicate that the signature of covariation can be easily identified using the mutational history approach.

While the previous results indicate a strong difference in the signature of covariation between know pairs and unpaired sites we might wish to determine how well the method does in predicting true pairs (true positives) relative to miss-assigning pairs (false positives). One commonly used measure is Mathew's correlation coefficient,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

where TN is the number of true negatives, TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives. In using this measure the first step is to rank each comparison MHIC value and then to break this into discrete intervals or thresholds. Values above the threshold are considered to be *paired* and values below to be *unpaired*. At each threshold the MCC is calculated and is a measure of the method's ability to correctly identify true pairs while minimizing false positives. Figure 6 shows the MCC values for the the tRNA_{Gly} and tRNA_{His}.

We evaluated overall performance by calculating the mean and 95% confidence intervals of MCC at each threshold value for all 21 tRNAs (Figure 7). The performance of the method, as measured by the MCC, is high with a maximum mean MCC value of 0.42,

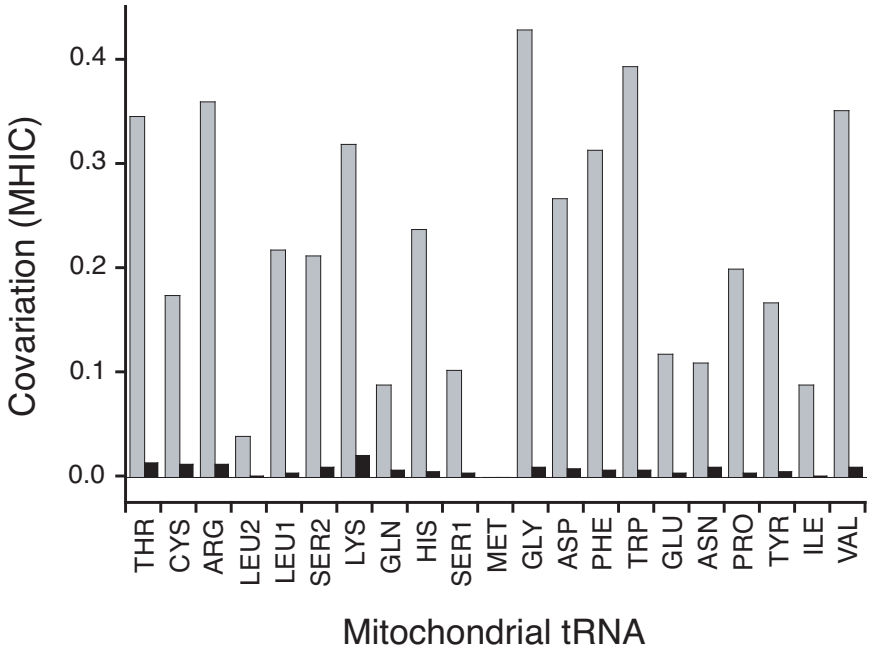

Figure 5: Nucleotide covariation measures for the 21 mammalian tRNAs. Grey boxes represent the mean MHIC for known pairs and black boxes are the mean of unpaired sites. A comparison of the standard deviations (not shown) indicate that in all but the tRNA_{Met} these differences are significant.

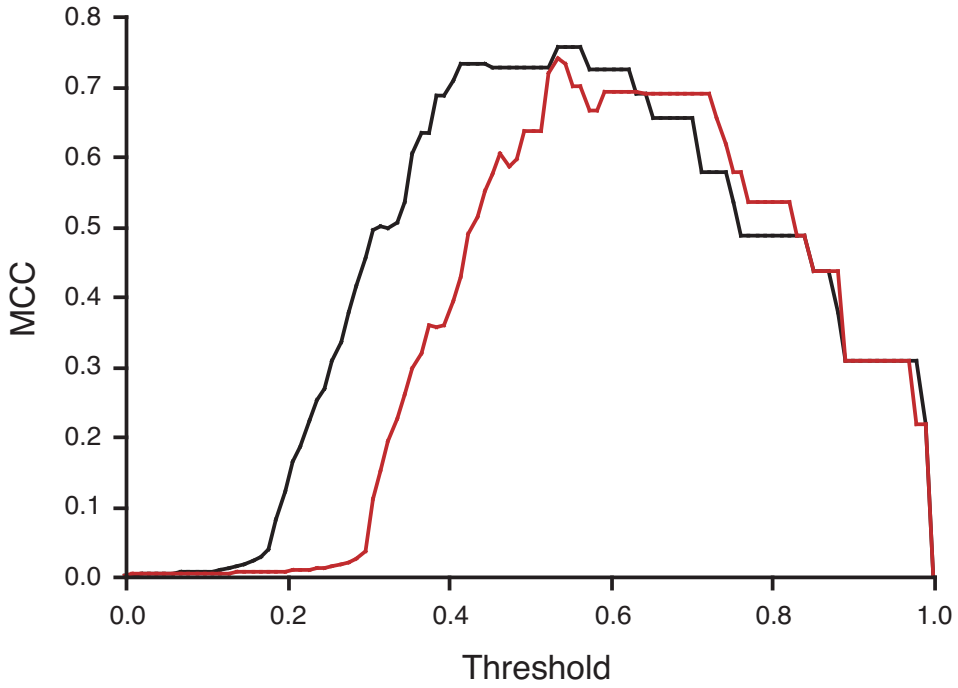

Figure 6: Mathew's correlation coefficient (MCC) for the mammalian tRNA_{Gly} (black) and tRNA_{His} (red).

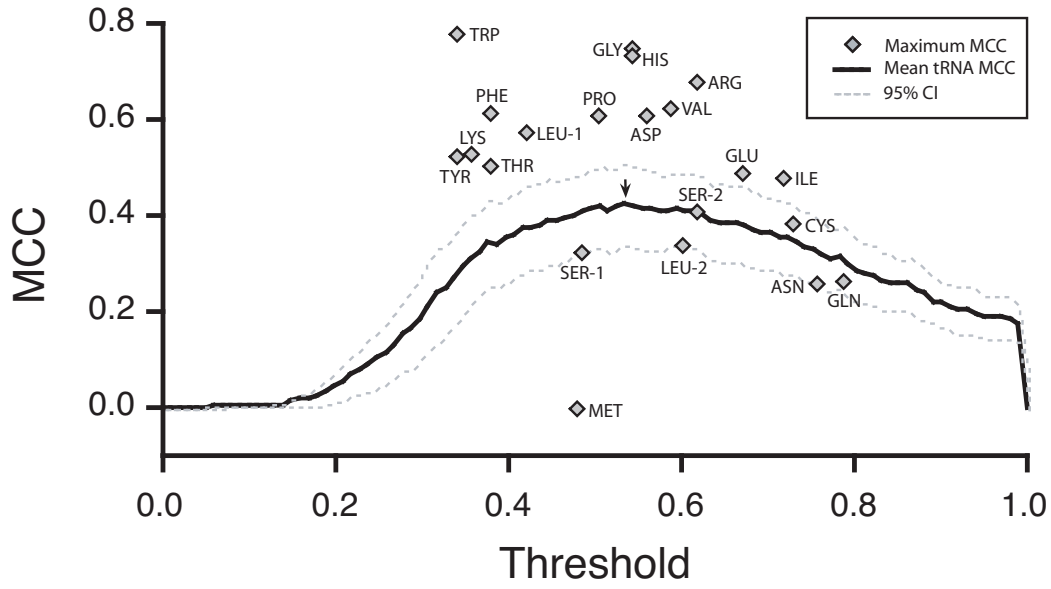

Figure 7: Mathew's correlation coefficient (MCC) for the 21 mammalian tRNAs.

averaged across the 21 tRNAs, and with 11 out of 21 (52%) of the tRNAs with individual maximum MCC values above 50%. The performance of this approach does decline with divergence as expected (data not shown).

5 Software

Few user friendly software packages have been written that generalize the mutational mapping method to different kinds of data (molecular and morphological) and a wide variety of biological questions. We introduce one such package, SIMMAP, which provides researchers the ability to address a wide variety of questions using either molecular or morphological data.

Briefly, SIMMAP is a software implementation of the indirect method of mapping mutational histories described in the first sections of this chapter. For example, SIMMAP can be used to sample character histories on phylogenies to address questions about general evolutionary patterns of change (trends), positive selection at focal sites or across a gene region, correlation among characters (as described above), and will generate the raw mutational histories for a custom analyses not available in the software package. In addition, SIMMAP can be used to calculate the posterior distribution of ancestral states in either a full hierarchical or empirical Bayesian framework. The software accommodates a wide variety of substitution models and priors. SIMMAP is free for academic use and a download can be obtained at `http://www.simmap.com`.

Acknowledgments

The author's would like to acknowledge the following funding agencies for support during the writing; a Danish FNU Grant (271-05-0599) to J.P.B and R.N, and a Carlsberg Foundation Grant (21-00-0680) to P.P.G.

References

Adkins, R. M. and Honeycutt, R. L. (1994). Evolution of the primate cytochrome c oxidase subunit II gene. *J Mol Evol*, 38(3):215–231.

Bishop, J. G., Dean, A. M., and Mitchell-Olds, T. (2000). Rapid evolution of plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proceedings of the National Academy of Science USA*, 97:5322–5327.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol*, 19:1171–1180.

Bollback, J. P. (2006). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7(88).

Brooks, D. R. and McLennan, D. A. (1991). *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology*. University of Chicago Press.

Chiu, D. K. and Kolodziejczak, T. (1991). Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, 7:347–352.

Dimmic, M. W., Hubisz, M. J., Bustamante, C. D., and Nielsen, R. (2005). Detecting coevolving amino acid sites using bayesian mutational mapping. *Bioinformatics*, 21 Suppl 1:126–126.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, 125:1–15.

Gillespie, J. (1991). *The Causes of Molecular Evolution*. Oxford University Press.

Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J., and Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic. Acids. Res.*, 20:5785–5795.

Harvey, P. H. and Pagel, M. D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press.

Huelsenbeck, J. P., Nielsen, R., and Bollback, J. P. (2003). Stochastic mapping of morphological characters. *Syst Biol*, 52:131–158.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics Applications Note*, 17:754–755.

Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol*, 42(2):313–320.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

Lanavé, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20:86–93.

Langley, C. H. and Fitch, W. M. (1974). An estimation of the constancy of the rate of molecular evolution. *J Mol Evol*, 3:161–177.

Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysomzymes. *Nature*, 385:151–154.

Mummidi, S., Bamshad, M., Ahuja, S. S., Gonzalez, E., Feuillet, P. M., Begum, K., Galvis, M. C., Kostecki, V., Valente, A. J., Murthy, K. K., Haro, L., Dolan, M. J., Allan, J. S., and Ahuja, S. K. (2000). Evolution of human and non-human primate cc chemokine receptor 5 gene and mrna. potential roles for haplotype and mrna diversity, differential haplotype-specific transcriptional activity, and altered transcription factor binding to polymorphic nucleotides in the pathogenesis of HIV-1 and simian immunodeficiency virus. *J Biol Chem*, 275(25):18946–18961.

Nielsen, R. (2001). Mutations as missing data: inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics*, 159:401–411.

Nielsen, R. (2002). Mapping mutations on phylogenies. *Syst Biol*, 51:729–732.

Nielsen, R. and Huelsenbeck, J. P. (2001). Detecting positively selected amino acid sites using posterior predictive p-values. In *Pacific Symposium on Biocomputing, Proceedings*, pages 576–588. World Scientific, Singapore.

Pagel, M. D. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401:877–884.

Paterlini, M. G. (2002). Structure modeling of the chemokine receptor ccr5: implications for ligand binding and selectivity. *Biophys J*, 83(6):3012–3031.

Pupko, T., Pe'er, I., Shamir, R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*, 17(6):890–896.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*, 20(10):1692–1704.

Suchard, M. A., Weiss, R. E., Sinsheimer, J. S., Dorman, K. S., Patel, P., and McCabe, E. R. B. (2003). Evolutionary similarity among genes. *Journal of the American Statistical Association: Theory and Methods*, 98(463):653–662.

Templeton, A. R. (1996). Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the cytochrome oxidase II gene in the hominoid primates. *Genetics*, 144:1263–1270.

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., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650.

Yoder, A. D. and Yang, Z. (2004). Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. *Mol Ecol*, 13(4):757–773.

Yu, J. and Thorne, J. L. (2006). Dependence among sites in RNA evolution. *In preparation.*