9

Markov Models of Protein Sequence Evolution

Matthew W. Dimmic

Department of Biological Statistics and Computational Biology,
Cornell University, Ithaca, NY 14853, USA, mwd8@cornell.edu

9.1 Introduction

Proteins play a vital role in almost every process of life. There are over 2100 known protein families; many are crucially involved in biochemical metabolism, cellular signaling and transport, reaction catalysis, cytoskeletal structure, immune recognition, and sensory input. Because single mutations in the amino acid sequences can have a drastic effect on the protein's function–and potentially on the fitness of the individual–proteins can be considered the primary unit of phenotypic expression. The complex relationship among protein chemistry, structure, function, and evolution is therefore a significant piece of the evolutionary puzzle, and models of protein evolution are used to test hypotheses about these relationships.

There are several applications where protein evolutionary models have had particular success. For example:

1. detecting and aligning remote homologs,
2. measuring divergence times between sequences and species,
3. inferring the evolutionary history of related proteins (the phylogenetic tree), and
4. determining the physicochemical factors that have been important to the function and evolution of a protein family.

This chapter focuses on models that can be used for the last two applications, specifically those that treat evolution as a Markov chain with transitions between amino acid states. When combined with the statistical toolbox of likelihood methods (Chapter 2), Markov models have proven to be a powerful tool for phylogenetic inference and hypothesis testing. Rather than attempting to provide an exhaustive description of all available models, this chapter will highlight a few that illustrate the important distinguishing features of protein sequence evolution.

9.2 Basic Features of Protein Sequences

The factors that are important to the evolution of a protein can be complex and subtle, and the study of protein evolution is a very active field. A detailed discussion of protein structure and evolution is beyond the scope of this chapter (see, e.g., references [10, 60]), but a few of the most relevant features from an evolutionary perspective follow.

- **Most proteins function natively with the amino acid chain folded into a stable three-dimensional structure.** This structure is called the *tertiary structure* and is thought to be determined primarily by the amino acid chain's interaction with a solvent and with itself (a proposition known as the Thermodynamic Hypothesis [4]). While the protein-folding pathway is still not well-understood for most sequences, some general principles of protein folding are known. For example, in aqueous solution, a combination of entropic and enthalpic factors combine to cause hydrophobic (oily) amino acids to fold into the interior of a protein, exposing charged and polar residues to solvent. These factors also give rise to stable substructures that occur ubiquitously in protein families (called *secondary structure*), such as alpha helices and beta strands. One consequence of this folded protein structure is that each residue in a protein sequence is exposed to a different local environment, potentially resulting in different evolutionary constraints at each site.
- **In general, protein structure is more conserved during evolution than protein sequence.** In the SCOP database of protein structure classification [49], there are over 2100 protein structural families of homologous sequences, while there are nearly 170,000 amino acid sequences in Pfam [8], a protein families database. A good rule of thumb is that two sequences with more than 30% sequence identity are likely to be homologous and fold into the same tertiary structure or domain, although homology can sometimes be inferred at a lower identity. The conservation of function is less clear-cut; sometimes proteins can perform the same function with less than 15% identity, while in other cases the function can be completely altered by the mutation of a few key amino acids.
- **Protein sequences are generally subject to greater selective constraint than noncoding DNA sequences.** For a protein to function properly in an organism, it must be transcribed and transported, fold, interact with its binding partners or substrate, perform its function efficiently (catalysis, recognition, transport, etc.), and be properly disposed of when no longer needed. If a mutation in the amino acid sequence affects any of these steps, it can potentially affect the function of the protein and therefore the fitness of the organism. The combination of these factors constrains the evolution of protein sequences much more than the evolution of most noncoding DNA sequences, an effect that can be observed as a low nonsynonymous/synonymous substitution rate ratio. This allows protein sequences to be used to infer homology among more distant evolutionary

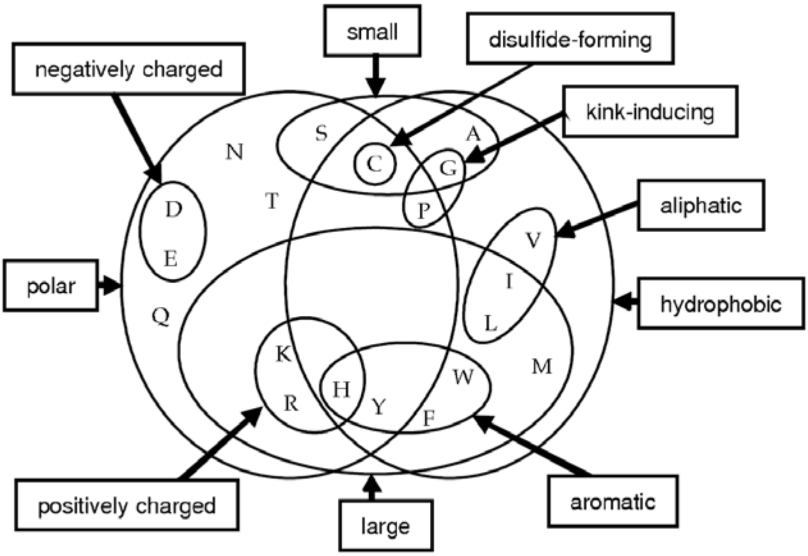

Fig. 9.1. Some commonly used amino acid properties. For other properties, a variety of empirically determined scales have been collected in the AAindex database [39].

relatives, but it also complicates phylogenetic inference because the effect of these constraints can be difficult to predict.

- **Amino acids each have unique properties that are utilized in different contexts within the protein.** Figure 9.1 shows several different sets of physicochemical properties; the importance of each property depends upon the local environment of the amino acid. For example, tyrosine is a bulky, aromatic amino acid that is often found in the same context as other large aromatics such as tryptophan and phenylalanine. But tyrosine also has a polar hydroxyl group like serine and threonine, and like those amino acids it can be involved in hydrogen-bonding interactions as well. In modeling protein evolution, it is therefore important to take into account not only the properties of the amino acids but also the context in which they are used.

9.3 The REV Model

A Markov model of protein evolution must at a minimum provide a substitution probability matrix $\mathbf{P}(t)$, where $P_{ij}(t)$ is the probability that an amino acid substitution $i \rightarrow j$ will occur in evolutionary time t. In the likelihood function (see Chapter 2), a different $\mathbf{P}$-matrix is needed for each branch of the tree. Rather than estimating each matrix separately, typically a single *instantaneous transition rate matrix*[1] **Q** is used for the entire tree, and the different **P**-matrices can be calculated using the standard equation for a continuous Markov process:

$$P_{ij}(t) = \left[e^{\mathbf{Q}t}\right]_{ij} \ . \tag{9.1}$$

The challenge for any model of protein evolution is to determine the best estimate of the transition rates in the **Q**-matrix given the available data. In general these rates are not equal; some amino acid substitutions are more likely than others, and this is directly related to the physical and chemical characteristics of the amino acids in protein structures.

9.3.1 Counting Methods for Model Estimation

Early models of protein evolution were limited by the computational issues associated with likelihood inference on phylogenetic trees. To overcome these limitations, empirical methods were devised to approximate the transition matrix by counting the number of inferred substitutions. The first widely used model of protein evolution was developed by Margaret Dayhoff and co-workers in the *Atlas of Protein Sequence and Structure* [17]. Using sets of closely related protein sequences (more than 85% similar), they used an assumption of parsimony to infer the ancestral amino acids at each site in the protein. Once the amino acid at each node has been determined, the internodal substitutions can be counted, resulting in a 20 × 20 symmetric matrix of amino acid replacement counts (the **A**-matrix).

These counts depend on the exposure of each amino acid during the evolutionary process; a rare but mutable amino acid can be indistinguishable from a common amino acid that rarely changes. To discriminate between these possibilities, Dayhoff defined the mutability of an amino acid m_i as the number of inferred changes for each amino acid divided by its total number of nodal appearances on the tree N_i:

$$m_i = \frac{\sum_{k \neq i} A_{ik}}{N_i} \ . \tag{9.2}$$

The mutabilities and the count matrix were then multiplied to calculate **M**, the "mutation probability matrix"[2],

$$M_{ij} = \lambda m_i \frac{A_{ij}}{\sum_{k \neq i} A_{ik}} \quad \text{for} \quad j \neq i \ , \tag{9.3}$$

[1] Because "transition" has a formal meaning when dealing with sequence evolution, the Markov transition rate matrix **Q** is sometimes called a mutation rate matrix or substitution rate matrix in the biological literature.

[2] For consistency with standard Markov model notation, the i, j indices have been reversed from Dayhoff's original notation.

where λ is a scaling factor that determines the average probability of a substitution in a specified unit of evolutionary time. Because this is a probability matrix, each row must sum to 1, and therefore the diagonals are

$$M_{ii} = 1 - \lambda m_i \ . \tag{9.4}$$

When λ is set so that the mean probability of substitution is 0.01, on average one substitution will be observed per 100 amino acid sites. This is called 1 PAM, or "point accepted mutation," a commonly used measure of distance between protein sequences.

The Dayhoff model was originally developed to aid in alignment of distant homologs and to help determine evolutionary distances between sequences. To convert it into a continuous-time Markov chain model that can be used for statistical inference and likelihood calculations [22, 41, 1], the matrix is commonly converted into a slightly different form, the symmetric "relative rate matrix" **R**:

$$R_{ij} = \frac{A_{ij}}{N_i N_j} \ . \tag{9.5}$$

Typically this is estimated empirically from the sequence alignment since in a stationary process with sufficient data $\pi_j^{obs} \to \pi_j$. The instantaneous transition matrix **Q** is then defined as

$$Q_{ij} = \delta R_{ij} \pi_j / s \quad \text{for} \quad j \neq i \tag{9.6}$$

and

$$Q_{ii} = - \sum_{k \neq i} Q_{ik} \ , \tag{9.7}$$

where π_j is the stationary frequency of amino acid j, δ is the number of expected substitutions per site in a unit of evolutionary time (typically 0.01), and s is a normalization constant,

$$s = \sum_{i,j,i \neq j} \pi_i \pi_j R_{ij} \ . \tag{9.8}$$

This **R**-matrix parameterization is generally called the REV model.[3] As the most general reversible amino acid model, it is analogous in form to the GTR model for nucleotide substitution. One important difference is that in contrast with the GTR nucleotide model, the REV matrix values are fixed and not estimated for each new dataset of interest. Similar empirical counting methods have been used to update the REV model parameters as more data have become available. Dayhoff's matrix represents data from 1572 counted substitutions; in 1992, Jones et al. updated the matrix parameters using 59,190

[3] The values of the Dayhoff **R**-matrix are different from the Dayhoff log-odds PAM matrix used for sequence alignment (although they use the same underlying **A** count matrix), so care should be taken not to confuse the two.

substitutions from 16,130 protein sequences in what is now commonly called the JTT model [37]. These sequences were generally globular proteins in an aqueous solvent; a model has also been tallied for transmembrane proteins and is called the tmREV matrix [38]. Each of these REV models–Dayhoff, JTT, and tmREV–differs only in the particular (fixed) values of their **R**-matrices.

Counting methods are relatively rapid, but they can only utilize the information from closely related sequence comparisons. To test the effects of this assumption, Benner, Cohen, and Gonnet [9] computed a set of matrices from proteins related by different PAM distances. When comparing these with extrapolated Dayhoff matrices, they found that the Dayhoff parameter values reflected the structure of the genetic code, while over long timescales the more accurate matrix values were better correlated with physicochemical properties of the amino acids. Even with methods that allow more divergent sequences to be used, the parsimony assumption inherent in the counting methods can cause bias in parameter estimation [16].

9.3.2 Likelihood Methods for Model Estimation

Maximum likelihood (ML) methods are a natural choice for optimizing models over divergent datasets [22, 76], as they can account for the probability of multiple substitutions over long branches and can be tested using a rigorous statistical framework. The likelihood equation for a phylogenetic tree utilizes a continuous-time Markov chain, which determines the probability of substitution over the evolutionary time t of each branch by exponentiating the **Q**-matrix as shown in equation (9.1). This equation can be approximated as

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \approx \mathbf{1} + \mathbf{Q}t + (\mathbf{Q}t)^2/2 + \dots . \tag{9.9}$$

The higher-order terms in this expansion account for the nonzero probability of multiple substitutions over long branches. As t increases and/or the off-diagonal terms in **Q** increase, the higher-order terms become more significant and the assumptions of parsimony no longer hold true.

The first use of ML estimation (MLE) methods to optimize REV model parameters was by Adachi and Hasegawa [1], who were interested in modeling the evolution of mitochondrial proteins. The Dayhoff and JTT matrices were developed as an average over many protein families from the nuclear genome, but mitochondrial proteins evolve under different selective constraints. Translated using a different genetic code with a different nucleotide compositional bias, most mitochondrial proteins also function in a lipid membrane rather than in aqueous solution. To account for these differences, the mtREV model [1] was developed on a tree of mitochondrial proteins from a diverse set of vertebrate species. Instead of using a counting method to infer the values of the **R**-matrix, each of the R_{ij} values was treated as a free parameter of the model and estimated using maximum likelihood. The MLE REV model was therefore estimated with 210 parameters: the 190 values of the symmetric **R**-matrix and the 20 amino acid frequencies. (The model has 208 degrees of freedom, because the **R**-matrix values are relative and the amino acid frequencies must sum to 1.) The resulting mtREV matrix had a significantly higher likelihood on mitochondrial proteins than the JTT matrix, and some of the known differences between mitochondrial and nuclear proteins are evident from mtREV's parameter values. For example, the Arg↔Lys substitution rate is much lower in mtREV than in JTT, a difference that is attributed to the fact that it requires two nucleotide mutations in the mitochondrial genetic code while only requiring one in the universal code. Other MLE REV models developed for specific datasets include the cpREV model for chloroplast proteins [2] and the rtREV model for retroviral polymerase proteins [19]. As with the Dayhoff model, all of these ML-estimated REV models have their **R**-matrix values fixed for further analyses; they are not adjusted for each new dataset.

For a general model applicable to many different protein families, the MLE equivalent of the Dayhoff and JTT matrices is the WAG matrix [74]. To create this matrix, 3905 protein sequences were divided into 182 protein families. A neighbor-joining tree was inferred for each family, and then the combined likelihood was maximized by adjusting the values of the **R**-matrix. Using the likelihood ratio test (LRT), the increase in likelihood over the former models was found to be statistically significant for all families in the analyzed dataset. In fact, the increase in likelihood from the JTT matrix to the WAG matrix is even greater than the increase from Dayhoff to JTT, despite the fact that WAG was optimized using fewer protein sequences than JTT, an indication of the power of the ML estimation method. Because ML estimation can be computationally expensive, approximate methods have been developed as a compromise between accuracy and speed [54].

The selective constraints acting on the amino acid level are reflected in the parameter values for these REV models. For example, in the universal genetic code, Ala is fourfold degenerate–represented by the codons GC*–while Trp is only represented by one codon (UGG). Therefore, there are six possible nucleotide mutations away from Ala and nine mutations away from Trp. If there were no selection on the amino acid level, one would predict from entropic principles that Trp would show a greater propensity for substitution than Ala since there are more "escape routes." But according to the mutabilities calculated for example by Jones et al. [37], tryptophan has the *lowest* mutability, while alanine has a mutability four times larger, an effect due in part to tryptophan's unique chemical characteristics. In the mtREV model, Cys is more mutable than it is in matrices optimized on proteins that function in an intracellular environment [1], probably because Cys-Cys disulfide bonds are not thought to be as important to membrane proteins as they are to aqueous proteins. The importance of such factors can be tested by comparison with the codon Poisson model [41], which disallows all single-step amino acid substitutions requiring more than one nucleotide mutation ($R_{ij} = 0$), while all other R_{ij} values are set to 1. This simple model is almost always statistically rejected in favor of models such as the Dayhoff model, an indication that simple nucleotide models are inadequate for modeling protein sequences because they do not account for the different properties of amino acid residues.

The simplicity and generality of the REV model are attractive, and its similarity in form to nucleotide models has made its implementation in phylogenetic software a fairly straightforward task. But this simplicity can be a limitation when attempting to understand the complex determinants of evolution at the protein level. Alternative models have concentrated on correcting some of these limitations, focusing on four features of amino acid sequence evolution: site heterogeneity, time heterogeneity, site dependence, and the physicochemical properties of amino acids. The rest of this chapter will be devoted to a discussion of these features and the models that address them.

9.4 Modeling Heterogeneity Across Sites

The parameter values in the REV models are typically an average over many amino acid sites from many different proteins. The implicit assumption is that every site in the protein is subject to the same evolutionary constraints, or at least that the constraints are evenly distributed about some mean value. But most proteins fold into an intricate three-dimensional structure, creating a different chemical environment for each amino acid residue. This heterogeneity in environments leads to heterogeneity in evolutionary constraints, which can have a dramatic effect on protein evolution and inference [56] (see Figure 9.2). The concept of a single transition matrix that can describe the evolutionary process at every site becomes difficult to justify.

9.4.1 Rate Heterogeneity Across Sites (RHAS)

One useful approximation for modeling site heterogeneity is the use of a distribution of evolutionary rates, or rate heterogeneity across sites (RHAS). According to the Neutral Theory [40], functionally important sites are under more stringent evolutionary constraints and will therefore exhibit a lower overall substitution rate. One of the simplest methods for adding rate heterogeneity to a phylogenetic model is to use a Gamma distribution of rates [77]. This is done exactly as with the nucleotide models (see Chapter 1), where each site is assigned an equal prior probability ϕ_k of evolving at rate λ_k. The possible values for λ_k are drawn from a discretized Gamma distribution with a specified number of categories K. The likelihood function in each column in the alignment D_n is determined by summing the conditional likelihood for each possible rate:

$$L\left(D_n|\theta'\right) \equiv \sum_{k=1}^{K} f\left(D_n|\lambda_k, \theta'\right) \phi_k \ . \tag{9.10}$$

In this case, θ' represents the other parameters in the model; for example, the **R**-matrix and amino acid frequencies. The shape of the Gamma distribution

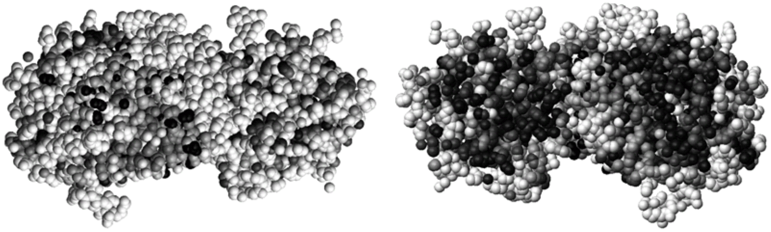

Fig. 9.2. An empirical Bayesian mapping of rates onto sites in the trypsin family. The darker the residue, the slower the rate of evolution at that site. Residues on the surface of the protein (*left*) tend to be less conserved than those seen in a cutaway view of the interior (*right*). The posteriors were calculated using CONSURF [63] and mapped onto the structure using MOLMOL [43].

can be adjusted with a single parameter, while the relative rates of substitution at any particular site are determined by a REV model. The result is a set of K rate categories, each related by a common REV model but multiplied by a different rate constant to determine the **P**-matrices:

$$P_{ij}^{k}(t) = \left[e^{\mathbf{Q}\lambda_k t}\right]_{ij} . \qquad (9.11)$$

The improvement in likelihood with a rate distribution is almost always significant relative to a site-homogeneous model even with just a few rate categories, making the Gamma distribution a commonly used approximation. This difference is not just a statistical nuance; failure to account for rate heterogeneity can also cause errors when inferring phylogenies and divergence times [13]. For this reason, the inclusion of some form of site heterogeneity–at least a REV+Γ model–is almost always recommended for phylogenetic analysis.

9.4.2 Pattern Heterogeneity Across Sites (PHAS)

The fact that rate heterogeneity is ubiquitous among protein sequences is evidence of the diversity of selective constraints operating on the amino acid level. Still, a rate distribution cannot account for variability in the *pattern* of evolutionary constraints. It allows a site to evolve more slowly or quickly, but a simple rate distribution does not, for example, allow a Gly→Ala substitution rate to be higher than a Gly→Pro substitution at one site but lower at another. Due to the diversity of amino acid environments in a folded protein, such differences can be pronounced. For example, glycine and alanine are the smallest amino acids, while proline (although still small) is bulkier. In the folded core of the protein, where steric constraints might preclude a bulky amino acid, the Gly→Pro substitution may be less favorable than the more conservative Gly→Ala substitution. But glycine and proline are also known to induce kinks in the protein chain, and these kinks are sometimes necessary for terminating alpha helices and creating turns. In these locations, a Gly→Pro substitution becomes the "conservative" one, accepted more often than a Gly→Ala substitution. This can only be modeled by a change in the *relative* rates of substitutions, not just the overall rate.

To account for this type of heterogeneity (called here pattern heterogeneity across sites, or PHAS), matrices have been estimated for specific structural classes of sites [59, 73, 44, 53]. For example, Koshi and Goldstein [44] divided sites into four different structural categories–helix (H), turn (T), strand/sheet (E), and coil (C)–and subdivided those into two accessibility categories: buried (b) and exposed (e). Then ML estimation was used along with an evolutionary tree to estimate structure-specific substitution matrices for each category.

The Koshi-Goldstein structure-based matrices were log-odds matrices designed for sequence alignment and structural prediction rather than Markov substitution matrices. To optimize matrices for phylogenetic use, Goldman and co-workers used an across-sites hidden Markov model (HMM) called the PASSML model [29]. PASSML begins with the assumption that any sequence site is in one of the eight structural categories mentioned above. These categories are further divided into a total of 38 possible classes by position along the sequence. For example, there are six each of the possible buried and exposed sheet classes [Eb_i, Ee_i ($i \in \{1, 2, ..., 6\}$)], ten each of the buried and exposed helix classes, two buried and two exposed turn classes, and one buried and one exposed coil class. This seemingly complicated division has an empirical basis: if each sequence site were completely independent (with no $i, i+1$ dependence) and transitions between structural categories were random, the length of each structure in a protein would be geometrically distributed with a mean of 1. This is physically unrealistic; helices and sheets by definition involve more than one amino acid. By adding in site dependence with an HMM, the PASSML model's mean structure length better resembles the empirically observed distribution.

A large training database of over 200 globular protein families with known structure was used to estimate PASSML's parameters, with the **R**-matrix for each site category estimated using a technique similar to the Dayhoff counting method. The result was a set of eight **Q**-matrices and their associated equilibrium amino acid frequencies (one set for each combination of secondary structure type and solvent accessibility):

$$\theta_{\text{passml}} = \{\mathbf{Q}_k, \boldsymbol{\pi}_k\} \text{ for } k \in \{1, 2, ..., 8\} \quad . \tag{9.12}$$

To model site dependence, the PASSML model also includes a set of ρ_{kl} parameters, the transition probabilities between the hidden classes that were estimated by empirical fit to the data. Once the model was estimated on the training set, all parameter values were then fixed for further analysis.

To apply PASSML to nontraining datasets, it is not necessary to know the protein's structure. Because this is an HMM, the true state of a site is

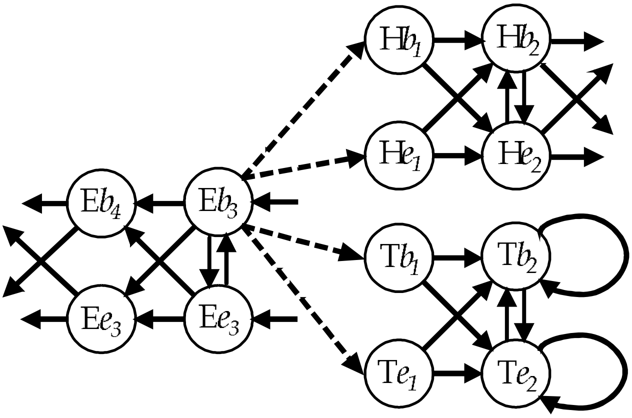

Fig. 9.3. An example of some allowed transitions within categories (solid lines) and between categories (dashed lines) in the PASSML model. The structural categories shown are helix, sheet, and turn (H, E, and T); the accessibility categories are buried and exposed (*b* and *e*). Not all transitions or states are shown. For more details, see Goldman et al. [29]

considered to be "hidden", and the likelihood is a sum over the conditional likelihood of each site class at each site,

$$L(S_n|T) = \sum_{k_n} f(S_n, k_n|T) \;, \tag{9.13}$$

where

$$f(S_n, k_n|T) = \sum_{k_{n-1}} f(D_n|k_n, T)\mathrm{P}(S_{n-1}, k_{n-1}|T)\rho_{k_{n-1}k_n} \;. \tag{9.14}$$

Here D_n denotes the nth column in the N-column alignment, S_n denotes the set of columns $\{D_x\}, x \in \{1, 2, ..., n\}$, k_n is the structural category for site n, $\rho_{k_{n-1}k_n}$ is the transition probability from a category at column $n-1$ to a category at column n, and T is the tree topology. The difference between this equation and an alignment-based HMM approach (see Chapter 14) is in the likelihood function; in this case $f(D_n|k_n, T)$ is the phylogenetic likelihood function for site n.

Using likelihood ratio tests on several different protein families, the authors found that simple models that did not include structural categories were always rejected in favor of those that did. Even HMMs with just eight site classes with no site-to-site dependence (all ρ_{kl} equal) yielded a much higher likelihood. (Each site class's mean rate is also variable, so rate heterogeneity is an implicit feature of the model.) The inclusion of additional solvent accessibility categories was also found to be significant, but the dependence among adjacent sites was a less important feature of the model. Using the same technique but different category designations, the PASSML-TM model [47] and

MT126 model [48] have been optimized for transmembrane and mitochondrial proteins, respectively.

An interesting twist on structure-based evolutionary modeling uses simulated evolution on a known protein structure to create substitution matrices. The IS-SCPE method [25] (for Independent Sites–Structurally Constrained Protein Evolution) requires a representative structure and sequence for the protein family of interest. The sequence is repeatedly mutated, and the mean field energies at each site in the structure are computed using prespecified energy potentials. The structural perturbation of the new sequence from the reference structure is then calculated, and mutations causing a perturbation smaller than a specified cutoff are accepted. Finally, the accepted mutations are tallied in a set of replacement count matrices that are categorized by structural class (for example, position in an alpha helix).

Matrices created using this method were found to have a significantly higher likelihood than the JTT+Γ model, another indication that rate heterogeneity alone can sometimes be insufficient for modeling and that pattern heterogeneity also plays a large role. The IS-SCPE method is promising for proteins of known structure, although its assumptions that energetic stability and local structural integrity are important evolutionary constraints should be kept in mind when this method is applied.

9.5 Mechanistic Models

The models discussed in Section 9.4 emphasize structural features that are often easily observable: alpha helices, beta sheets, solvent accessibility, structural stability, etc. The ubiquity and strong conservation of such features favor this assumption, but protein evolution can also be affected by subtleties that are difficult to ascertain a priori. Transient recognition binding patches, allosteric regulatory networks, and dynamic hinge regions are just a few examples of evolutionary constraints that may be crucial but not obvious, even when the protein structure is available. In fact, there is evidence indicating that many functional regions of proteins are disordered and do not exist in a single stable structural state [20].

When little is known about the structural determinants of evolution in a protein family, ideally one would prefer to estimate the model's parameters for each dataset of interest. This is typical when applying nucleotide models such as the Jukes-Cantor or GTR models, where the ML estimates of the parameters are jointly estimated while searching for the ML tree topology. Contrast this with all the amino acid models mentioned thus far, which have been trained on a set of reference sequences or a reference structure and then the parameters are fixed for further analysis. The reasons for this are both theoretical and practical. The GTR model has only six parameters (ten if the nucleotide frequencies are also estimated), while the full REV model has 190 parameters (or 210 with estimated frequencies). When using a PHAS model, this number is multiplied by the number of site classes. A model with eight site classes, each represented by a REV matrix, can have over 1600 parameters! Precise estimation with so many degrees of freedom requires an extremely large dataset. Even when hundreds of sequences are used, the time required for model estimation can be prohibitive, especially when simultaneously estimating the tree topology and branch lengths.

One way to reduce the number of parameters is to use a mechanistic model. The parameters of the models discussed previously can all be considered to be somewhat empirical: substitutions are tallied in parameters that have more statistical convenience than physical meaning. In reality, these relative rates are an aggregate measure of the physicochemical characteristics of the amino acids and their interactions with their local environment. In contrast, mechanistic models explicitly utilize these physicochemical characteristics, facilitating the testing of hypotheses related to these properties. This reparameterization reduces the degrees of freedom, allowing the use of a realistic number of site classes while estimating mechanistic model parameters for each dataset of interest. Mechanistic models are frequently used in combination with multiple site classes; in these cases they are a type of PHAS model.

Several examples of mechanistic models can be summarized as physicochemical amino acid fitness models [45, 79]. In these types of models the $\mathbf{Q}$-matrix for each site class k is divided into a mutation rate λ and an amino acid substitution function Ω_{ij}^k:

$$Q_{ij}^k = \lambda \Omega_{ij}^k(F_i^k, F_j^k) \ . \tag{9.15}$$

F_i^k and F_j^k are amino acid fitnesses,[4] parameters that are explicitly dependent upon the physicochemical properties of the amino acids. For example, in the model of Koshi and Goldstein [45] (called the FIT-PC model here), these fitnesses are determined as quadratic functions of the amino acid's hydrophobicity (h) and volume (v):

$$F_i^k = a_k \left(h_i - h_o^k\right)^2 + b_k \left(v_i - v_o^k\right)^2 \ . \tag{9.16}$$

In this model, a_k, b_k, h_o^k, and v_o^k are all parameters of the model and estimated from the data using maximum likelihood. The first two parameters (a_k and b_k) determine the strength of the selective pressure from each chemical characteristic, while h_o^k and v_o^k determine the optimal value of that characteristic in the site class. The substitution rate for nonsynonymous changes in the FIT-PC model is determined by a fitness function,

[4] These parameters have been called fitnesses as an analogy to fitness functions on an energy landscape rather than as fitnesses in the genetic sense of the term. Nevertheless, they could be made mathematically equivalent with the proper choice of fitness function. Also, in a reckless abuse of notation, a superscript k will indicate that the parameter is particular to that site class, not that k is a numerical exponent.

$$\Omega_{ij}^k = \begin{cases} \omega_k e^{\left(F_j^k - F_i^k\right)}, & \text{if } F_j^k < F_i^k, \\ \omega_k, & \text{if } F_j^k \geq F_i^k. \end{cases} \tag{9.17}$$

The value of the parameter ω_k is estimated using ML, and it can be regarded as a general selective disadvantage for making any nonsynonymous change (or as an adaptive advantage if $\omega_k > 1$). This type of fitness function is also known as a Metropolis-Hastings function; its form is chosen for its mathematical convenience, as it allows the straightforward derivation of the equilibrium frequencies for each site class as a function of the fitnesses:

$$\pi_i^k = \frac{e^{F_i^k}}{\sum_{i'} e^{F_{i'}^k}} \ . \tag{9.18}$$

Qualitatively, favorable mutations to "more fit" amino acid are all accepted at the same rate, while unfavorable mutations are tolerated depending on the difference between the amino acid properties. The larger the difference, the lower the substitution probability.

There are several alternatives for calculating λ in equation (9.15). One possibility is to set λ as an estimated parameter. This can only be done if ω_k is fixed, as they are indistinguishable on the amino acid level (one is actually estimating $\{\lambda\omega\}_k$). Another possibility is to use a Gamma rate distribution to subdivide each site class into r rate categories,

$$Q_{ij}^{kr} = \Omega_{ij}^k \lambda^{kr} \ , \tag{9.19}$$

where each λ^{kr} is determined from category r of a discretized $\Gamma(\alpha, \omega_k\alpha)$ distribution that has a mean ω_k (see Chapter 5).

A third possibility, suggested by Yang et al. [79], is to specify λ as a weighted sum of the mutation rates on the codon level, independent of site class k but dependent on the set of codons $u \in i$ and $v \in j$ coding for amino acids i and j, respectively:

$$\lambda_{ij} = \sum_{u \in i} \sum_{v \in j} \lambda_{uv} \left(\frac{\pi_u}{\sum_{u' \in i} \pi_{u'}} \right) . \tag{9.20}$$

The frequency of codon u, π_u, can be estimated empirically from the data. λ_{uv} can itself be set as a function of mechanistic parameters on the nucleotide level such as the transition/transversion rate ratio κ:

$$\lambda_{uv} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_v, & \text{for transversion,} \\ \kappa\pi_v, & \text{for transition.} \end{cases} \tag{9.21}$$

Because these fitness models do not make prior assumptions about which fitness characteristics best describe each specific site, they must deal with the issue of how to assign the site classes. For example, a large amino acid may be appropriate for some sites in the protein (represented by high values of v_o^k and b_k), while a small amino acid might be important at another site. To avoid any prior assumptions about which sites are represented by which selective constraints, *hidden site classes* are used. These are analogous to the hidden rate classes used by the Gamma distribution and to the hidden Markov classes of the PASSML model, albeit without site dependence. Rather than explicitly assigning classes to sites, all sites instead have a prior probability ϕ_k of being modeled by each site class k. The likelihood at each column in the alignment D_n is the sum of the conditional probabilities weighted by these prior distributions:

$$L\left(D_n|\theta\right) \equiv \sum_k f(D_n|\theta_k, k)\phi_k \ . \tag{9.22}$$

Typically the prior distribution are set equal for all classes and all sites (flat priors), although they can be specified on a site-by-site basis if prior information about each site is to be included in the model.

Under the simplifying assumption that $\lambda = 1$, the FIT-PC model has just five parameters per site class (a_k, b_k, h_o, v_o, and ω_k) compared with over 200 for a REV model. This allows the values of the parameters to be ML-estimated for each dataset of interest rather than estimated on a training set and then fixed as with the REV models. The FIT-PC model generally yields higher likelihoods than site-homogeneous REV models once a moderate number of site classes is specified (generally five or more) [45, 18]. The fact that higher likelihoods can be achieved even with such a simplified model is further evidence of the importance of site heterogeneity in protein modeling.

Other parameterizations of the fitnesses and fitness function have also been explored [79, 78]. For example, in the DIST-PC model [79], the fitnesses can more accurately be called distances, where the distance d_{ij} between amino acids is (in their example) based on polarity (p) and volume (v) [52]:

$$d_{ij} = \sqrt{(p_i - p_j)^2/\sigma^2_{\Delta p} + (v_i - v_j)^2/\sigma^2_{\Delta v}} \ . \tag{9.23}$$

Here $\sigma^2_{\Delta p}$ and $\sigma^2_{\Delta v}$ are the standard deviations of $|p_i - p_j|$ and $|v_i - v_j|$, respectively. The substitution rate is an exponential function of this distance:

$$\Omega^k_{ij} = \omega_k e^{-(b_k d_{ij}/d_{\max})} \ . \tag{9.24}$$

The parameter b_k is a measure of the strength of selection upon the particular physical properties of the amino acid; a larger value of b_k indicates that more radical amino acid changes are less likely to be accepted as substitutions. The λ parameter can be specified as described above in the FIT-PC model. The differences between the Ω functions of the two models reflect two distinct philosophies about the manner in which evolution proceeds. The DIST-PC function can be thought of as a neutral walk through the fitness landscape; what matters most is not the direction of changes but whether or not they are conservative or radical. Given enough mutational steps, an amino acid position could change from small to large with little penalty. As a result, the equilibrium amino acid frequencies for the Ω function are all equal in the DIST-PC model if reversibility is assumed. By contrast, the FIT-PC model assumes that the protein site has an optimal set of physicochemical properties, and the favorability of an amino acid change is measured relative to both the former amino acid and this ideal value. Favorable mutations are all accepted at the same rate, while unfavorable ones are tolerated depending on their distance from the former amino acid's properties. In reality, the sites in a protein probably evolve in a mixture of these two regimes, and so a mixture of site classes or fitness functions can be appropriate [68].

The FIT-PC and DIST-PC models relax assumptions about which protein structural types are important, but they still require the specification of particular amino acid characteristics. Hydrophobicity, bulk, and polarity have been shown to be three of the most dominant [52, 44, 70], but other characteristics are sometimes crucial, such as the turn-inducing properties of glycine and proline or the delocalized electrons of the aromatic amino acids. To avoid any assumptions about which characteristics are important, a general fitness model can be used [18]. FIT-GEN is nested with the FIT-PC model, instead setting each F_i in equation (9.17) as a free parameter rather than as a function of physicochemical properties. This yields 21 parameters per site class: the 20 amino acid fitnesses F_i^k and the nonsynonymous rate ω_k (there are 20 free parameters because the fitnesses are relative). With adequate data, FIT-GEN is better able to capture the nuances of evolution than FIT-PC at the cost of some simplicity, while still using 188 fewer parameters per site class than a REV model. FIT-GEN can be used in an iterative manner with FIT-PC; general fitnesses can first be determined, and these can then be correlated with physicochemical characteristics. The dominant characteristics can then be utilized in a FIT-PC or DIST-PC model for later analysis on the same protein family.

The FIT-GEN model still assumes a specific number of site classes; the most general approach would be to assign one site class per location in the protein. Bruno [11] used an EM algorithm to obtain site-specific amino acid frequencies, with one frequency vector per site. The obstacle then becomes a lack of data; at short evolutionary distances, the inferred substitutions at each site may be just a fraction of the allowable substitutions, so a large, diverse sequence set is required. Although the parameters from this method are not directly applicable to phylogenetic analysis, they can provide a starting point for further site classifications, such as by principle component analysis [45] or as initial groupings in a FIT-GEN model.

Part of the power of these types of mechanistic hidden site class models is that they lend themselves well to empirical Bayesian mapping [58]. In this technique, the posterior probability of each site class k at each site n can easily be calculated using the likelihood and the prior distributions:

$$\Pr(k|D_n, \theta) = \frac{f(D_n|\theta_k, k)\phi_k}{\sum_{k'} f(D_n|\theta_k, k')\phi_{k'}} \ . \tag{9.25}$$

These posteriors probabilities can then be mapped onto the sequence alignment or protein structure of interest to determine which sites are more likely to be evolving under the different selective constraints [78, 5, 68, 69]. For example, site class 1 could model a fitness function based on polarity, site class 2 on bulk, and so on. When the posterior probabilities are mapped onto the sequence alignment, sites where bulk has been more important to evolution than polarity will have a higher posterior probability for site class 2. When mapped onto the protein structure, these posterior probabilities can reveal important evolutionary features such as transmembrane regions and dimerization interfaces [69]. Empirical Bayesian mapping is not limited to PHAS models; it has also been applied to RHAS models to map rate heterogeneity onto protein structures [63] (see Figure 9.2).

9.6 Modeling Heterogeneity over Time

The phylogenetic models discussed above assume that the rate and pattern of evolution have remained constant over the entire evolutionary tree, an assumption called *homotachy* [50]. This assumption can be violated when a protein is adapting to a new function or structure; according to the Neutral Theory, sites that are involved in the change in function will appear to evolve at a different rate. By developing models that allow a change in rate over time, these types of functional shifts can be detected.

The concept of an explicitly heterotachous model (or RHAT model, for Rate Heterogeneity Across Time) was first outlined in a maximum parsimony framework as the covarion model [23]. With this model, only a fraction of the sites in a protein-coding gene are “on” and can accept mutations: the concomitantly variable codons. All others are “off”, and no substitutions are observed; these sites are assumed to be completely functionally constrained. A site may switch from “on” to “off” (and vice versa) with a certain persistence time, indicating that the site has acquired (or lost) functional significance.

More recently, covarion-like models for proteins have been cast into a likelihood framework, allowing the application of likelihood ratio tests for hypothesis testing. In 1999, Gu developed a time-heterogeneous ML method and applied it to the detection of functional divergence between gene duplicates [34]. Gene duplication is thought to be a factory for evolutionary diversification; one copy of the gene can continue to perform its native function, while the other can be adapted for a distinct task [15]. This adaptation results in different rates of substitution for each of the two paralogous protein families, a phenomenon dubbed type I divergence.

Consider the subfamilies in Figure 9.4, with two possible states: S_0 and S_1. S_0 is the null hypothesis that there are no altered functional constraints

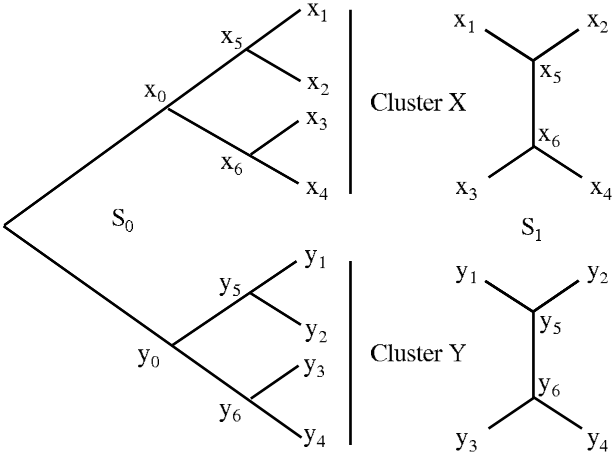

Fig. 9.4. In state S_0, both subfamilies are scaled by the same overall rate λ. In state S_1, each family subtree may have a different overall rate. Adapted from Gu [35].

in either subfamily. In this state, the substitution rates for each subfamily are completely correlated: $\lambda_X = \lambda_Y = \lambda$. The other possibility, S_1, is that the function of either or both subfamilies has diverged since the common ancestor, and therefore λ_X and λ_Y are treated as independent. The other parameter requiring estimation is θ_{12}, the probability that the site is in state S_1 (also called the coefficient of type I divergence).

To calculate the likelihoods, it is assumed that the subtrees are statistically independent, so that $f(X_n|\lambda_X)$ and $f(Y_n|\lambda_Y)$ are the likelihoods at site n for the unrooted subfamilies X and Y, respectively, conditional upon the rates for each subfamily. Since it is a difference in rate that is important and not the absolute rates, the values of λ_X and λ_Y are not explicitly specified but integrated out by using a Gamma rate distribution [77],

$$p(X_n) = E[f(X_n|\lambda)] = \sum_{\lambda'} f(X|\lambda')\phi(\lambda') \;, \tag{9.26}$$

where $\phi(\lambda')$ is the probability of each λ' partition from the Gamma distribution. The joint probabilities conditional on being in either state S_0 or S_1 can then be written as

$$\begin{aligned} f^*\left(X_n, Y_n|S_0\right) &= \sum_{\lambda'} f(X_n|\lambda')f(Y_n|\lambda')\phi(\lambda') \\ &= E\left[f(X_n|\lambda)f(Y_n|\lambda)\right] \;, \\ f^*\left(X_n, Y_n|S_1\right) &= p(X_n)p(Y_n) \\ &= E\left[f(X_n|\lambda_X)\right] \times E\left[f(Y|\lambda_Y)\right] \;. \end{aligned} \tag{9.27}$$

Finally, the full joint probability for the two subtrees at this site is

$$p^*(X_n, Y_n) = (1 - \theta_{12})\, f^*\left(X_n, Y_n|S_0\right) + \theta_{12} f^*\left(X_n, Y_n|S_1\right) \ , \tag{9.28}$$

where θ_{12} is a parameter called the coefficient of divergence. Over the whole tree, the likelihood is

$$L(X, Y|\text{data}) = \prod_n p^*(X_n, Y_n) \ . \tag{9.29}$$

The null hypothesis is that $\theta_{12} = 0$, while the alternate hypothesis of functional divergence is that $\theta_{12} > 0$; these can be compared with a likelihood ratio test.[5] As the support for a rate change in the data increases, so should the θ_{12} parameter.

This RHAT model has been shown to successfully detect functional divergence on a variety of protein families, including COX enzymes [35] and tyrosine kinases [33]. It has been extended to the comparison of multiple clusters [35] and for the detection of type II divergence, where the evolutionary rate immediately after duplication is different from that in either subfamily. Empirical Bayesian mapping has been applied to detect the specific sites most likely to be involved in the functional change [35, 42, 28], and a faster approximate method has been devised that uses the ML estimates of the substitution counts in each subfamily to test for significance [34].

Note that the RHAT model, like the RHAS model, specifies only the rate parameter in conjunction with a REV matrix such as JTT and does not address differences in the pattern of mutations. This could be important if, for example, a particular site may evolve at the same rate in two subfamilies but with positively charged residues selected in one subfamily and negatively charged residues selected in the other. These types of questions have been addressed for nucleotide models [72, 27, 36] and in qualitative fashion for proteins [69], but they have not yet been applied in a rigorously testable "PHAT" context for proteins. As an example, one could set S_0 as the null hypothesis that sites in the subtrees evolve with the same mechanistic site class ($k_{X_n} = k_{Y_n} = k$) and evaluate S_1 as the alternative hypothesis that k_{X_n} and k_{Y_n} are independent.

The existence of rate heterogeneity between widely divergent sequences and between paralogous protein subfamilies is generally well-accepted. But there is mounting evidence indicating that heterotachy, like site heterogeneity, may be quite common even within protein families where function is largely maintained [28]. For example, Lopez and co-workers [50] performed a thorough analysis of heterotachy on over 3000 sequences of vertebrate cytochrome b, a protein whose function in the electron transport pathway is generally conserved throughout vertebrates. They used a modified RHAT model to examine several evolutionary groupings of cytochrome b, finding significant evidence of protein heterotachy among birds, mammals, and fish, as well as among

[5] Since the θ_{12} parameter is at the boundary of its state space in the null model, the corrected χ^2 test should be used for significance testing [30].

four different groupings of murids. This significance was not caused by just a few extremely adaptive locations but was instead seen at a large percentage of sites. The fact that rates could vary significantly through time within a protein family with ostensibly conserved function is an indication that heterotachy may be an important component of protein models.

9.7 Modeling Correlated Evolution Between Sites

Most models of protein evolution treat sites independently, but this is mainly a mathematical convenience that helps to keep the likelihood equations tractable. In reality, a protein sequence does not generally function as an extended floppy chain, but as a globular structure where the amino acids pack tightly against one another. Since it is these interactions that determine the structure and function of a protein, there is significant interest in modeling correlated evolution between sites.

Correlation between sites can be classified as indirect or direct. Indirect correlation occurs when sites are in the same structural category and therefore subject to the same selective constraints. Their rates and patterns may be correlated, but a substitution at one site does not necessarily affect the rate of substitution at another. This type of correlation is the basis for models such as the PASSML models discussed previously. To measure the strength of correlation between adjacent sites, Gonnet and co-workers used a 400×400 dipeptide matrix [31]. This matrix was created using a parsimony-counting method similar to the Dayhoff method, but in this case there are 400 character states, one representing each two-residue pair. The resulting matrix was significantly different from matrices created by assuming site independence, indicating that nearby sites can undergo correlated evolution. For example, on average, conservation at the first position was likely to be correlated with conservation in the second, a reflection of the fact that nearby residues tend to be in the same types of environments.

Direct correlation, or coevolution, occurs when a substitution at one site alters the fitness landscape at other sites, potentially creating an adaptive evolutionary regime.[6] For example, the salt bridge is one type of stabilizing interaction in proteins, potentially formed when a positively charged amino acid residue is in the proximity of a negatively charged residue. If one member of the salt bridge mutates into an oppositely charged residue, it can destabilize the protein structure or disrupt its function. Assuming the mutation is accepted as a substitution, the salt bridge can be reestablished by a compensatory mutation at the other site, and the substitution rate can temporarily increase as a result [24]. Other possibilities for compensatory coevolution include small-large amino acid pairs and a polar-polar to nonpolar-nonpolar compensation.

[6] This is sometimes called "covariation," but that term is avoided here to minimize confusion with the covarion model mentioned in Section 9.6.

The degree to which such coevolution occurs is still debated. Most evidence seems to indicate that compensatory substitution does occur but that the prevalence is low [57, 66, 7, 46, 51, 32]. Some possible explanations for this weak signal are: (a) the first mutation is generally so deleterious that no chance for compensatory change is allowed, (b) an unfavorable substitution can be effectively compensated by subtle shifts in the protein structure, and/or (c) compensatory substitutions are important but occur at just a few sites in a particular protein, making them hard to detect among the many comparisons that must be made. Because of the potential for predicting protein structures and interactions, there has been significant interest in developing methods to detect the sites that may be strongly coevolving. Most of these methods have been primarily based upon detecting mutually informative sites in the alignment [67, 46, 6, 3, 21], but these types of methods can be misled unless proper correction due to evolutionary relationships is taken into account [61, 71, 75]. Even in methods that do explicitly utilize the phylogenetic tree, the tests are generally not based on a particular model of evolution [14, 26].

As an example of a coevolutionary Markov model, the site-independent fitness models in Section 9.5 are readily applicable to a coevolutionary framework by adding a correlation term:

$$F^{AB}(a,b) = F^{A}_{\text{ind}}(a) + F^{B}_{\text{ind}}(b) + F^{AB}_{\text{dep}}(a,b) \ . \tag{9.30}$$

In this equation, $F^{A}_{\text{ind}}(a)$ is the fitness for amino acid a if site A evolved independently of site B; for example, the fitness in equation (9.16) can be used. $F^{AB}_{\text{dep}}(a,b)$ is the coevolution term, an increase or decrease in the fitness of amino acid a due to the presence of amino acid b at site B. This dependent fitness function can itself be made mechanistic:

$$F^{AB}_{\text{dep}}(a,b) = \rho_{AB}\psi_{ab} \ . \tag{9.31}$$

ρ_{AB} is the strength of interaction between the site pairs, and $\mathbf{\Psi}$ is a symmetric interaction matrix, where ψ_{ab} describes the interaction between amino acids a and b. In the salt-bridge example given above, $\psi_{ab} > 0$ when a and b are of opposite charge, and $\psi_{ab} < 0$ when their charge has the same sign. Assuming $\rho_{AB} > 0$, the overall fitness F^{AB} will be increased when $\psi_{ab} > 0$–when the interaction between a and b is favorable. If a mutation is attempted at site A to amino acid j, when $\psi_{jb} > \psi_{ib}$, the result will potentially be an increase in Ω_{aj} and therefore Q_{aj}, the transition rate (equations (9.17) and (9.15)). When specifying $\mathbf{\Psi}$ using for example empirically determined contact energies, this model can be nested with the FIT-PC model; the two are equivalent when $\rho_{AB} = 0$ for a site pair. This model is similar to a codon-based model described in [64].

While such a coevolutionary fitness model is theoretically attractive, it is computationally impractical when performing full-likelihood calculations. One of the barriers to the development of any coevolutionary Markov model is the size of the state space. Instead of the 20 amino acid states in the site-independent model, there are $20 \times 20 = 400$ possible pairs of amino acids, leading to a 400 × 400 **Q**-matrix. Even if the data were available to estimate the ρ_{AB} parameter for each site pair, it is computationally expensive to exponentiate such a matrix and to use it in the phylogenetic likelihood function. (RNA coevolutionary models, with only $4 \times 4 = 16$ possible states, have had more success because they do not suffer from this limitation [55, 65].)

To simplify the state space, Pollock and co-workers [62] created a coevolutionary Markov model by reducing amino acids to two states (designated A and a or B and b, depending on their position in the pair). For example, all large amino acids might be designated A and small residues called a, or the split could be based on amino acid charge (positive or negative). There are then four possible states at an amino acid site pair—AB, Ab, aB, and ab—and the transition matrix is

$$\mathbf{Q} = \begin{array}{c} AB \\ Ab \\ aB \\ ab \end{array} \left\{ \begin{array}{cccc} -\sum_{AB} & \lambda_B\pi_{Ab}/\pi_A & \lambda_A\pi_{aB}/\pi_B & 0 \\ \lambda_B\pi_{AB}/\pi_A & -\sum_{Ab} & 0 & \lambda_A\pi_{ab}/\pi_B \\ \lambda_B\pi_{AB}/\pi_B & 0 & -\sum_{aB} & \lambda_B\pi_{ab}/\pi_a \\ 0 & \lambda_A\pi_{Ab}/\pi_b & \lambda_B\pi_{aB}/\pi_a & -\sum_{ab} \end{array} \right\} , \tag{9.32}$$

where λ_A and λ_B are the rates at each site and the π's are the stationary frequencies for each possible pair. The independent frequencies π_A and π_B are constrained by the pairwise frequencies:

$$\pi_A = \pi_{AB} + \pi_{Ab} \; . \tag{9.33}$$

The coevolutionary model has six parameters per site pair:

$$\theta_{\mathrm{coev}} = \{\lambda_A, \lambda_B, \pi_{AB}, \pi_{Ab}, \pi_{aB}, \pi_{ab}\} \; . \tag{9.34}$$

(Since the π values must sum to 1, one of them is constrained by the others, and there are five free parameters). This yields one degree of freedom in comparison with the site-independent model, which assumes that $\pi_{xy} = \pi_x\pi_y$. The degree of correlation can be examined as a residue disequilibrium value, $RD = \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}$, where a higher RD value indicates greater correlation. Pollock et al. found that the likelihood ratios did not fit the usual chi-squared distribution, so they used simulation to determine significance levels.

When applied to myoglobin, their model indicated the presence of coevolution, especially among neighboring sites, but as with most studies, the signal is weak. For example, they tested 2259 site pairs for coevolution using a charge metric to determine the character states. Due to Type I error, at the 5% significance level one would expect to erroneously report 113 pairs as false positives where no coevolution actually occurred. Pollock et al. found 158 significant pairs, indicating that 43 truly coevolving site pairs are probably mixed in with those 113. This is an example of the multiple testing problem that arises when testing all site pairs in a protein: the number of comparisons increases as the square of the number of sites, threatening to swamp the small number of true positives with false positives. Therefore, it is often important to reduce the number of comparisons either by making some prior assumptions about which sites are to be tested, by combining the data from groups of sites using total likelihood, or by making only relative comparisons.

9.8 Final Notes

Although protein evolutionary models have taken great strides since the formation of the Dayhoff matrix, their development and implementation are still nascent when compared with nucleotide models. For example, there is no commonly accepted hierarchy of nested protein models, and most of the more complex models detailed here have not been incorporated into popular tree-searching software packages. Therefore, to perform ML tree estimation on amino acid sequences, REV matrices such as JTT and WAG are generally the only options in commonly used software. At the very least, it is important to include rate heterogeneity among sites, such as with the $+\Gamma$ option, as failure to do so can cause errors in topology and divergence estimation [13]. Studies seem to tentatively indicate that the tree topology is fairly robust to model misspecification, as long as some site heterogeneity is included [12] in either RHAS or PHAS form. Therefore, it may be an adequate approximation to choose a credible set of trees using a REV$+\Gamma$ model and then test more detailed hypotheses with the specialized models. Nevertheless, the full potential of recent advances in protein modeling will not be realized until these models are better integrated with tree-searching methods.

Another practical decision is whether to use amino acid or codon models. Codons contain information about the underlying mutation rate, and this information can be valuable for detecting selection at a particular site or along a particular lineage (see Chapter 5). But with this increase in information comes a decrease in computational speed. The Felsenstein pruning algorithm for likelihood calculation [22] is $\mathcal{O}(N^3)$; computational time increases as the cube of the number of states. Since codon state space is over three times larger than amino acid state space, computations with amino acid models are generally about 27 times faster than with codon models. For larger datasets and/or longer divergence times, amino acid models are often a more pragmatic choice and may provide more information about the origin of evolutionary constraints such as protein structure and amino acid characteristics. For smaller, more closely related sets of sequences, codon models offer higher fidelity and may provide more information about the different "directions" of Darwinian selection (purifying or adaptive) that act upon the evolution of the protein.

One practical constraint on the development of protein phylogenetic models has been the computational time involved in ML estimation and significance testing. Bayesian phylogenetic methods hold great promise for alleviating these concerns. Bayesian methods can provide estimates of the variance on the parameters of interest and integrate over the uncertainty in other parameters, allowing models that are more complex than those estimated using ML methods. With recent computational strides in this field, it is possible that Bayesian methods may facilitate the combination of site dependence with rate, pattern, and time heterogeneity into a unified framework.

References

[1] J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol*, 42(4):459–468, Apr 1996.

[2] J. Adachi, P. J. Waddell, W. Martin, and M. Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J Mol Evol*, 50(4):348–358, Apr 2000.

[3] D. A. Afonnikov, D. Y. Oshchepkov, and N. A. Kolchanov. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. *Bioinformatics*, 17(11):1035–1046, Nov 2001.

[4] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, Jul 1973.

[5] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*, 19(6):950–958, Jun 2002.

[6] W. R. Atchley, K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol Biol Evol*, 17(1):164–178, Jan 2000.

[7] E. Azarya-Sprinzak, D. Naor, H. J. Wolfson, and R. Nussinov. Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Eng*, 10(10):1109–1122, Oct 1997.

[8] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res*, 30(1):276–280, Jan 2002.

[9] S. A. Benner, M. A. Cohen, and G. H. Gonnet. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*, 7(11):1323–1332, Nov 1994.

[10] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Publishing, New York, 1999.

[11] W. J. Bruno. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol Biol Evol*, 13(10):1368–1374, Dec 1996.

[12] T. R. Buckley. Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. *Syst Biol*, 51(3):509–523, Jun 2002.

[13] T. R. Buckley, C. Simon, and G. K. Chambers. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst Biol*, 50(1):67–86, Feb 2001.

[14] G. Chelvanayagam, A. Eggenschwiler, L. Knecht, G. H. Gonnet, and S. A. Benner. An analysis of simultaneous variation in protein structures. *Protein Eng*, 10(4):307–316, Apr 1997.

[15] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann. Evolution of the protein repertoire. *Science*, 300(5626):1701–1703, Jun 2003.

[16] T. M. Collins, P. H. Wimberger, and G. J. P. Naylor. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Sys Biol*, 43:482–496, 1994.

[17] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, chapter 22, pages 345–352. National Biomedical Research Foundation, Washington, DC, 1978.

[18] M. W. Dimmic, D. P. Mindell, and R. A. Goldstein. Modeling evolution at the protein level using an adjustable amino acid fitness model. In *Pacific Symposium on Biocomputing*, pages 18–29. World Scientific, Singapore, 2000.

[19] M. W. Dimmic, J. S. Rest, D. P. Mindell, and R. A. Goldstein. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J Mol Evol*, 55(1):65–73, Jul 2002.

[20] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, May 2002.

[21] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, Suppl 5:157–162, 2001. Evaluation Studies.

[22] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[23] W. M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet*, 4(5):579–593, Oct 1970.

[24] K. M. Flaherty, D. B. McKay, W. Kabsch, and K. C. Holmes. Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. *Proc Natl Acad Sci USA*, 88(11):5041–5045, Jun 1991.

[25] M. S. Fornasari, G. Parisi, and J. Echave. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol*, 19(3):352–356, Mar 2002, letter.

[26] K. Fukami-Kobayashi, D. R. Schreiber, and S. A. Benner. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J Mol Biol*, 319(3):729–743, Jun 2002.

[27] N. Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol*, 18(5):866–873, May 2001.

[28] E. A. Gaucher, X. Gu, M. M. Miyamoto, and S. A. Benner. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci*, 27(6):315–321, Jun 2002.

[29] N. Goldman, J. L. Thorne, and D. T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149(1):445–458, May 1998.

[30] N. Goldman and S. Whelan. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol Biol Evol*, 17(6):975–978, Jun 2000, letter.

[31] G. H. Gonnet, M. A. Cohen, and S. A. Benner. Analysis of amino acid substitution during divergent evolution: The 400 by 400 dipeptide substitution matrix. *Biochem Biophys Res Commun*, 199(2):489–496, Mar 1994.

[32] S. Govindarajan, J. E. Ness, S. Kim, E. C. Mundorff, J. Minshull, and C. Gustafsson. Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. *J Mol Biol*, 328(5):1061–1069, May 2003.

[33] J. Gu, Y. Wang, and X. Gu. Evolutionary analysis for functional divergence of Jak protein kinase domains and tissue-specific genes. *J Mol Evol*, 54(6):725–733, Jun 2002.

[34] X. Gu. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol*, 16(12):1664–1674, Dec 1999.

[35] X. Gu. Mathematical modeling for functional divergence after gene duplication. *J Comput Biol*, 8(3):221–234, 2001.

[36] J. P. Huelsenbeck. Testing a covariotide model of DNA substitution. *Mol Biol Evol*, 19(5):698–707, May 2002.

[37] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3):275–282, Jun 1992.

[38] D. T. Jones, W. R. Taylor, and J. M. Thornton. A mutation data matrix for transmembrane proteins. *FEBS Lett*, 339(3):269–275, Feb 1994.

[39] S. Kawashima and M. Kanehisa. AAindex: Amino acid index database. *Nucleic Acids Res*, 28(1):374, Jan 2000.

[40] M. Kimura. *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers*. University of Chicago Press, Chicago, 1994.

[41] H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol*, 30:151–160, 1990.

[42] B. Knudsen and M. M. Miyamoto. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci USA*, 98(25):14512–14517, Dec 2001.

[43] R. Koradi, M. Billeter, and K. Wuthrich. MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graph*, 14(1):51–55, Feb 1996.

[44] J. M. Koshi and R. A. Goldstein. Context-dependent optimal substitution matrices. *Protein Eng*, 8(7):641–645, Jul 1995.

[45] J. M. Koshi and R. A. Goldstein. Models of natural mutations including site heterogeneity. *Proteins*, 32(3):289–295, Aug 1998.

[46] S. M. Larson, A. A. Di Nardo, and A. R. Davidson. Analysis of covariation in an SH3 domain sequence alignment: Applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J Mol Biol*, 303(3):433–446, Oct 2000.

[47] P. Lió and N. Goldman. Using protein structural information in evolutionary inference: Transmembrane proteins. *Mol Biol Evol*, 16(12):1696–1710, Dec 1999.

[48] P. Lió and N. Goldman. Modeling mitochondrial protein evolution using structural information. *J Mol Evol*, 54(4):519–529, Apr 2002.

[49] L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. SCOP: A structural classification of proteins database. *Nucleic Acids Res*, 28(1):257–259, Jan 2000.

[50] P. Lopez, D. Casane, and H. Philippe. Heterotachy, an important process of protein evolution. *Mol Biol Evol*, 19(1):1–7, Jan 2002.

[51] Y. Mandel-Gutfreund, S. M. Zaremba, and L. M. Gregoret. Contributions of residue pairing to beta-sheet formation: Conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J Mol Biol*, 305(5):1145–1159, Feb 2001.

[52] T. Miyata, S. Miyazawa, and T. Yasunaga. Two types of amino acid substitutions in protein evolution. *J Mol Evol*, 12(3):219–236, Mar 1979.

[53] K. Mizuguchi and T. Blundell. Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics*, 16(12):1111–1119, Dec 2000.

[54] T. Muller, R. Spang, and M. Vingron. Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol*, 19(1):8–13, Jan 2002.

[55] S. V. Muse. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics*, 139(3):1429–1439, Mar 1995.

[56] G. J. Naylor and W. M. Brown. Structural biology and phylogenetic estimation. *Nature*, 388(6642):527–528, Aug 1997, letter.

[57] E. Neher. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA*, 91(1):98–102, Jan 1994.

[58] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, Mar 1998.

[59] J. Overington, D. Donnelly, M. S. Johnson, A. Sali, and T. L. Blundell. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci*, 1(2):216–226, Feb 1992.

[60] L. Patthy. *Protein Evolution.* Blackwell Science, London, 1999.

[61] D. D. Pollock and W. R. Taylor. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng*, 10(6):647–657, Jun 1997.

[62] D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: Maximum likelihood identification and relationship to structure. *J Mol Biol*, 287(1):187–198, Mar 1999.

[63] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 (Suppl 1):71–77, Jul 2002.

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

[65] M. Schoeniger and A. von Haeseler. Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *J Mol Evol*, 49(5):691–698, Nov 1999.

[66] O. Schueler and H. Margalit. Conservation of salt bridges in protein families. *J Mol Biol*, 248(1):125–135, Apr 1995.

[67] I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. *Protein Eng*, 7(3):349–358, Mar 1994.

[68] O. Soyer, M. W. Dimmic, R. R. Neubig, and R. A. Goldstein. Using evolutionary methods to study G-protein coupled receptors. In *Pacific Symposium on Biocomputing*, pages 625–636. World Scientific, Singapore, 2002.

[69] O. Soyer, M. W. Dimmic, R. R. Neubig, and R. A. Goldstein. Dimerization in aminergic G-protein coupled receptors: Application of a hidden site-class model of evolution. *Biochemistry*, 42(49):14522–14531, Dec 2003.

[70] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng*, 9(1):27–36, Jan 1996.

[71] P. Tufféry and P. Darlu. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol Biol Evol*, 17(11):1753–1759, Nov 2000.

[72] C. Tuffley and M. Steel. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci*, 147(1):63–91, Jan 1998.

[73] H. Wako and T. L. Blundell. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J Mol Biol*, 238(5):682–692, May 1994.

[74] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, 18(5):691–699, May 2001.

[75] K. R. Wollenberg and W. R. Atchley. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci USA*, 97(7):3288–3291, Mar 2000.

[76] Z. Yang. Estimating the pattern of nucleotide substitution. *J Mol Evol*, 39(1):105–111, Jul 1994.

[77] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol*, 39(3):306–314, Sep 1994.

[78] Z. Yang. Relating physicochemical properties of amino acids to variable nucleotide substitution patterns among sites. In *Pacific Symposium on Biocomputing*, pages 81–92. World Scientific, Singapore, 2000.

[79] Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15(12):1600–1611, Dec 1998.