Amino Acid Substitution Matrices

Dannie Durand

Overview

In prior lectures, we introduced Markov models of nucleotide substitution. We derived expressions for the probability that nucleotide x will change to nucleotide y after elapsed time t. Further, we used the model to account for multiple substitutions, by estimating the number of actual substitutions that occurred, given the number of observed mismatches.

Here, we focus on Markov models of amino acid replacement and their use in deriving amino acid substitution matrices. An amino acid substitution matrix assigns a score to a pair of aligned amino acids, j and k. A good substitution matrix should have the following properties:

- *Biophysical properties of residues:* Amino acids differ in size and charge. Some are acidic, some are basic, some have aromatic side chains. Generally, replacement of an amino acid with another amino acid with similar properties is less likely to break the protein or cause dramatic changes in function than replacement with an amino acid with different properties. A substitution matrix should reflect this.
- *Evolutionary divergence:* The observation of identical or functionally conservative amino acids at the same site is more surprising in highly diverged protein families than in families characterized by little sequence divergence. The best results are obtained using a substitution matrix based on the statistics of amino acid replacements typical of the degree of evolutionary divergence of the proteins under consideration. Therefore, a family of matrices that is parameterized by sequence divergence is desired.
- *Multiple substitutions:* The score associated with an amino acid pair, j and k, should reflect the probability of observing j aligned with k, taking into account the possibility of multiple replacements at the same site.

There are two commonly used families of amino acid substitution matrices that have these properties, the PAM matrices (Dayhoff *et al.*, 1978) and the BLOSUM matrices (Henikoff and Henikoff, 1992.) Both substitution matrix families are parameterized by sequence divergence. The PAM matrices are based on a formal Markov model of sequence evolution. The BLOSUM matrices use an *ad hoc* approach. Both families were derived according to the following general approach, although the details of each step differ between the two methods.

1. Use a set of "trusted" multiple sequence alignments (ungapped) to infer model parameters.
2. Count observed amino acid pairs in the trusted alignments, correcting for sample bias.
3. Estimate substitution frequencies from amino acid pair counts.
4. Construct a log odds scoring matrix from substitution frequencies.

PAM matrices

The PAM matrices were developed by Margaret Dayhoff and her colleagues in 1978. A PAM is a unit of evolutionary distance. The term "PAM" means "percent accepted mutation." We say the divergence between two sequences is n PAMs, if, on average, n amino acid replacements per 100 residues (including multiple substitutions at the same site) occurred since their separation.

The Dayhoff matrices are parameterized by PAM distance. Dayhoff used the following strategy to obtain amino acid substitution matrices that are parameterized by evolutionary distance:

- Construct a Markov chain to model amino acid substitution at a single site i. This chain has twenty states, one for each possible amino acid at that site. If the chain is in state j at time t, we say that we see amino acid j at site i at time t. Note that this model assumes site independence.
- For this Markov chain, we derive the PAM-1 transition probability, $P_{jk}^{(1)}$, from closely related alignments that are assumed to contain no multiple substitutions. $P_{jk}^{(1)}$ is the probability of observing amino acid k at site i at time $t+1$, given that we observed amino acid j at site i at time t; in other words, the probability that amino acid j will be replaced by amino acid k in sequences separated by 1 PAM of evolutionary distance.
- The PAM-n transition probability, $P_{jk}^{(n)}$, is obtained by extrapolating from the PAM-1 transition probability. This is the probability that j will be replaced with k after n time steps. We can also think of $P_{jk}^{(n)}$ as the probability of observing amino acid j aligned with amino acid k in sequences that are n PAM units apart.

Dayhoff's implementation of the general approach given above is as follows:

1. As training data, Dayhoff *et al* used a set of ungapped, global multiple sequence alignments of 71 groups of closely related sequences. Within each group, the sequence identity was 85% or greater. The rationale is that sequences with at least 85% identity will contain no site that has sustained more than one mutation.

2. Observed amino acid pair frequencies were tabulated from the 71 multiple alignments. Sample bias was corrected by counting the minimum number of changes required to fit the data to a tree, according to a parsimony model. The counts were averaged over all most parsimonious trees. For each tree, T, we calculate A_{jk}^T by counting the number of edges connecting j and k, for $j \neq k$. Note that $A_{jk}^T = A_{kj}^T$, since every edge connecting j with k also connects k with j. We define A_{jj}^T to be twice the number of edges connecting j and j. This is because the edges connecting two dissimilar residues are also counted twice, once in the jk direction and once in the kj direction. The overall counts are obtained by averaging over all trees:

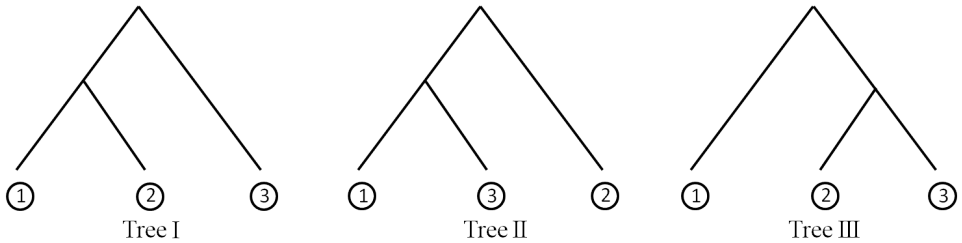
$$A_{jk} = \frac{1}{n_T} \sum_T A_{jk}^T,$$

where n_T is the number of trees with an optimal parsimony score.

To see how this works in practice, suppose we have an alignment of three sequences, each of which has two amino acids:

```
1:  VV
2:  II
3:  VI
```

There are three rooted tree topologies with three leaves

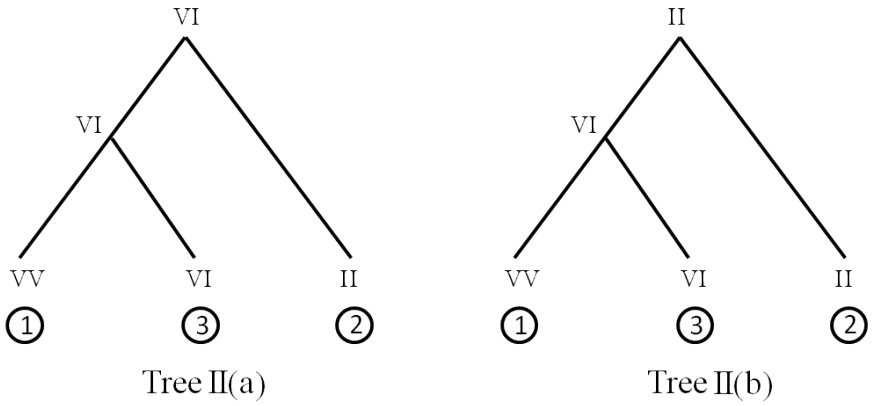


Note that although all three trees appear to have the same shape, the leaf labels differ, corresponding to different evolutionary hypotheses. For example,Tree I corresponds to the hypothesis that

sequence (1) is more closely related to sequence (2) than to sequence (3), while Tree II says that sequences (1) and (3) are most closely related.

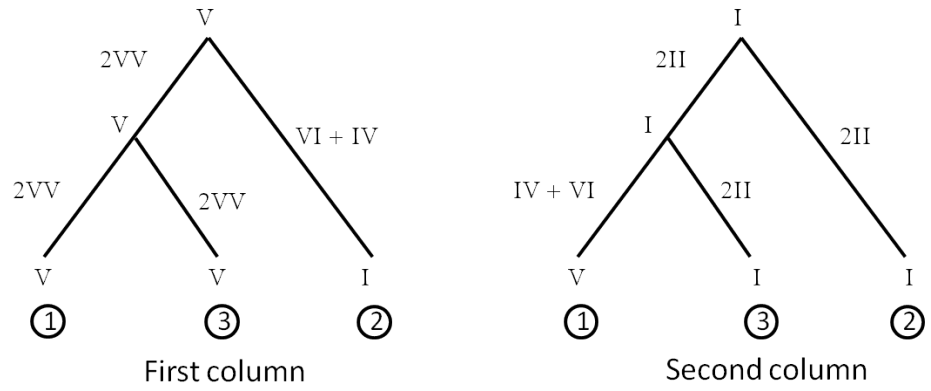
In order to determine A_{jk}^{T}, the count for tree T, each leaf is annotated with the corresponding present-day sequences (*i.e. VV, II*, and *VI*). The sequences on internal nodes are unknown, since they correspond to ancestral sequences. Dayhoff inferred the sequences on the internal nodes according to the *parsimony criterion*, which states that the best hypothesis is the hypothesis that requires the fewest mutations to explain the data. In other words, she assigned sequences to the internal nodes of each tree in such a way that the total number of changes along branches of the tree is minimal. In general, there can be more than one way to assign sequences to internal nodes that minimizes the total change. In our example, there are two possible assignments of ancestral sequences for both Tree II and Tree III. Tree I has a unique set of internal labels.

The two possible internal node labelings of Tree II are shown here:



Since there is no way of knowing which set of inferred ancestral sequences is correct, we must consider all possibilities. When the internal labels are taken into account in this example, there are five most parsimonious trees, one for Tree I and two each for Trees II and III. In the Dayhoff framework, we estimate the pair counts on each of the five trees and take the average.

The figure below illustrates how state changes are counted for Tree II(a) in our example.



We need to determine the number of pairs of each type in each column and then sum over all columns. The number of `VV`, `VI`, `IV`, and `II` pairs are shown on the left for the first column and on the right for the second column of the alignment. Combining these counts, we obtain

$$A_{VV} = 6, \qquad A_{VI} = 2, \qquad A_{IV} = 2, \qquad A_{II} = 6$$

for Tree II(a).

3. The transition matrix $P_{jk}^{(1)}$ is derived from the counts, A_{jk}, obtained in step 2, as follows:

$$P_{jk}^{(1)} = m_j \frac{A_{jk}}{\sum_{h \neq j} A_{jh}}, \quad j \neq k \tag{1}$$

$$P_{jj}^{(1)} = 1 - m_j \tag{2}$$

Here, m_j is the "mutability" of amino acid j and is defined to be

$$m_j = \frac{1}{Lp_j z} \sum_{l \neq j} A_{jl}, \tag{3}$$

where p_j is the background frequency of j and L is the length of the alignment. We select the normalization factor, z, so that

$$\sum_{j=1}^{20}(p_j m_j) = \frac{1}{100}. \tag{4}$$

The training alignments are sufficiently conserved to contain no multiple substitutions, but the frequency of replacements in each alignment may not be exactly one in a hundred. This normalization step guarantees that the transition matrix will correspond to exactly 1 PAM. We obtain an expression for the normalization factor, z, by substituting the right hand side of equation (3) for m_j in equation (4) and solving for z. This yields

$$z = \frac{100}{L} \sum_{j=1}^{20} \sum_{l \neq j} A_{jl}. \tag{5}$$

We now replace the z in equation (3) with the right hand side of equation (5) to obtain the mutability of j:

$$m_j = \frac{0.01}{p_j} \frac{\sum_{l \neq j} A_{jl}}{\sum_h \sum_{l \neq h} A_{hl}}.$$

Substituting the expression for m_j into the right hand side of equation (1), we obtain the PAM1 transition probability

$$P_{jk}^{(1)} = \frac{0.01}{p_j} \frac{A_{jk}}{\sum_h \sum_{l \neq h} A_{hl}}.$$

Note that $P_{jk}^{(1)}$ in equation (1) is consistent with the definition of a Markov chain: the transition matrix rows sum to 1 and it is history independent. This Markov chain is finite, aperiodic and irreducible ("connected"). Therefore, it has a stationary distribution.

We now consider the PAM-2 transition matrix. Note that the residue at site i can change from a j to a k in two time steps via several state paths: $j \to j \to k, j \to k \to k$, or $j \to l \to k$, where l is a third amino acid, not equal to j or k. Recall that the probability of changing from a j to a k in two time steps is

$$P_{jk}^{(2)} = \sum_l P_{jl}^{(1)} P_{lk}^{(1)}$$

$P^{(2)}$ can also be derived by squaring the matrix $P^{(1)}$ by matrix multiplication.

Similarly, we can use matrix multiplication to derive the PAM-n transition matrix for any $n \geq 2$ as follows:

$$P^{(n)} = \left(P^{(1)}\right)^n.$$

4. We obtain a log odds scoring matrix from the transition probability matrix as follows. Let $q_{jk}^{(n)} = p_j P_{jk}^{(n)}$ be the probability that we see amino acid j aligned with amino acid k at a given position in an alignment of sequences with n PAMs of divergence; i.e., that amino acid j has been replaced by amino acid k after n PAMs of mutational change. Then, we define the PAM-n scoring matrix to be

$$\begin{aligned} S^n[j,k] &= \lambda \log \frac{q_{jk}^{(n)}}{p_j p_k} && (6) \\ &= \lambda \log \frac{P_{jk}^{(n)}}{p_k}, && (7) \end{aligned}$$

where λ is a constant chosen to scale the matrix to a convenient range. Typically $\lambda = 10$ and the entries of S^n are rounded to the nearest integer. Note that equation (7) is a log odds ratio, where $q_{jk}^{(n)}$ is the probability of seeing j and k aligned under the alternate hypothesis that j and k share common ancestry and $p_j p_k$ is the probability that j and k are aligned by chance.

It is easy to verify that the PAM-n transition matrix is not symmetric; that is, $P_{jk}^{(n)} \neq P_{kj}^{(n)}$. This makes sense since replacing amino acid j with amino acid k may have different consequences than replacing k with j.

In contrast, the substitution matrix *is* symmetric; that is, $S^n[j,k] = S^n[k,j]$. This makes sense because in an alignment, we cannot determine direction of evolution, so we assign the same score to j aligned with k and to k aligned with j.