Amino Acid Substitution Matrices
Dannie Durand

Overview

In the last lecture, we introduced a Markov model of substitution in nucleotide sequences and used that model to estimate the number of substitutions, taking multiple substitutions into account. In this lecture, we focused 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 a dramatic change 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) 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, 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 \( k \) in \( n \) 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.

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 \( A_{jk}^T = A_{kj}^T \). 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.

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

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

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

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

$$m_j = \frac{1}{np_j z} \frac{\sum_{l \neq j} A_{jl}}{\sum_h \sum_{l \neq h} A_{hl}},$$  \hspace{1cm} (1)

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

$$\sum_{j=1}^{20} (p_j m_j) = \frac{1}{100}$$  \hspace{1cm} (2)

in order to guarantee that we obtain a transition matrix corresponding to exactly 1 PAM. We obtain an expression for the normalization factor, $z$, by substituting the right hand side of equation (1) for $m_j$ in equation (2) and solving for $z$. This yields

$$z = \frac{100}{n} \sum_{j=1}^{20} \sum_{l \neq j} A_{jl}.$$  \hspace{1cm} (3)

We now replace the $z$ in equation (1) with the right hand side of equation (3) to obtain the mutability of $j$:

$$m_j = 0.01 \frac{1}{p_j} \frac{\sum_{l \neq j} A_{jl}}{\sum_h \sum_{l \neq h} A_{hl}}.$$  

Note that $P_{jk}^{(1)}$ is consistent with the definition of a Markov chain. The rows sum to 1 and it is history independent. This Markov chain is finite, aperiodic and irreducible. 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: $j \rightarrow j \rightarrow k, j \rightarrow k \rightarrow j, j \rightarrow l \rightarrow k$, where $l$ is a third amino acid, not equal to $j$ or $k$. 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:

$$P^{(n)} = (P^{(1)})^n.$$ 

4. We obtain a log odds scoring matrix from the transition probability matrix as follows. Let $q^{(n)}_{jk} = p_j P^{(n)}_{jk}$ 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

$$S^n[j, k] = \lambda \log \frac{q^{(n)}_{jk}}{p_j p_k}$$

(4)

$$= \lambda \log \frac{P^{(n)}_{jk}}{p_k}$$

(5)

where $\lambda$ is a constant. Note that equation (5) is a log odds ratio, where $q^{(n)}_{jk}$ 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. Typically $\lambda = 10$ and the entries of $S^n$ are rounded to the nearest integer.

It is easy to verify that the PAM-n transition matrix is not symmetric; that is, $P^{(n)}_{jk} \neq P^{(n)}_{kj}$. 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 is 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 $j$. 