Amino Acid Substitution Matrices

Tuesday, October 30th

Dannie Durand

Overview

Goal: Amino acid similarity matrices that take into account

- biophysical properties of residues,
- evolutionary divergence and
- multiple substitutions.

There are two commonly used families of amino acid substitution matrices that have these properties:

- PAM - Dayhoff *et al.*, 1978
- BLOSUM - Henikoff S., Henikoff JG., 1992.

Each family is parameterized by evolutionary distance. Both use the following general approach

1. Find a set of "trusted" MSA's (ungapped)
2. Count amino acid changes, correcting for sample bias in choice of sequences
3. Estimate substitution frequencies
4. Construct log odds scoring matrix

PAM matrices

PAM is a unit of evolutionary distance. The term "PAM" means "accepted point mutation" or "percent accepted mutation." We say two sequences are n PAMs apart if every 100 residues contain, on average, n actual changes (including multiple substitutions) between them. Our goal is to construct a family of matrices that are parameterized by PAM distance.

Dayhoff's implementation of the general approach given above is as follows:
1. The training data of a set of ungapped, global multiple sequence alignments of 71 groups of closely related sequences. Within each group, the sequence similarity is $\geq$85%.

2. Count replacements in the alignments, correcting for sample bias in choice of sequences by averaging 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. The overall counts are obtained by averaging over all trees:

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

3: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, $P^1[j,k]$, from closely related alignments (no multiple substitutions). $P^1[j,k]$ 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.
- We then extrapolate to obtain the PAM-n transition probability, $P^n[j,k]$. This is the probability that j will be replaced with k in n time steps. We can also think of $P^1[j,k]$ as the probability of observing amino acid j and amino acid k aligned in sequences that are n PAM units apart.

Specifically, from the counts, A_{jk}, obtained in step 2, the transition matrix $P^1[j,k]$ is derived as follows:

$$\begin{aligned} p^1[j,k] &= m_j \frac{A_{jk}}{\sum_{i \neq j} A_{ji}}, \quad j \neq k \\ p^1[j,j] &= 1 - m_j \end{aligned}$$

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

$$m_j = \frac{1}{np_j z} \quad \frac{\sum_{l \neq j} A_{jl}}{\sum_h \sum_{l \neq h} A_{hl}}, \tag{1}$$

where p_j is the background frequency of j and n is the length of the MSA. We select the nomalization factor, z, so that

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

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 (1) for m_j in equation 2 and solving for z. This yields

$$z = \frac{100}{n} \sum_{j=1}^{20} \sum_{l \neq j} A_{jl} \tag{3}$$

We now replace z with the right hand side of (3) in equation 1 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 $P^1[j,k]$ is consistent with the definiction 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: $j \to j \to k, j \to k \to k$, or $j \to l \to k$, where l is a third amino acid, not equal to j or k. The probability of changing from a j to a k in two time steps is

$$P^2[j,k] = \sum_k \sum_l P^1[j,l] P^1[l,k].$$

$P^2[j,k]$ can also be derived by squaring the matrix $P^1[k,l]$ by matrix multiplication.

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

$$P^n[j,k] = (P^1[j,k])^n$$

3. We obtain a log odds scoring matrix from the transition probability matrix as follows. Let $q^n[j,k] = P_j P^n[j,k]$ 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 is replaced by amino acid k after n PAMs of mutational change. Then, we define the PAM n scoring matrix to be

$$S^n[j,k] = \lambda \log \frac{q^n[j,k]}{p_j p_k} \tag{4}$$

$$= \lambda \log \frac{P^n[j,k]}{p_k} \tag{5}$$

where λ is a constant. Note that Equation 5 is a log odds ratio, where $q^n[j,k]$ is the probability of seeing j and k aligned under the alternate hypotheses 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, $P^n[j,k] \neq P^n[k,j]$. 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.