Chapter 3

Amino Acid Substitution Matrices

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, x and y. A good substitution matrix should have the following properties:

- *Evolutionary divergence:* The subsitution matrix should be appropriate for the degree of evolutionary divergence of the proteins under consideration. The observation of identical or functionally similar 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 amino acid replacement frequencies that are typical of the protein family. Therefore, a set of matrices that is parameterized by sequence divergence is desired.

- *Multiple substitutions:* The score associated with an amino acid pair, x and y, should reflect the probability of observing x aligned with y, taking into account the possibility of multiple replacements at the same site.

- *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.

There are several families of amino acid substitution matrices that have these properties. Two that are widely used are the PAM matrices (Dayhoff *et al.*, 1978) and the BLOSUM

matrices (Henikoff and Henikoff, 1992.) Both of these families of substitution matrices are parameterized by sequence divergence. The PAM matrices account for evolutionary divergence using a formal Markov model of sequence evolution. The BLOSUM matrices use an *ad hoc* approach. Although the details differ, both matrix families were derived according to the following general strategy:

1. Use a "trusted" set of ungapped, multiple sequence alignments 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 likelihood scoring matrix from substitution frequencies.

3.1 A log likelihood ratio framework for scoring alignments

Before introducing the PAM and BLOSUM matrices, we briefly introduce the log likelihood framework in which these matrices were developed. Suppose $\alpha^{\kappa}(s_1, s_2)$ is an ungapped alignment of sequences s_1 and s_2 of length n. Under the assumption of positional independence, we can assign a similarity score to $\alpha^{\kappa}(s_1, s_2)$ by adding the similarities of the symbols in each position in the alignment,

$$\mathcal{S} = \sum_{i=1}^{n} p(s_1[i], s_2[i]), \tag{3.1}$$

where $p(x, y)$ is a quantitative measure of the similarity of x and y. Recall that earlier in the semester, we used a simple scoring scheme with a single match score, $p(x, x) = M$, $\forall x \in \Sigma$, and a single mismatch score, $p(x, y) = m$, $\forall x, y \in \Sigma$ such that $x \neq y$. Since all matches (respectively, mismatches) have the same score, with this scoring scheme

$$\mathcal{S} = \hat{m} \cdot m + (n - \hat{m}) \cdot M,$$

where $\hat{m}$ is the number of mismatches in α^{κ}.

This simple scoring scheme has limitations, especially for amino acids. First, since all mismatches are assigned the same score, it cannot reflect differences in the biochemical similarity of various amino acid pairs. Second, if M and m are chosen arbitrarily, then alignment scores have no intuitive meaning in an absolute sense. For example, if I tell you that a given alignment has a score of 14, you know that it is better than some other alignment of the same sequences that has a score of 12, but you have no way of assessing whether the alignment is inherently good or bad.

Third, this scoring scheme does not take the evolutionary divergence of s_1 and s_2 into account. If we are testing the hypothesis that s_1 and s_2 are related and have changed very little since they diverged from their common ancestor, then we might interpret any

mismatch as evidence that s_1 and s_2 are unrelated, even if the mismatch is a conservative replacement (i.e., involves amino acids with similar biochemical properties). In contrast, if we are testing the hypothesis that s_1 and s_2 are related and have changed a great deal since their divergence, then we might interpret mismatches that represent conservative replacements as evidence s_1 and s_2 are indeed related. In order to capture these nuances, we require a scoring method that is parametized by evolutionary divergence.

One way of assessing whether an alignment is good in an absolute sense is to ask whether $\alpha^\kappa(s_1, s_2)$ reflects more similarity than we expect to see by chance. Let H_0 be the null hypothesis that s_1 and s_2 are unrelated sequences. The alternate hypothesis, H_A, is that s_1 and s_2 are related sequences with a given amount of evolutionary divergence. We can assess whether $\alpha^\kappa(s_1, s_2)$ reflects more than chance similarity by calculating the ratio of the probabilities of the alignment under H_A and H_0:

$$\mathcal{LR}(\alpha^\kappa) = \frac{p(\alpha^\kappa|H_A)}{p(\alpha^\kappa|H_0)}. \tag{3.2}$$

This *likelihood ratio* will be less than one, if the alignment of s_1 and s_2 represents less similarity than expected by chance, and greater than 1, if the alignment represesnts is more similarity than expected by chance. If the ratio is much greater than 1, then we have strong evidence that the sequences share common ancestry.

Under the assumption of positional independence, the probability of the alignment is equivalent to the product of probabilities of the individual positions in the alignment

$$\mathcal{LR}(\alpha^\kappa) = \prod_{i=1}^{n} \frac{p(\alpha^\kappa[i]|H_A)}{p(\alpha^\kappa[i]|H_0)}, \tag{3.3}$$

where $\alpha^\kappa[i]$ is the alignment of $s_1[i]$ and $s_2[i]$. This formulation provides a way to assess alignments based on the probabilities of individual amino acid pairs in the alignment. (Recall that α^κ is an ungapped alignment.) However, it requires calculating the product of a sequence of numbers between 0 and 1, with the concomitant challenge of working with smaller and smaller numbers as the length of the alignment increases.

This problem can be addressed by calculating the *log* of the likelihood ratio, instead of the likelihood ratio, itself. Note that since $\log(x)$ increases monotonically with x, the alignment that maximizes $\mathcal{LR}(\alpha^\kappa)$, also maximizes $\log \mathcal{LR}(\alpha^\kappa)$. Thus, $\log \mathcal{LR}(\alpha^\kappa)$ can also be used to assess the extent to which $\alpha^\kappa(s_1, s_2)$ represents more than chance similarity. Taking the log of both sides of Equation 3.3 yields

$$\log \mathcal{LR}(\alpha^\kappa) = \log \prod_{i=1}^{n} \frac{p(\alpha^\kappa[i]|H_A)}{p(\alpha^\kappa[i]|H_0)} \tag{3.4}$$

$$= \sum_{i=1}^{N} \log \frac{p(\alpha^\kappa[i]|H_A)}{p(\alpha^\kappa[i]|H_0)}. \tag{3.5}$$

Since $\mathcal{LR}(\alpha^\kappa)$ is non-negative, $\log \mathcal{LR}(\alpha^\kappa)$ ranges from $-\infty$ to ∞. If $\log \mathcal{LR}(\alpha^\kappa) > 0$, then $\alpha^\kappa(s_1, s_2)$ reflects more than chance similarity than expected by chance; if $\log \mathcal{LR}(\alpha^\kappa) > 0$, then $\alpha^\kappa(s_1, s_2)$, then $\alpha^\kappa(s_1, s_2)$ reflects less similarity than expected.

The right hand side of this equation looks very similar to the right hand side of Equation 3.1: in both cases, we have a sum of values, one for each position in the alignment. In Equation 3.1, the i^{th} entry in the sum is measure of similarity of $s_1[i]$ and $s_2[i]$; in Equation 3.5, the i^{th} entry is the probability, relative to chance, of observing $s_1[i]$ aligned with $s_2[i]$. This suggests that we can use the log likelihood ratios to define a scoring scheme. By defining the similarity score of x aligned with y to be

$$p(x, y) = \log \frac{p(\binom{x}{y}|H_A)}{p(\binom{x}{y}|H_0)},$$

we obtain an alignment score that is equivalent to the log of the ratio of the probabilities of that alignment under the alternate and null hypotheses:

$$\mathcal{S} = \log \mathcal{LR}(\alpha^\kappa).$$

This yields a scoring scheme that has a natural, biological interpretation, that can be adjusted to account for evolutionary diverence, and that can be interpreted in an absolute, as well as a relative, context.

To define similarity scores in this way, requires estimates of $p(\binom{x}{y}|H_A)$ and $p(\binom{x}{y}|H_0)$, for a range of evolutionary distances. For amino acid substitution matrices, these quantities are estimated from trusted amino acid alignments. In the following sections, we discuss amino acid pair probabilities are estimated in derivation of the PAM matrices and the BLOSUM matrices.

3.2 PAM matrices

In 1978, Margaret Dayhoff and her colleagues developed a family of substitution matrices that are parameterized by PAM distance, a unit of evolutionary divergence. The term "PAM" is an abbreviation of "percent accepted mutation." The divergence between two sequences is N PAMs, if, on average, N amino acid replacements (possibly at the same site) per 100 residues occurred since their separation. Note that this is distinct from percent identit, which reflects the number of matches per 100 residues.

The derivation of these matrices requires estimating amino acid pair frequencies in sequences that are diverged by N PAMs, for a range of values of N. Given alignments of sequences that are separated by N PAMs, amino acid pair frequencies can be estimated simply by tabulating the number of instances of each amino acid pair in those alignments. However, it is not clear how to obtain such alignments, because determining the PAM distance associated with a given alignment is not straightforward. The number of *mismatches*

can easily be determined by inspection, but inferring the number of *replacements* that occurred requires a method for estimating multiple replacements at the same site. To address this problem, Dayhoff first constructed a model of amino acid replacement using alignments with high levels of sequence similarity, in which multiple substitutions at the same site are unlikely. She then used higher-order Markov models to obtain models of amino acid replacements in more diverged sequences.

Dayhoff developed this model using the four step approach described above. Specifically:

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. This requires inferring the unrooted tree that describes the evolutionary relationships between the sequences in each aligned family and then estimating the number of amino acid replacements that occured on each branch of that tree.

We will demonstrate how this works in practice using the following alignment of four amino acid sequnces of length four:

```
1:  AEIR
2:  DEIR
3:  QKLH
4:  AHLH
```

For an alignment with four sequences, there are three unrooted trees with four leaves, shown in Fig. 3.1. Tree I corresponds to the hypothesis that Sequence (1) is more closely related to Sequence (2) than to either Sequence (3) or Sequence (4). According to Tree II, Sequence (1) and Sequence (3) are most closely related, while Tree III says that Sequence (1) and Sequence (4) are closest. For each tree, the leaves are annotated with the corresponding present-day sequences. The sequences on internal nodes are unknown, since they correspond to ancestral sequences.

First, we will illustrate how to estimate the number of substitutions, given the evolutionary tree. Then, we will return to the question of how to infer the tree that best explains a given alignment. Dayhoff inferred the sequences on the internal nodes according to the *parsimony criterion*, which states that the best hypothesis is the one that requires the fewest amino acid replacements to explain the data. Consistent with this criterion, sequences were assigned to the internal nodes of each tree in such a way that the total number of changes along branches of the tree is minimized.

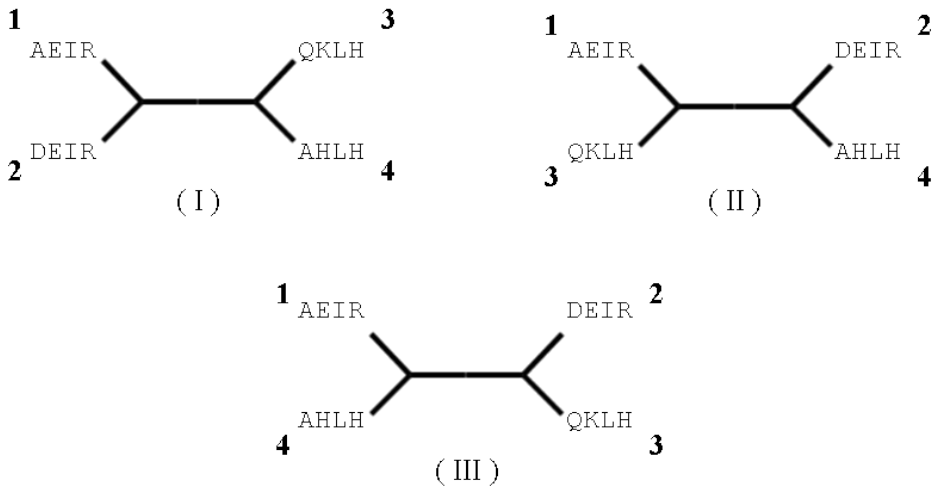

Figure 3.1: Three unrooted trees representing the three possible hypotheses for evolution of four sequences. The leaves of each tree is labeled with the corresponding present-day sequences. The internal nodes are not labeled. The sequences associated with internal nodes correspond to ancestral sequences and are unknown.

For example, suppose that we have determined that Tree I is the best hypothesis for the evolutionary history of the four sequences in the alignment. Ancestral sequences that satisfy the parsimony criterion for Tree I are shown Fig. 3.2. With these ancestral sequences, six substitutions (shown on their respective branches) are required to explain the evolution of the four present day sequences. Convince yourself that there is no assignment of labels to the internal nodes that allows for fewer than six substitutions.

Once ancestral sequences have been inferred, the counts for each amino acid pair are tabulated. A_{xy}, the number of x,y pairs observed, is determined by by counting the number of edges connecting x and y, for $x \neq y$. Note that $A_{xy} = A_{yx}$, since every edge connecting x with y also connects y with x. A_{xx} is defined to be twice the number of edges connecting x and x. This is because the edges connecting two dissimilar residues are also counted twice, once in the xy direction and once in the yx direction. For example, there are 6 `EE` pairs in Fig. 3.2: Two counts are contributed by each of the three edges connecting `AEIR` and `AEIR`, `AEIR` and `DEIR`, and `AEIR` and `AELH`. The tabulated counts for all amino acid pairs are given in the table in Fig. 3.3.

In general, there can be more than one way to assign sequences to internal nodes such that the total change is minimized. Each most parsimonious set of internal node labels will result in different amino acid pair counts. In our example, there are two additional

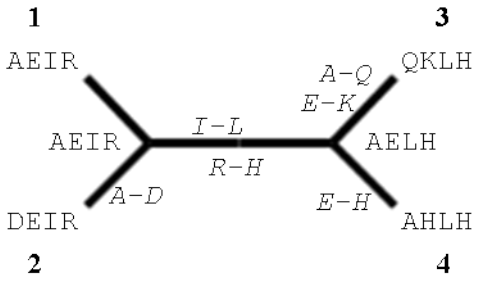

Figure 3.2: Tree I from Fig. 3.1 with ancestral sequences inferred according to the parsimony criterion. The associated amino acid replacements are shown on the branches of the tree. Six replacements are required to explain the present day sequences. This set of most parsimonous ancestral sequences is not unique. There are two other most parsimonious hypotheses for the ancestral sequences, shown in Fig. 3.4.

assignments of ancestral sequences for which six substitutions are sufficient to explain the present-day sequences, shown in Fig. 3.4. The pair counts resulting from these two alternate sets of labels are given in the tables in Fig. 3.5. Since there is no way of knowing which set of inferred ancestral sequences is the best estimate, all possibilities must be considered. Dayhoff does this by averaging the counts over all most parsimonious labelings. For our example, Fig. 3.6 shows the average of the pair counts in Figs. 3.3 and 3.5.

Comparison of the original multiple alignment with the pair counts derived from the trees in Figs. 3.2 and 3.4 demonstrates how this approach compensates for sample bias and leads to different amino acid pair statistics. If we derived amino acid pairs directly from the

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			6	1		1			
H			1	4					1
I					4		1		
K			1						
L					1		4		
Q	1								
R				1					4

Figure 3.3: Amino acid pair counts derived according to Dayhoff's counting scheme from the tree in Fig. 3.2. Only amino acids that are present in at least one sequence are shown in the table.

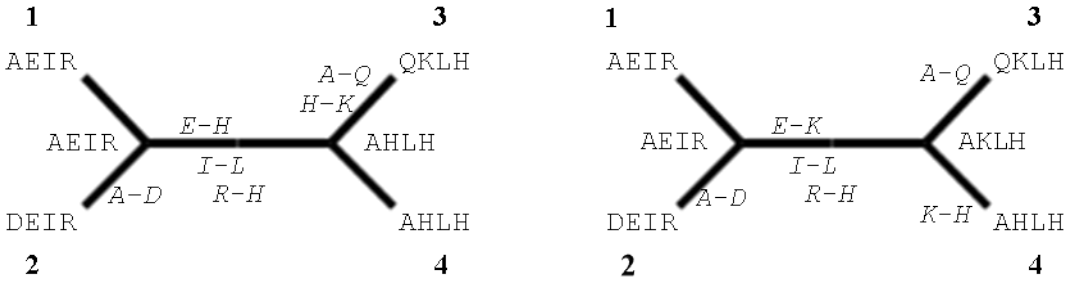

Figure 3.4: Two other sets of most parsimmonious ancestral sequences for Tree I from Fig. 3.1. The associated amino acid replacements are shown on the branches of the tree.

alignment, each sequence would be compared to three other sequences, effectively counting the replacement of the same amino acid more than once. In contrast, when counting amino acid pairs on a tree, each sequence is compared to one other sequence, i.e., the inferred ancestral sequence. For example, since `D` and `Q` both appear in the first column of the alignment, obtaining amino acid pair counts directly from the alignment would result in a non-zero value of A_{DQ}. However, no `D-Q` replacement appears on the branches of the labeled trees in Figs. 3.2 and 3.4 and $A_{DQ} = 0$ in the table in Fig. 3.5.

Having demonstrated how to infer ancestral sequences for a given evolutionary tree, we return to the question of how to infer the tree that is the best hypothesis for the aligned sequences. Dayhoff also used the parsimony principle to select the tree. For a given tree, the minimum number of changes required to explain the present day sequences, over all possible internal labelings, is called the *parsimony score* of that tree. Tree I has a parsimony score of 6, for example. Given an alignment of a family of k sequences, all unrooted trees

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			4	1					
H			1	6		1			1
I					4		1		
K				1					
L					1		4		
Q	1								
R				1					4

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			4			1			
H				4		1			1
I					4		1		
K			1	1		2			
L					1		4		
Q	1								
R				1					4

Figure 3.5: Amino acid pair counts derived according to Dayhoff's counting scheme from the trees in Fig. 3.4.

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			4.7	1		0.7			
H			1	4.7		0.7			1
I					4		1		
K			0.7	0.7		0.7			
L					1		4		
Q	1								
R				1					4

Figure 3.6: Average amino acid pair counts. Each entry represents the mean of the corresponding entries in the tables in Figs. 3.3 and 3.5.

with k sequences were considered and the parsimony score was estimated for each tree. In general, there can be more than one most parsimonious tree for a given set of pressent-day sequences, although for our example there is only one. (Convince yourself that for Trees II and III, it is not possible to assign sequences to the internal nodes that require six or fewer replacements.) Having found the set of most parsimonious trees, Dayhoff estimated amino acid pair frequencies by averaging the counts over all most parsimonious labelings of all most parsimonious trees, yielding

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

where n_T is the number of labeled trees with an optimal parsimony score and T is an indicator variable that enumerates such trees.

3. To estimate substitution frequencies from amino acid pair counts, Dayhoff constructed a family of Markov models representing evolution at a single site, i, in an amino acid sequence (Note that this model assumes site independence.) All models in the family have twenty states, one for each amino acid. If the model visits state x at time t, we say that the amino acid at site i was an x at time t. The models differ in their transmission probability matrices, which reflect the propensity for amino acid replacement at various evolutionary divergences.

Dayhoff derived $P_{xy}^{(1)}$, transition matrix for the 1 PAM model, from closely related alignments that may be assumed to contain no multiple substitutions. $P_{xy}^{(1)}$ is the probability that amino acid x will be replaced by amino acid y in sequences separated by 1 PAM

of evolutionary distance. Next, Dayhoff derived the PAM-N transition matrix, $P_{xy}^{(N)}$, by extrapolating from the PAM-1 transition probability, as described in detail below: .

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

$$P_{xy}^{(1)} = m_x \frac{A_{xy}}{\sum_{h \neq x} A_{xh}}, \quad x \neq y \tag{3.6}$$

$$P_{xx}^{(1)} = 1 - m_x \tag{3.7}$$

Here, m_x is the "mutability" of amino acid x and is defined to be

$$m_x = \frac{1}{L p_x z} \sum_{l \neq x} A_{xl}, \tag{3.8}$$

where p_x is the background frequency of x, L is the length of the alignment, and z is a scaling that guarantees that the transition matrix will correspond to exactly 1 PAM. We select the scaling factor, z, so that

$$\sum_{x=1}^{20} (p_x m_x) = \frac{1}{100}. \tag{3.9}$$

This scaling factor is required because although 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.

We obtain an expression for the scaling factor, z, by substituting the right hand side of Equation 3.8 for m_x in equation (3.9) and solving for z. This yields

$$z = \frac{100}{L} \sum_{x=1}^{20} \sum_{l \neq x} A_{xl}. \tag{3.10}$$

We now replace the z in Equation 3.8 with the right hand side of Equation 3.10 to obtain the mutability of x,

$$m_x = \frac{0.01}{p_x} \frac{\sum_{l \neq x} A_{xl}}{\sum_h \sum_{l \neq h} A_{hl}}.$$

Substituting the expression for m_x into the right hand side of Equation 3.6, we obtain the PAM1 transition probability

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

Note that $P_{xy}^{(1)}$ in Equation 3.6 is consistent with the definition of a Markov chain: the rows of the transition matrix 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 derive the PAM-2 transition matrix. Note that the residue at site i can change from x to y in two time steps via several state paths: $x \to x \to y$, $x \to y \to y$, or $x \to l \to y$, where l is a third amino acid, not equal to x or y. Recall that the probability of changing from x to y in two time steps is

$$P_{xy}^{(2)} = \sum_l P_{xl}^{(1)} P_{ly}^{(1)}$$

$P^{(2)}$ can be derived by squaring the matrix $P^{(1)}$ by matrix multiplication. This is the transition probability of a second order Markov chain that models amino acid replacements that occur in two time steps. 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 likelihood scoring matrix from the transition probability matrix as follows. Let $q_{xy}^{(N)} = p_x P_{xy}^{(N)}$ be the probability that we see amino acid x aligned with amino acid y at a given position in an alignment of sequences with N PAMs of divergence; i.e., that amino acid x has been replaced by amino acid y after N PAMs of mutational change. Then, we define the PAM-N scoring matrix to be

$$S^N[x,y] = \lambda \log \frac{q_{xy}^{(N)}}{p_x p_y} \tag{3.11}$$

$$= \lambda \log \frac{P_{xy}^{(N)}}{p_y}, \tag{3.12}$$

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 3.12 is a log likelihood ratio, where $q_{xy}^{(N)}$ is the probability of seeing x and y aligned under the alternate hypothesis that x and y share common ancestry with divergence N and $p_x p_y$ is the probability that x and y are aligned by chance.

It is easy to verify that the PAM-N transition matrix is not symmetric; that is, $P_{xy}^{(N)} \neq P_{yx}^{(N)}$. This makes sense since replacing amino acid x with amino acid y may have different consequences than replacing y with x. In contrast, the substitution matrix *is* symmetric; that is, $S^N[x,y] = S^N[y,x]$. This makes sense because in an alignment, we cannot determine direction of evolution, so we assign the same score when x is aligned with y, and when y is aligned with x.

3.3 BLOSUM Matrices

The BLOSUM (BLOck SUbstitution Matrices) matrices were derived by Steven and Jorja Henikoff in 1992 [1]. They were based on a much larger data set than the PAM matrices, and used conserved local alignments or "blocks," rather than global alignments of very closely related sequences. The *"trusted" alignments* used to construct the BLOSUM matrices consisted of roughly 2000 blocks of conserved regions representing 500+ groups of proteins.

Here, we discuss the procedure for constructing a substitution matrix in the BLOSUM framework from a single aligned block. In reality, the BLOSUM matrices were constructed from many blocks. See Ewens and Grant, Section 6.5.2, for a detailed treatment of the BLOSUM matrices, including a discussion of how pair frequencies from multiple blocks are combined. Their treatment includes a worked example with more than one block. Note that their notation is somewhat different from the notation we use in class.

BLOSUM matrix construction uses clustering rather than an explicit evolutionary model, to account for different degrees of sequence divergence. Clustering with different values of N, ranging from 45% to 90%, produces a parameterized set of matrices representing different degrees of sequence divergence. In order to construct a BLOSUM-N matrix, the sequences in each block are first grouped into clusters, such that the percent identity of any pair of sequences from different clusters is less than N. Next, for every pair of clusters, amino acids pairs consisting of one amino acid from each cluster are tabulated. Pairs of amino acids within the same cluster are ignored. Amino acid pair counts are normalized by cluster size so that all clusters contribute equally to the pair statistics.

The clustering step in BLOSUM matrix construction has two purposes: parameterizing evolutionary divergence and accounting for sample bias. First, since only amino acids pairs sampled from two different clusters are tabulated, the data used to construct the matrix consists of amino acid pairs observed in sequences with a particular divergence (i.e., sequences that are less than N identical). Second, to control for sample bias, the contribution of each residue in a cluster is normalized by the number of sequences in that cluster. As a result, each cluster contributes the same amount of information to the estimation of amino acid pair frequencies, even though clusters may contain different numbers of sequences.

The specific procedure for BLOSUM matrix construction is as follows:

Partitioning sequences into clusters with N% identity: The clustering step takes as input a block of k sequences of length L (no gaps) and generates C non-overlapping clusters. The ith cluster, C_i, has k_i sequences of length L, where $k = \sum k_i$. The sequences in the block are partitioned in such a way that every sequence in a cluster is at least N% identical to at least one other sequence in the cluster.

[1] *Amino acid substitution matrices from protein blocks*, PNAS, 1992 Nov 15;89(22):10915-9

One way to obtain such a clustering is to represent the block as a weighted graph, where the nodes correspond to sequences. The nodes for each pair of sequences are connected by an edge that is weighted by their percent identity. To obtain clusters with an $N\%$ identity threshold, all edges with weights lower than $N\%$ are removed, resulting in one or more connected components. Each connected component corresponds to a cluster. If N is greater than the greatest edge weight, then each cluster will contain a single sequence. If N is smaller than the lowest edge weight, then all sequences will be in a single cluster. If this happens, it is not possible to construct a BLOSUM matrix for this value of N.

Amino acid pair counts: Following the clustering step, the *observed* frequency of amino acid x aligned with amino acid y is calculated as follows. For each pair of clusters, C_i and C_j, we determine the number of x, y and y, x pairs, where x and y are in the same column, but in different clusters. Let $N_l(C_i, x)$ be the number of times that residue x appears in the l^{th} column of cluster C_i. Then, the total number of pairs in column l involving an x in one cluster and a y in the other cluster is

$$N_l(C_i, x) \cdot N_l(C_j, y) + N_l(C_i, y) \cdot N_l(C_j, x).$$

However, each of the clusters contributes only one count per column, so we must down weight the number of pairs by the product of the size of the clusters. Suppose clusters C_i and C_j contain k_i and k_j sequences, respectively. Then, the contribution of column l in clusters C_i and C_j to the pair count for x and y is

$$\frac{N_l(C_i, x) \cdot N_l(C_j, y) + N_l(C_i, y) \cdot N_l(C_j, x)}{k_i \cdot k_j}.$$

To obtain the total x, y pair count from this block, we sum over all pairs of clusters and over all columns, yielding

$$A_{xy}^N = \sum_{i=1}^{C} \sum_{j=i+1}^{C} \sum_{l=1}^{L} \frac{N_l(C_i, x) \cdot N_l(C_j, y) + N_l(C_i, y) \cdot N_l(C_j, x)}{k_i \cdot k_j}, \tag{3.13}$$

where $x \neq y$. We use the superscript N to indicate that these are pair counts for a BLOSUM-N matrix, where N is the threshold used in the clustering. When $x = y$, the pairs are only counted in one direction:

$$A_{xx}^N = \sum_{i=1}^{C} \sum_{j>i} \sum_{l=1}^{L} \frac{N_l(C_i, x) \cdot N_l(C_j, x)}{k_i \cdot k_j} \tag{3.14}$$

Estimating substitution frequencies: The frequencies of amino acid pairs are derived from the pair counts by normalizing by the total number of possible pairs; that is, by the

product of the number of sites in the block and the number of pairs of clusters:

$$q_{xy}^N = \frac{A_{xy}^N}{L \cdot \binom{C}{2}}.$$

Estimating the expected pair frequencies: The expected frequency of x aligned with y is the product of the *background* probabilities of observing x and y independently. In PAM matrix construction, the background frequency of an amino acid is assumed to be the frequency of that amino acid in typical proteins, for example, as tabulated by Robinson and Robinson[2]. In contrast, in BLOSUM matrix construction, the expected frequencies are estimated from the BLOCK data and adjusted for the current value of N.

In order to get the *expected* frequency of x aligned with y, we first estimate the frequencies of the individual residues in the current block, again using the clusters to correct for sample bias. As above, the counts from each cluster are "discounted" by a factor of $1/k_i$, and then normalized by the total number of elements, $L \cdot C$, to obtain the amino acid background frequency:

$$p_x = \frac{1}{L \cdot C} \sum_{i=1}^{C} \sum_{l=1}^{L} \frac{N_l(C_i, x)}{k_i}.$$

The expected pair frequencies are then obtained from the products of the background frequencies:

$$\begin{aligned} E_{xy} &= p_x p_y + p_y p_x \\ E_{xx} &= p_x^2. \end{aligned}$$

Finally, the BLOSUM-N *log likelihood scoring matrix* is calculated from the ratios of the observed and expected frequencies:

$$S^N[x, y] = 2 \log_2 \frac{q_{xy}^N}{E_{xy}}.$$

3.4 Comparing PAM and BLOSUM Matrices

We began this endeavor with the goal of deriving substitution matrices that are parameterized by evolutionary divergence. In other words, a given alignment should be scored with a matrix

[2] *Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins*, PNAS, 1991 Oct;88:8880-4

	PAM	BLOSUM
Evolutionary model	Explicit evolutionary model	None
Data	Full length MSAs	Conserved blocks
Bias correction	Trees	Clustering
Multiple substitutions	Markov model: $P^N = (P^1)^N$	Implicitly represented in data
Evolutionary distance	Markov model: $P^N = (P^1)^N$	Clustering
Matrices	Transition & log likelihood scoring matrices	Log likelihood scoring matrix only
Parameter N	Distance increases with N	Distance decreases with N
Biophysical properties	Derived indirectly from data	Derived indirectly from data

Table 3.1: Properties of the PAM and BLOSUM matrices.

with scores that are appropriate for the evolutionary divergence of the sequences being compared. In addition, these scores should implicitly account for multiple substitutions per site, consistent with the typical evolutionary divergence associated with each matrix in the family. A further goal is that the matrices should reflect the biophysical properties of amino acids. The scores for amino acid pairs with similar biophysical properties (i.e., conservative replacements) should be greater than scores for amino acid pairs with divergent biophysical properties (i.e., non-conservative or radical replacements).

The PAM and BLOSUM matrices were both constructed in an explicit log-likelihood framework, with entries of the form

$$S^N[x, y] = c\ log_2 \frac{q^N_{xy}}{p_x p_y},$$

where the numerator, q^N_{xy}, is the frequency of the amino acid pair (x, y) in alignments of related sequences with divergence N and the denominator, $p_x p_y$, is the frequency with which the pair (x, y) will occur if amino acids are sampled according to their background frequencies. The constant c is a scaling factor chosen for convenience. Multiplying every entry in the matrix by a constant changes the value of the entries in an absolute sense, but does not change the ratio between any two entries of the matrix. As a result, the constant does not change the extent to which one amino acid pair is preferred over another. Scaling a matrix with a constant, c, can be used to obtain scores in a convenient range, *e.g.* between 1 and 20.

Although the PAM and BLOSUM matrices have the general log-likelihood framework in common, they differ in many aspects of their construction, as summarized in Table 3.1. In both cases, the frequencies of amino acid pairs, q^N_{xy}, were estimated from amino acid pair counts in *"trusted"* alignments, but these trusted alignments are different in nature. In

contrast to the PAM alignments, the BLOSUM matrices are based on locally conserved regions (ungapped blocks) in multiple alignments of sequences that were not highly conserved along their entire length. The PAM matrices were constructed from full length alignments of closely related sequences with at least 85% identity. These sequences are assumed to contain no site at which more than one substitution has occurred. The trusted alignments used to construct the BLOSUM matrices consisted of roughly 2000 blocks of conserved regions representing 500+ groups of proteins. In other words, some protein families contribute more than one block.

Both matrix families are parameterized by sequence divergence, but this is achieved using very different methods. The PAM matrices are based on a Markov chain that models amino acid replacement explicitly. The use of a Markov model allowed Dayhoff and her colleagues to address several challenges in matrix construction. A PAM-1 transition matrix is constructed from amino acid pair counts obtained from the trusted alignments. The effect of sample bias on these pair frequencies was mitigated by counting changes on the branches of maximum parsimony trees. Dayhoff accounted for both evolutionary divergence and multiple substitutions by deriving higher order Markov chains from the PAM-1 transition matrix. With PAM matrices, the divergence parameter increases with evolutionary divergence. A rough equivalence between PAMs and percent identity can be determined through simulations, as shown in Table 3.2.

The BLOSUM matrices have no underlying mathematical model. In BLOSUM matrix construction, clustering is used to address sample bias and to obtain different degrees of divergence. Sequences with at least $N\%$ identity are placed in the same cluster. Amino acid pairs are only counted across clusters, not within clusters. In contrast to the PAM matrices, the BLOSUM divergence parameter *decreases* as evolutionary divergence increases. BLOSUM matrices can also be roughly calibrated by percent identity using empirical methods, providing an approximate mapping between the PAM divergence scale and the BLOSUM divergence scale (Table 3.2).

Neither matrix family explicitly considers biophysical properties. The PAM and BLOSUM matrices are constructed from aligned sequences that are conserved because the amino acids in each column are under selective constraints. Nevertheless, the matrices favor amino acid pairs that share biochemical properties. Inspection of the BLOSUM62 matrix, for example, shows that alignments of residues in the same biochemical group tend to have positive log likelihood scores. These residues are more likely to be observed together in alignments of related sequences than by chance. Residues from different biochemical groups tend to have negative scores. These residues are less likely to be observed together in related sequences than in chance alignments. A score of zero means that this pair of residues is equally likely in related and chance alignments.

Sequence identity	PAM	BLOSUM
83%	20	-
-	30	-
63%	60	-
-	70	-
43%	100	90
38%	120	80
30%	160	60
25%	200	50
20%	250	45

Table 3.2: Correspondance between percent identity and the divergence of PAM and BLOSUM matrices.

