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.

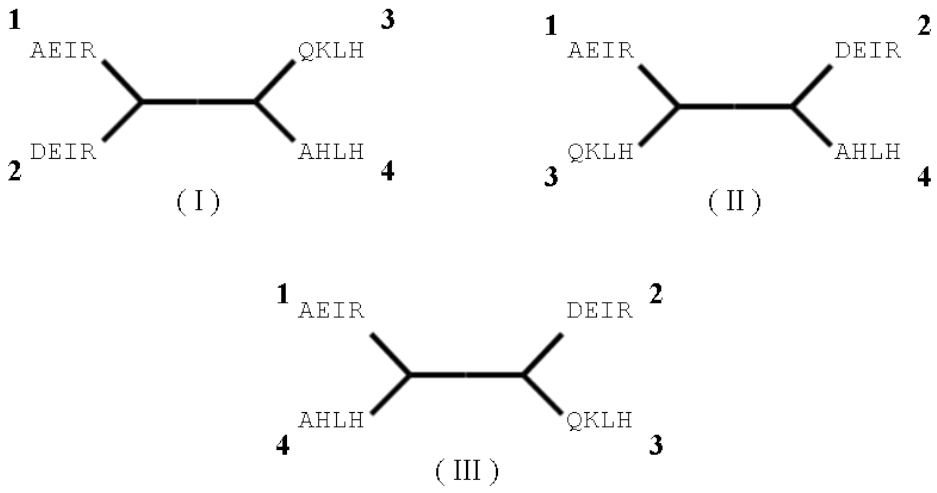

Figure 3.1: Three unrooted trees representing the three possible hypotheses for evolution of four sequences. The leaves of each tree is labeled with the corresponding present-day sequences. The internal nodes are not labeled. The sequences associated with internal nodes correspond to ancestral sequences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

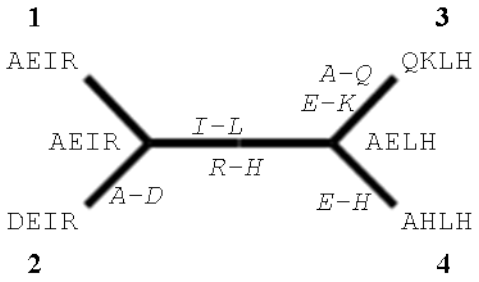

Figure 3.2: Tree I from Fig. 3.1 with ancestral sequences inferred according to the parsimony criterion. The associated amino acid replacements are shown on the branches of the tree. Six replacements are required to explain the present day sequences. This set of most parsimonous ancestral sequences is 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.

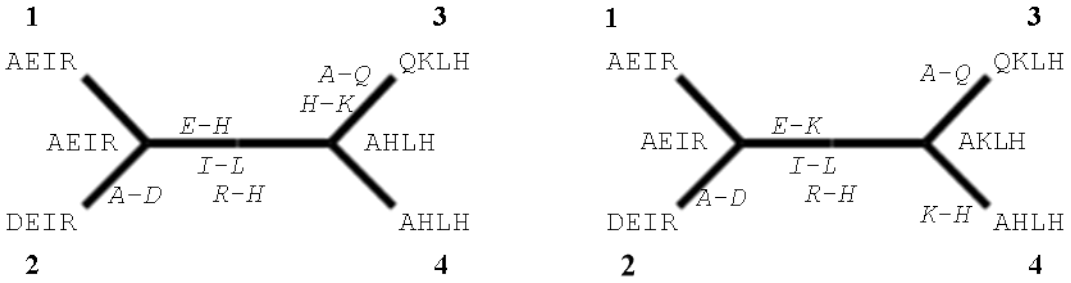

Figure 3.4: Two other sets of most parsimmonious ancestral sequences for Tree I from Fig. 3.1. The associated amino acid replacements are shown on the branches of the tree.

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.

