Lecture 3: Substitution Matrices
Substitution matrices

Algorithms in Sequence Analysis
Substitution scores $S(i,j)$

Given an alignment, what is its score?

Substitution score in sequence alignment:

\[
S(L, L) \\
S(P, G) \\
S(I, L)
\]

\[
M[i,j] = \max \begin{cases} 
M[i-1,j-1] + \text{score}(X[i], Y[j]) \\
M[i,j-1] - g \\nM[i-1,j] - g \end{cases}
\]
Definition

- Two-dimensional matrix with score values describing the probability of one amino acid or nucleotide being replaced by another during sequence evolution.
What is simplest substitution table you can think of for:

- DNA sequences
- Protein sequences
Scoring matrices for nucleotide sequences

• Can be simple:
  – e.g. positive value for match and zero (or -1) for mismatch.
  – frequencies of mutation are equal for all bases (Jukes-Cantor model).

• Can be more complicated:
  – taking into account transitions (> ) and transversions ( < ) (Kimura model).
Scoring matrices for nucleotide sequences

- **Simple model**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Kimura**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>α</td>
<td>β</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>β</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>β</td>
<td>β</td>
<td>β</td>
</tr>
</tbody>
</table>

- **Jukes-Cantor**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>α</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>β</td>
</tr>
</tbody>
</table>
Question:

What sequence similarity do you expect for 2 random DNA sequences?
Sequence similarity measures

• There are various models to correct for the fact that the true rate of evolution cannot be observed through nucleotide (or amino acid) exchange patterns (e.g. due to back mutations)
• Saturation level for nucleotide sequences is ~75%, higher real mutations are no longer observable; e.g. observed A->T through A->C->G->T or observed G->G through G->C->G (back mutation)
• Saturation level for protein sequences is ~94%
Similarity criterion for phylogeny

Observed protein sequence distance
(e.g. percent difference)

Evolutionary modeled protein sequence distance
(e.g. PAM)

94%
Evolutionary models

The Jukes–Cantor (left) and Kimura (right) models for DNA substitutions.

In the Jukes–Cantor model, all nucleotides have equal substitution rates ($\alpha$).

In the Kimura model, there are unequal rates of transitions ($\alpha$) and transversions ($\beta$). The probability values for identical matches are shaded because evolutionary distances only count different residue positions.
Mutation Model

Let $p_{ij}(t)$ be the probability that $i$ changes into $j$ in time $t$

$$P(t) = \begin{pmatrix}
p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\
p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\
p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\
p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \\
\end{pmatrix}$$

Let $Q_{ij}$ be the rate of change of substitution from $i$ to $j$

$$Q = \begin{pmatrix}
-\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\
\mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\
\mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\
\mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \\
\end{pmatrix}$$

Let $Q_{ij}$ be the rate of change of substitution from $i$ to $j$

Note that columns sum to zero: $\mu_x = \sum_{y \neq x} \mu_{xy}$
Mutations over time

For each position in a DNA molecule there are four nucleotide probabilities:

\[ P(t) = (p_A(t), p_G(t), p_C(t), p_T(t))^T \]

Then the probability of A changes proportionally to \( \mu_A \) over a small time step \( \Delta t \):

\[ p_A(t + \Delta t) = p_A(t) - p_A(t)\mu_A \Delta t + \sum_{x \neq A} p_x(t)\mu_{xA} \Delta t \]

- The same holds for the other three nucleotide types
- Formula keeps track of lost A's (negative term) and newly created A's

Writing the same using the Q matrix:

\[ P(t + \Delta t) = P(t) + QP(t)\Delta t \]
Mutations over time

Now writing the formula for $P(t+\Delta t)$ in continuous time:

$$P'(t) = QP(t)$$

Can solve this, to get the matrix $P(t)$
**JC69 model (Jukes and Cantor, 1969)**

JC69 is the simplest substitution model. There are several assumptions. It assumes equal base frequencies \((\pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{4})\) and equal mutation rates. The only parameter of this model is therefore \(\mu\), the overall substitution rate.

\[
Q = \begin{pmatrix}
* & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\
\frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\
\frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\
\frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & *
\end{pmatrix}
\]

\[
P = \begin{pmatrix}
\frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\
\frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\
\frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\
\frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu}
\end{pmatrix}
\]

Distance between two sequences is given by

\[
d = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right)
\]

where \(p\) is the proportion of sites that differ between the two sequences.
Jukes and Cantor, 1969: evolutionary distance

\[ d = \frac{3}{4} \ln \left( 1 - \frac{4}{3} \mu \right) \]

Note:

3/4 t \( \mu \) = number of expected substitutions (i \( \neq \) j)

\[ P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-t \mu} \quad \text{for i \( \neq \) j} \]
Similarity criterion for phylogeny

$\text{Observed sequence nucleotide distance (percent difference)}$

$\text{Evolutionary modeled nucleotide sequence distance (Jukes-Cantor)}$

$$d = -\frac{3}{4} \ln(1 - \frac{4}{3p})$$
Calculating true rate of DNA evolution

- **Jukes-Cantor model** (nucleotides):
  \[ d_{AB} = -(3/4) \ln[1 - (4/3)p_{AB}] \]

- E.g. \( p_{AB} = 0.3 \) (i.e. observed sequence difference is 30%)
  \[ \Rightarrow d_{AB} = -3/4 \ln[1 - (4/3 \times 0.3)] = 0.38 \]

- Model only works for closely related sequences
Scoring matrices for nucleotide sequences

- **Simple mode**
  
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Jukes-Cantor**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>α</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>β</td>
</tr>
</tbody>
</table>

- **Kimura**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>β</td>
<td>β</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>β</td>
<td>β</td>
<td>α</td>
</tr>
</tbody>
</table>

Transversions are less common than transitions.
Calculating true rate of DNA evolution

• **Kimura model** (*nucleotides*):
  \[ d_{AB} = -(1/2) \ln(1 - 2p_{ti} - p_{tv}) - (1/4) \ln(1 - 2p_{tv}) \]

• An example: let sequences A and B differ by 30%. If 20% of changes are a result of transitions \((t_i)\) and 10% of changes are a result of transversions \((t_v)\), the evolutionary distance can be calculated using

\[ d_{AB} = -1/2 \ln(1 - 2 \times 0.2 - 0.1) - 1/4 \ln(1 - 2 \times 0.1) = 0.40 \]
Calculating true rate of evolution

• For **protein sequences**: use sequence identity with *Kimura correction*:
  – Express sequence distance as \((1 – \text{fraction identity})\)
  – Protein sequences:

\[
d_{AB} = -\ln(1.0 - p_{AB} - (p_{AB})^2/5.0),
\]

where \(d_{AB}\) is the corrected distance and \(p_{AB}\) is percentage divergence between the two aligned sequences \(A\) and \(B\)

e.g. \(p_{AB} = 0.3 \rightarrow d_{AB} = -\ln(1.0 - 0.3 - (0.3)^2/5.0) = 0.38\)

• Artifact: distances can become negative (set to zero)

• The popular multiple sequence alignment method ClustalW uses sequence identity with Kimura (1983) correction
Other evolutionary models
The Jukes-Cantor and Kimura models assume equal base rates (stationary frequencies becoming equal in infinite evolutionary time)

There are more complex models, such as
• F81 model (Felsenstein 1981)
• HKY85 (Hasegawa, Kishino and Yano 1985)
• T92 (Tamura 1992)
• TN93 model (Tamura and Nei 1993)
• GTR: Generalised time-reversible

these do not assume equal base rates and take many more parameters into consideration. However, these more complex models are normally not used in practice because the calculations are complicated and the variance levels resulting from their formulae are typically high.
How are these models used in sequence alignment

The main use in alignment is ‘correcting’ the alignment scores (Jukes-Cantor and Kimura models for DNA; (another) Kimura model for proteins)

However, the majority of (multiple) alignment methods do not use this correction, since it is based upon crude sequence identity scores that only score residue matches

An important use however is in Phylogeny, where attempts are made to reconstruct the most-likely evolutionary decendance of sequences, particularly in Maximum Likelihood methods (see later lecture)
What is better to align? DNA or protein sequences?

1. Many mutations within DNA are synonymous ⇒ divergence overestimation
2. Evolutionary relationships can be more accurately expressed using a **20×20 amino acid exchange table**

3. DNA sequences contain **non-coding regions**, which should be avoided in homology searches.

4. Still an issue when translating into (six) protein sequences through a codon table.

5. Searching at protein level: **frameshifts** can occur, leading to stretches of incorrect amino acids and possibly elongation.

   However, frameshifts normally result in stretches of highly unlikely amino acids.
So?

Rule of thumb:

⇒ if ORF (open reading frame) exists, then align at protein level
Scoring matrices
4 nucleotides - 20 amino acids

purines
pyrimidines

Acidic and amide side chains
- Aspartate
- Asparagine

Aromatic side chains
- Tryptophan
- Phenylalanine

Basic side chains
- Lysine
- Histidine

Aliphatic side chains
- Valine
- Isoleucine

Hydroxyl or sulfur-containing side chains
- Serine
- Methionine

Cyclic side chain
- Proline
Scoring matrices for amino acid sequences

- Are complicated, scoring has to reflect:
  - Physio-chemical properties of aa’s
  - Likelihood of residues being substituted among truly homologous sequences

- Certain aa with similar properties can be more easily substituted: preserve structure/function

- “Disruptive” substitution is less likely to be selected in evolution (non functional proteins)
Taylor’s Venn diagrams (Taylor, 1986)
Example of environment:

Cysteines are very common in metal binding motifs

Zinc finger (@wiki)
Now let’s think about alignments

• Let’s consider a simple alignment: ungapped global alignment of two (protein) sequences, x and y, of length n.

• In scoring this alignment, we would like to assess whether these two sequences have a common ancestor, or whether they are aligned by chance.

\[
\frac{Pr(x, y \mid M)}{Pr(x, y \mid R)}
\]

\( \Rightarrow \) sequences have common ancestor (Match)

\( \Rightarrow \) sequences are aligned by chance (Random)

• We therefore want our amino acid substitution table (matrix) to score an alignment by estimating this ratio (= improvement over random).

• In brief, each substitution score is the log-odds probability that amino acid a could change (mutate) into amino acid b through evolution, based on the constraints of our evolutionary model.
Target and background probabilities

• Background probability

If $q_a$ is the frequency of amino acid $a$ in one sequence and $q_b$ is the frequency of amino acid $b$ in another sequence, then the probability of the alignment being random is given by:

$$\Pr(x, y \mid R) = \prod_i q_{x_i} \prod_i q_{y_i}$$

• Target probability

If $p_{ab}$ is now the probability that amino acids $a$ and $b$ have derived from a common ancestor, then the probability that the alignment is due to common ancestry is given by:

$$\Pr(x, y \mid M) = \prod_i p_{x_i, y_i}$$
Source of target and background probabilities: 
high confidence alignments

• Target frequencies
  – The “evolutionary true” alignments allow us to get statistics on biologically permissible amino acid mutations and derive the frequencies of observed pairs. These are the TARGET frequencies (20x20 combinations).

• Background frequencies
  – The BACKGROUND frequencies are simply the frequencies at which each amino acid type is observed in these “trusted” data sets (20 values).
Log-odds

- Substitution matrices apply logarithmic conversions to describe the probability of amino acid substitutions.

- The converted values are the so-called log-odds scores.

- So they are simply the logarithmic ratios of the observed mutation frequency divided by the probability of substitution expected by random chance (target – background).
Formulas

• **Odds-ratio** of two probabilities

\[
\frac{\Pr(x, y \mid M)}{\Pr(x, y \mid R)} = \frac{\prod_i p_{x_iy_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \frac{\prod_i p_{x_iy_i}}{\prod_i q_{x_i} q_{y_i}}
\]

• **Log-odds** probability of an alignment being random is therefore given by

\[
\log \frac{\Pr(x, y \mid M)}{\Pr(x, y \mid R)} = \sum \log \left( \frac{p_{x_iy_i}}{q_{x_i} q_{y_i}} \right)
\]

\[
\log \left( \prod_i x \right) = \sum \log x
\]
So... for a given substitution matrix:

$$\log \frac{\Pr(x, y \mid M)}{\Pr(x, y \mid R)}$$

- **a positive score**
  means that the frequency of amino acid substitutions found in the high-confidence alignments is greater than would have occurred by random chance

- **a zero score**
  … that the frequency is equal to that expected by chance

- **a negative score**
  … that the frequency is less than that expected by chance
Alignment score

• The alignment score \( S \) is given by the sum of all amino acid pair substitution scores:

\[
S = \sum_i s(x_i, y_i) = \log \frac{\Pr(x, y \mid M)}{\Pr(x, y \mid R)}
\]

• Where the substitution score for any amino acid pair \([a,b]\) is given by:

\[
s(a, b) = \log \frac{p_{ab}}{q_a q_b}
\]
Alignment score

• The total score of an alignment:

\[ EAAS \]
\[ VF - T \]

• would be:

\[ S = s(E,V) + s(A,F) + \gamma(1) + s(S,T) \]
Empirical matrices

• Are based on surveys of actual amino acid substitutions among related proteins

• Most widely used: PAM and BLOSUM
The PAM series

- The first systematic method to derive amino acid substitution matrices was done by Margaret Dayhoff et al. (1978) *Atlas of Protein Structure*.

- These widely used substitution matrices are frequently called Dayhoff, MDM (Mutation Data Matrix), or PAM (Point Accepted Mutation) matrices.

- **Key idea**: trusted alignments of closely related sequences provide information about biologically permissible mutations.
The PAM design

1. Step 1. Dayhoff used 71 protein families (each family having closely related family members), made hypothetical phylogenetic trees and recorded the number of observed substitutions (along each branch of the tree) in a 20x20 target matrix.

Mutations are counted in both directions, i.e. $A \rightarrow B = B \rightarrow A$.
This leads to a symmetrical mutation matrix $A$ (holding accepted point mutations).
The PAM design

- **Step 2.** The target matrix was then converted to frequencies by dividing each cell \((a,b)\) over the sum of all other substitutions of \(a\).

\[
\Pr(b \mid a) = \frac{A_{ab}}{\sum_c A_{ac}}
\]

- **Step 3.** The target matrix was normalized so that the expected number of substitutions covered 1% of the protein (PAM-1).

\[
\Pr(b \mid a, t = 1)
\]

- **Step 4.** Determine the final substitution matrix.

\[
s(a, b \mid t) = \log \frac{p_{ab}}{q_a q_b} = \log \frac{P(b \mid a, t)}{q_b}
\]
So, how does it work with real data?

- the PAM calculation steps -
PAM: Phylogenetic Tree
inferred ancestral sequences and derived mutations
# PAM: Accepted Point Mutation

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

$A_{ij}$: number of times amino acid $j$ mutates to amino acid $i$. A mutation could go in both directions, therefore the tally of mutation $i$-j enters both $A_{ij}$ and $A_{ji}$ entries, while the tally of conservation $i$-i enters $A_{ii}$ entry twice.
Mutability of Residue $j$

$$m_j = 1 - \frac{A_{jj}}{\sum_{i=1,20} A_{ij}} = \frac{\sum_{i=1,20; i \neq j} A_{ij}}{\sum_{i=1,20} A_{ij}}$$

$m_j$ is the probability that amino acid $j$ will change in a given evolutionary interval. The absolute values of $m_j$ depend on how similar the sequences used to tally $A_{ij}$ are,

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser</td>
<td>149</td>
</tr>
<tr>
<td>Met</td>
<td>122</td>
</tr>
<tr>
<td>Asn</td>
<td>111</td>
</tr>
<tr>
<td>Ile</td>
<td>110</td>
</tr>
<tr>
<td>Glu</td>
<td>102</td>
</tr>
<tr>
<td>Ala</td>
<td>100</td>
</tr>
<tr>
<td>Gln</td>
<td>98</td>
</tr>
<tr>
<td>Asp</td>
<td>90</td>
</tr>
<tr>
<td>Thr</td>
<td>90</td>
</tr>
<tr>
<td>Trp</td>
<td>22</td>
</tr>
<tr>
<td>Val</td>
<td>80</td>
</tr>
<tr>
<td>Lys</td>
<td>57</td>
</tr>
<tr>
<td>Pro</td>
<td>56</td>
</tr>
<tr>
<td>His</td>
<td>50</td>
</tr>
<tr>
<td>Gly</td>
<td>48</td>
</tr>
<tr>
<td>Phe</td>
<td>45</td>
</tr>
<tr>
<td>Arg</td>
<td>44</td>
</tr>
<tr>
<td>Leu</td>
<td>38</td>
</tr>
<tr>
<td>Tyr</td>
<td>34</td>
</tr>
<tr>
<td>Cys</td>
<td>27</td>
</tr>
<tr>
<td>Gap</td>
<td>84</td>
</tr>
</tbody>
</table>

The value of Ala ($m_{\text{Ala}}$) has been set arbitrarily to 100 and the values of all other amino acids scaled accordingly. (Adapted from Table 21. Atlas of Protein Sequence and Structure, Suppl 3, 1978, M.O. Dayhoff, ed. National Biomedical Research Foundation, 1979.)
Total Mutation Rate

\[ P_j \text{ is the probability of random occurrence of amino acid } j \]

\[ P_j = \frac{\sum_{i=1,20} A_{ij}}{\sum_{i=1,20} \sum_{j=1,20} A_{ij}} \]

\[ \sum_{j=1,20} P_j m_j \text{ is the total mutation rate of all amino acids} \]
Normalize Total Mutation Rate to 1%

$\lambda$ is a scaling constant to make sure that the total mutation rate is $1\%$

$$\lambda \cdot \sum_{j=1,20} P_j m_j = 1\% \implies \text{solve for } \lambda$$

This defines an evolutionary period: the period during which the $1\%$ of all sequences are mutated (accepted of course)
Mutation Probability Matrix
normalized such that the
Total Mutation Rate is 1%

\[ M_{ij} \ (i \neq j) : \text{Probability of amino acid } j \text{ changing into } i \text{ in the evolutionary period} \]

\[ M_{ij} = \lambda \frac{A_{ij}}{\sum_{i=1,20} A_{ij}} \quad \text{-- Normalised using frequency of amino acid } j \]

\[ M_{jj} : \text{Probability of amino acid } j \text{ not changing in PAM-1} \]

\[ M_{jj} = 1 - \sum_{i=1,20; i \neq j} M_{ij} = 1 - \lambda m_j \]
### Mutation Probability Matrix

(transposed) $M^{*100000}$

<table>
<thead>
<tr>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>L</th>
<th>K</th>
<th>M</th>
<th>F</th>
<th>P</th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>Y</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9867</td>
<td>2</td>
<td>9</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>17</td>
<td>21</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>35</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>9913</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>19</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>4</td>
<td>1</td>
<td>9822</td>
<td>36</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>21</td>
<td>3</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td>9</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>0</td>
<td>42</td>
<td>9859</td>
<td>0</td>
<td>6</td>
<td>53</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9973</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Q</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>9876</td>
<td>27</td>
<td>1</td>
<td>23</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>56</td>
<td>0</td>
<td>35</td>
<td>9865</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>21</td>
<td>1</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>9935</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>21</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>18</td>
<td>8</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>9912</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>9872</td>
<td>9</td>
<td>2</td>
<td>12</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>33</td>
<td>9</td>
</tr>
<tr>
<td>L</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>22</td>
<td>9947</td>
<td>2</td>
<td>45</td>
<td>13</td>
<td>3</td>
<td>13</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>37</td>
<td>25</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>9926</td>
<td>20</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>9874</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>9946</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>P</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>9926</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>28</td>
<td>11</td>
<td>34</td>
<td>7</td>
<td>11</td>
<td>4</td>
<td>6</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>9840</td>
<td>38</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>22</td>
<td>2</td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>32</td>
<td>9871</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>V</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>57</td>
<td>11</td>
<td>1</td>
<td>17</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

**Note**: This matrix (based on 1% mutation) is conservative.  
**Note**: This matrix is NOT symmetrical (why not?)
$M^{(1)} = \text{PAM1 Mutation Probability Matrix}$

$M^{(2)} = \text{PAM2 Mutation Probability Matrix}$

-- Mutations that happen in twice the evolution period of that for a PAM1
PAM Matrix: Assumptions

- The likelihood of amino acid Y replacing X is the same as that of X replacing Y:
  \[ P(X \rightarrow Y) = P(Y \rightarrow X) \]

- Very closely related proteins are used to decrease the mediated mutations such as \( X \rightarrow Z \rightarrow Y \)

- Replacement at any site depends only on the amino acid at that site and the probability given by a Markov Model; all positions in a protein are equally mutable

- All sequences have average amino acid composition
In two PAM1 periods:

- $\{A \rightarrow R\} = \{A \rightarrow A \text{ and } A \rightarrow R\}$ or
  $\{A \rightarrow N \text{ and } N \rightarrow R\}$ or
  $\{A \rightarrow D \text{ and } D \rightarrow R\}$ or
  ... or
  $\{A \rightarrow V \text{ and } V \rightarrow R\}$
Entries in a PAM-2 Mutation Probability Matrix

\[ \text{Pr}(A \rightarrow R \text{ in 2 periods}) = \]
\[ \text{Pr}(A \rightarrow A \text{ in 1st period}) \times \text{Pr}(A \rightarrow R \text{ in 2nd period}) + \]
\[ \text{Pr}(A \rightarrow N \text{ in 1st period}) \times \text{Pr}(N \rightarrow R \text{ in 2nd period}) + \]
\[ \text{Pr}(A \rightarrow D \text{ in 1st period}) \times \text{Pr}(D \rightarrow R \text{ in 2nd period}) + \]
\[ \ldots \]

\[ P_{AR}^{(2)} = P_{AA} \cdot P_{AR} + P_{AN} \cdot P_{NR} + P_{AD} \cdot P_{DR} + \ldots \]
PAM-k Mutation Prob. Matrix

\[ M^{(2)} = M^{(1)} \times M^{(1)} \]

\[ M^{(K)} = \{ M^{(1)} \}^K \]
PAM units

• One PAM unit is defined as 1% of the amino acids positions that have been changed.

• E.g. to construct the PAM1 substitution table, a group of closely related sequences with mutation frequencies corresponding to one PAM unit is chosen. One PAM corresponds to about 1 million years of evolutionary time.

• How do you extend from PAM1 (1 Myears) to PAM2 (2 Myears), PAM3 (3 Myears), etc.
  – A->C in 2 Myears develops via A->B->C (over all possible B residues)
But there is a whole series of matrices: PAM10 … PAM250

- These matrices are extrapolated from PAM1 matrix (by matrix multiplication)

Self-multiply Matrix $N$ times to make PAM ‘$N$’; then take the Log

- So: a PAM is a relative measure of evolutionary distance
  - 1 PAM = 1 accepted mutation per 100 amino acids
  - 250 PAM = 250 mutations per 100 amino acids, so 2.5 accepted mutations per amino acid!
PAM-k log-likelihood matrix

\[ S_{ij} = 10 \log_{10} \left( \frac{(M^K)_{ij}}{P_i} \right) \]

Note that only the frequency of amino acid type \( i \) (\( P_i \)) is used to normalise the mutation probabilities, because the \( M_{ij} \) values are already normalised with respect to the frequency of amino acid type \( j \).

\[ P_i = \frac{\sum \ A_{ij}}{\sum_{j=1,20} \sum_{i=1,20} A_{ij}} \]

\( P_i \) is the probability of random occurrence of amino acid \( i \)

\( S \) is a symmetric matrix

The thus converted scores (after rounding) represent the PAM matrices as we know them.
**PAM numbers vs. observed amino-acid mutational rates**

<table>
<thead>
<tr>
<th>PAM Number</th>
<th>Observed Mutation Rate (%)</th>
<th>Sequence Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>99</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>75</td>
</tr>
<tr>
<td>80</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>110</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>200</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>250</td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>

*Note* Think about intermediate “substitution” steps ...
The PAM250 matrix

|   | A 2 | R -2 6 | N 0 0 2 | D 0 -1 2 4 | C -2 -4 -4 -5 12 | Q 0 1 1 2 -5 4 | E 0 -1 1 3 -5 2 4 | G 1 -3 0 1 -3 -1 0 5 | H -1 2 2 1 -3 3 1 -2 6 | I -1 -2 -2 -2 -2 -2 -3 -2 5 | L -2 -3 -3 -4 -6 -2 -3 -4 -2 2 6 | K -1 3 1 0 -5 1 0 -2 0 -2 -3 5 | M -1 0 -2 -3 -5 -1 -2 -3 -2 2 4 0 6 | F -4 -4 -4 -6 -4 -5 -5 -5 -2 1 2 -5 0 9 | P 1 0 -1 -1 -3 0 -1 -1 0 -2 -3 -1 -2 -5 6 | S 1 0 1 0 0 -1 0 1 -1 -1 -3 0 -2 -3 1 2 | T 1 -1 0 0 -2 -1 0 0 -1 0 -2 0 -1 -3 0 1 3 | W -6 2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4 0 -6 -2 -5 17 | Y -3 -4 -2 -4 0 -4 -4 -5 0 -1 -1 -4 -2 7 -5 -3 -3 0 10 | V 0 -2 -2 -2 -2 -2 -1 -2 4 2 -2 2 -1 -1 -1 0 -6 -2 4 | B 0 -1 2 3 -4 1 2 0 1 -2 -3 1 -2 -5 -1 0 0 -5 -3 -2 2 | Z 0 0 1 3 -5 3 3 -1 2 -2 -3 0 -2 -5 0 0 -1 -6 -4 -2 2 3 |

ARNDCEQEGHILKMFPSTWYVBDZ

W- R exchange is too large (due to paucity of data)
Scoring matrices for amino acid sequences

- Are complicated, scoring has to reflect:
  - Physio-chemical properties of aa’s
  - Likelihood of residues being substituted among truly homologous sequences
- Certain aa with similar properties can be more easily substituted: preserve structure/function
- “Disruptive” substitution is less likely to be selected in evolution (e.g. rendering non-functional proteins)
PAM model

- The scores derived through the PAM model are an accurate description of the information content (or the relative entropy) of an alignment (Altschul, 1991).

- **PAM1** corresponds to about 1 million years of evolution.

- **PAM120** has the largest information content of the PAM matrix series: “best” for general alignment.

- **PAM250** is the traditionally most popular matrix: “best” for detecting distant sequence similarity.
Summary Dayhoff’s PAM-matrices

• Derived from global alignments of closely related sequences.

• Matrices for greater evolutionary distances are extrapolated from those for smaller ones
  – Evolutionary clock (mutation rate) is supposed to be uniform (equal over evolutionary time)

• The number with the matrix (PAM40, PAM100) refers to the evolutionary distance; greater numbers are greater distances.

• Attempts to extend Dayhoff's methodology or re-apply her analysis using databases with more examples:
  – Jones, Thornton and coworkers used the same methodology as Dayhoff but with modern databases (CABIOS 8:275)
  – Gonnett and coworkers (Science 256:1443) used a slightly different (but theoretically equivalent) methodology
The BLOSUM series

• BLOSUM stands for: **BLOcks SUbstitution Matrices**

• Created by Steve Henikoff and Jorja Henikoff (PNAS 89:10915).

• Derived from local, un-gapped alignments of distantly related sequences (the BLOCKS database)

• All matrices are directly calculated; no extrapolations (PAM) are used.

• As with PAM: compare observed freqs of each pair to expected freqs -- Then: Log-odds matrix.
The **Blocks** database

- The Blocks Database contains multiple alignments of conserved regions in protein families.

- Blocks are multiply aligned un-gapped segments corresponding to the most highly conserved regions of proteins.

- The blocks for the BLOCKS database are made automatically by looking for the most highly conserved regions in groups of proteins represented in the PROSITE database. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the random distribution of matches. It is these calibrated blocks that make up the BLOCKS database.

- Gapless alignment blocks

---

![Gapless alignment blocks](image-url)
The BLOSUM series

- **BLOSUM30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, 90.**

- The number after the matrix (BLOSUM62) refers to the minimum percent identity of the blocks (in the BLOCKS database) used to construct the matrix (for BLOSUM62 all blocks have >=62% sequence identity);
  - e.g., for BLOSUM30 blocks are used with >= 30% sequence identity

- No extrapolations are made in going to higher evolutionary distances
  - Each time another selection of the BLOCKS alignments is used

- High number - closely related sequences
  - Low number - distant sequences

- BLOSUM62 is the most popular exchange matrix: best for general alignment.
A toy example of constructing a BLOSUM matrix from 4 training sequences
Constructing a BLOSUM matr.

1. Counting mutations

VVAPV
AAAPA
PVAPV
PAAAV

\[ N_{AA} = 0 + 1 + (4 \times 3/2) + 0 + 0 = 7 \]
\[ N_{VV} = 0 + 1 + 0 + 0 + (3 \times 2)/2 = 4 \]
\[ N_{PP} = 1 + 0 + 0 + (3 \times 2)/2 + 0 = 4 \]

\[ N_{AV} = N_{VA} = 1 + 2 \times 2 + 0 + 0 + 3 = 8 \]
\[ N_{AP} = N_{PA} = 2 + 0 + 0 + 3 + 0 = 5 \]
\[ N_{PV} = N_{VP} = 2 + 0 + 0 + 0 + 0 + 0 = 2 \]

\( N_{VP} \) is the number of V-P pairs
2. Tallying mutation frequencies

\[
q_{ij} = \begin{array}{cccc}
  & A & V & P \\
A & 14 & 8 & 5 \\
V & 8 & 8 & 2 \\
P & 5 & 2 & 8 \\
\end{array}
\]

\( q_{ij} \): number of times amino acid \( j \) mutates to amino acid \( i \).
A mutation could go in both directions, therefore the tally of A-V pair enters both \( q_{AV} \) and \( q_{VA} \) entries, while the tally of A-A pair enters \( q_{AA} \) entry twice.
3. Matrix of mutation probs.

$p_{ij}$ is the probability that a mutation occurs between amino acid $i$ and amino acid $j$

$$p_{ij} = \frac{q_{ij}}{\sum_i \sum_j q_{ij}}$$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>V</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>14/60</td>
<td>8/60</td>
<td>5/60</td>
</tr>
<tr>
<td>V</td>
<td>8/60</td>
<td>8/60</td>
<td>2/60</td>
</tr>
<tr>
<td>P</td>
<td>5/60</td>
<td>2/60</td>
<td>8/60</td>
</tr>
</tbody>
</table>
4. Calculate abundance of each residue (Marginal prob)

\[ p_i = \frac{\sum_{j=1,20} q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}} \]

\[ p_i = \begin{array}{c}
A \\
9/20
\end{array} \quad \begin{array}{c}
V \\
6/20
\end{array} \quad \begin{array}{c}
P \\
5/20
\end{array} \]

\[ (27/60) \quad (18/60) \quad (15/60) \]
5. Obtaining a BLOSUM matrix

The BLOSUM log-likelihood matrix:

\[ S_{ij} = 2 \log_2 \frac{p_{ij}}{p_i p_j} \]

\[
\begin{array}{cccc}
  & A & V & P \\
A & 0.409 & & \\
V & -0.036 & 1.134 & \\
P & -0.866 & -2.34 & 2.19 \\
\end{array}
\]
Constructing the real BLOSUM62 Matrix
1.2.3. Mutation Frequency Table

\[ P_{ij} \cdot 1000 \]

|   | A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 215| 23 | 178|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| R | 23 | 20 | 141|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| N | 19 | 20 | 141|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| D | 22 | 16 | 37 | 213|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| C | 16 | 4  | 4  | 4  | 119|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Q | 19 | 25 | 15 | 16 | 3  | 73 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| E | 30 | 27 | 22 | 49 | 4  | 35 | 161|    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| G | 58 | 17 | 29 | 25 | 8  | 14 | 19 | 378|    |    |    |    |    |    |    |    |    |    |    |    |    |
| H | 11 | 12 | 14 | 10 | 2  | 10 | 14 | 10 | 93 |    |    |    |    |    |    |    |    |    |    |    |    |
| I | 32 | 12 | 10 | 12 | 11 | 9  | 12 | 14 | 6  | 184|    |    |    |    |    |    |    |    |    |    |    |
| L | 44 | 24 | 14 | 15 | 16 | 16 | 20 | 21 | 10 | 114| 371|    |    |    |    |    |    |    |    |    |    |
| K | 33 | 62 | 24 | 24 | 5  | 31 | 41 | 25 | 12 | 16 | 25 | 161|    |    |    |    |    |    |    |    |    |
| M | 13 | 8  | 5  | 5  | 4  | 7  | 7  | 7  | 4  | 25 | 49 | 9  | 40 |    |    |    |    |    |    |    |    |
| F | 16 | 9  | 8  | 5  | 5  | 9  | 12 | 8  | 30 | 54 | 9  | 12 | 183|    |    |    |    |    |    |    |    |
| P | 22 | 10 | 9  | 12 | 4  | 8  | 14 | 14 | 5  | 10 | 14 | 16 | 4  | 5  | 191|    |    |    |    |    |    |
| S | 63 | 23 | 31 | 28 | 10 | 19 | 30 | 38 | 11 | 17 | 24 | 31 | 9  | 12 | 17 | 126|    |    |    |    |    |
| T | 37 | 18 | 22 | 19 | 9  | 14 | 20 | 22 | 7  | 27 | 33 | 23 | 10 | 12 | 14 | 47 | 125|    |    |    |    |
| W | 04 | 3  | 2  | 2  | 1  | 2  | 3  | 4  | 2  | 4  | 7  | 3  | 2  | 8  | 1  | 3  | 3  | 65 |    |    |    |
| Y | 13 | 9  | 7  | 6  | 3  | 7  | 9  | 8  | 15 | 14 | 22 | 10 | 6  | 42 | 5  | 10 | 9  | 9  | 102|    |    |
| V | 51 | 16 | 12 | 13 | 14 | 12 | 17 | 18 | 6  | 120| 95 | 19 | 23 | 26 | 12 | 24 | 36 | 4  | 15 | 196|

A R N D C Q E G H I L K M F P S T W Y V
4. Calculate Amino Acid Abundance

\[ p_i = \sum_j p_{ij} \]

A: 0.074  R: 0.052  N: 0.045  D: 0.054  C: 0.025
Q: 0.034  E: 0.054  G: 0.074  H: 0.026  I: 0.068
L: 0.099  K: 0.058  M: 0.025  F: 0.047  P: 0.039
S: 0.057  T: 0.051  W: 0.013  Y: 0.032  V: 0.073
### 5. Obtaining BLOSUM62 Matrix

A  4
R -1  5
N -2  0  6
D -2 -2  1  6
C  0 -3 -3 -3  9
Q -1  1  0  0 -3  5
E -1  0  0  2 -4  2  5
G  0 -2  0 -1 -3 -2 -2  6
H -2  0  1 -1 -3  0  0 -2  8
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W -3 -3 -4 -4 -2 -2 -3 -2 -3 -2 -3 -1  1 -4 -3 -2  1  1
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
        A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V

\[
S_{ij} = 2 \cdot \log_2 \frac{p_{ij}}{p_i p_j}
\]
The log-odds matrix for **BLOSUM62**

<table>
<thead>
<tr>
<th>A</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>-1 5</td>
</tr>
<tr>
<td>N</td>
<td>-2 0 6</td>
</tr>
<tr>
<td>D</td>
<td>-2 -2 1 6</td>
</tr>
<tr>
<td>C</td>
<td>0 -3 -3 -3 9</td>
</tr>
<tr>
<td>Q</td>
<td>-1 1 0 0 -3 5</td>
</tr>
<tr>
<td>E</td>
<td>-1 0 0 0 -4 2 5</td>
</tr>
<tr>
<td>G</td>
<td>0 -2 0 -1 -3 -2 -2 6</td>
</tr>
<tr>
<td>H</td>
<td>-2 0 1 -1 -3 0 0 -2 8</td>
</tr>
<tr>
<td>I</td>
<td>-1 -3 -3 -3 -1 -3 -3 -4 -3</td>
</tr>
<tr>
<td>L</td>
<td>-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4</td>
</tr>
<tr>
<td>K</td>
<td>-1 2 0 -1 -3 1 1 -2 -1 -3 -2 5</td>
</tr>
<tr>
<td>M</td>
<td>-1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5</td>
</tr>
<tr>
<td>F</td>
<td>-2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6</td>
</tr>
<tr>
<td>P</td>
<td>-1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7</td>
</tr>
<tr>
<td>S</td>
<td>1 -1 1 0 -1 0 0 0 0 -1 -2 -2 0 -1 -2 -1 4</td>
</tr>
<tr>
<td>T</td>
<td>0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -2 -1 1 5</td>
</tr>
<tr>
<td>W</td>
<td>-3 -3 4 -4 2 -2 -3 2 -2 -3 2 3 -1 1 -4 3 2 11</td>
</tr>
<tr>
<td>Y</td>
<td>-2 -2 2 -3 -2 -1 -2 -3 2 1 -1 -2 -1 3 -3 -2 -2 2 7</td>
</tr>
<tr>
<td>V</td>
<td>0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4</td>
</tr>
<tr>
<td>X</td>
<td>0 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 0 0 -2 -1 -1 -1</td>
</tr>
</tbody>
</table>

**Positive for chemically similar substitution**

**Common amino acids have low weights**

**Rare amino acids have high weights**

Higher random chance to become aligned (observed freqs versus expected freqs)
PAM versus BLOSUM

- Based on an explicit evolutionary model
- Derived from small, closely related proteins with ~15% divergence
- Higher PAM numbers to detect more remote sequence similarities
- Errors in PAM 1 are scaled 250X in PAM 250

- Based on empirical frequencies
- Uses much larger, more diverse set of protein sequences (30-90% ID)
- Lower BLOSUM numbers to detect more remote sequence similarities
- Errors in BLOSUM arise from errors in alignment
Comparing exchange matrices

To compare amino acid exchange matrices, the "Entropy" value can be used. This is a relative entropy value (H) which describes the amount of information available per aligned residue pair.

$$H = \sum s_{ij} \log_2(s_{ij} / p_i p_j)$$

<table>
<thead>
<tr>
<th>NAME</th>
<th>CLUSTER</th>
<th>#blocks</th>
<th>#PAIRS</th>
<th>ENTROPY</th>
<th>EXPECTED</th>
<th>NAME</th>
<th>CLUSTER</th>
<th>#blocks</th>
<th>#PAIRS</th>
<th>ENTROPY</th>
<th>EXPECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blosum30</td>
<td>30%</td>
<td>261</td>
<td>70,394</td>
<td>.1424</td>
<td>-.1074</td>
<td>Pam160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam350</td>
<td></td>
<td></td>
<td></td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam340</td>
<td></td>
<td></td>
<td></td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum35</td>
<td>35</td>
<td>439</td>
<td>136,328</td>
<td>.2111</td>
<td>-.1550</td>
<td>Pam150</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam330</td>
<td></td>
<td></td>
<td></td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam290</td>
<td></td>
<td></td>
<td></td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum40</td>
<td>40</td>
<td>672</td>
<td>227,179</td>
<td>.2851</td>
<td>-.2090</td>
<td>Pam130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam280</td>
<td></td>
<td></td>
<td></td>
<td>.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam250a</td>
<td></td>
<td></td>
<td></td>
<td>.36</td>
<td>-.28</td>
<td>Pam120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum45</td>
<td>45</td>
<td>900</td>
<td>357,558</td>
<td>.3795</td>
<td>-.2789</td>
<td>Blosum80</td>
<td>80</td>
<td>1941</td>
<td>2,907,119</td>
<td>.9868</td>
<td>-.7442</td>
</tr>
<tr>
<td>Pam240</td>
<td></td>
<td></td>
<td></td>
<td>.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam210</td>
<td></td>
<td></td>
<td></td>
<td>.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum50</td>
<td>50</td>
<td>1161</td>
<td>557,053</td>
<td>.4808</td>
<td>-.3573</td>
<td>Pam110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam200</td>
<td></td>
<td></td>
<td></td>
<td>.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam190</td>
<td></td>
<td></td>
<td></td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum55</td>
<td>55</td>
<td>1331</td>
<td>753,984</td>
<td>.5637</td>
<td>-.4179</td>
<td>Pam90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam180</td>
<td></td>
<td></td>
<td></td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam170</td>
<td></td>
<td></td>
<td></td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum60</td>
<td>60</td>
<td>1529</td>
<td>1,087,140</td>
<td>.6603</td>
<td>-.4917</td>
<td>Blosum100</td>
<td>100</td>
<td>2089</td>
<td>11,066,715</td>
<td>1.4516</td>
<td>-.10948</td>
</tr>
<tr>
<td>Blosum62</td>
<td>62</td>
<td>1572</td>
<td>1,245,852</td>
<td>.6979</td>
<td>-.5209</td>
<td>Pam70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam500</td>
<td></td>
<td></td>
<td></td>
<td>.70</td>
<td>-.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum63</td>
<td>63</td>
<td>1606</td>
<td>1,322,720</td>
<td>.7218</td>
<td>-.5389</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum65</td>
<td>65</td>
<td>1671</td>
<td>1,503,060</td>
<td>.7576</td>
<td>-.5675</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam140</td>
<td></td>
<td></td>
<td></td>
<td>.76</td>
<td>-.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam70</td>
<td></td>
<td></td>
<td></td>
<td>.882</td>
<td>-.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum70</td>
<td>70</td>
<td>1803</td>
<td>1,930,219</td>
<td>.8391</td>
<td>-.6313</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam130</td>
<td></td>
<td></td>
<td></td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum75</td>
<td>75</td>
<td>1896</td>
<td>2,365,918</td>
<td>.9077</td>
<td>-.6845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam120</td>
<td></td>
<td></td>
<td></td>
<td>.98</td>
<td>-.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum80</td>
<td>80</td>
<td>1941</td>
<td>2,907,119</td>
<td>.9868</td>
<td>-.7442</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam110</td>
<td></td>
<td></td>
<td></td>
<td>1.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum85</td>
<td>85</td>
<td>1993</td>
<td>3,861,389</td>
<td>1.0805</td>
<td>-.8153</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam100</td>
<td></td>
<td></td>
<td></td>
<td>1.18</td>
<td>-.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum90</td>
<td>90</td>
<td>2022</td>
<td>5,196,613</td>
<td>1.1806</td>
<td>-.8887</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum95</td>
<td>95</td>
<td>2050</td>
<td>7,185,859</td>
<td>1.2926</td>
<td>-.9743</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam90</td>
<td></td>
<td></td>
<td></td>
<td>1.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam80</td>
<td></td>
<td></td>
<td></td>
<td>1.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosum100</td>
<td>100</td>
<td>2089</td>
<td>11,066,715</td>
<td>1.4516</td>
<td>-.10948</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blosumna</td>
<td></td>
<td></td>
<td></td>
<td>1.5172</td>
<td>-.1484</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pam70</td>
<td></td>
<td></td>
<td></td>
<td>1.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evolution and Matrix “landscape”

- Recent evolution → identity matrix
- Ancient evolution → convergence to random model
A note on reliability

• All these matrices are designed using empirical evolutionary models.

• Circular problem

• It is important to understand that evolution is not the same for all proteins, not even for the same regions of proteins.
In Practice

• No single matrix performs best on all sequences. Some are better for sequences with few gaps, and others are better for sequences with fewer identical amino acids.

• Therefore, when aligning sequences, applying a general model to all cases is not ideal. Rather, re-adjustment can be used to make the general model better fit the given data.

• In practice, matrices such as BLOSUM62 (or sometimes BLOSUM50) are used in the vast majority of cases.
Pair-wise alignment quality *versus* sequence identity

- Vogt et al., JMB 249, 816-831, 1995

Pairwise alignments were made of sequence pairs for which the 'true' alignment was known from 3D-structural information, so the correctness of the alignments could be checked.
Take-home messages - 1

• If ORF exists, then align at protein level.

• Amino acid substitution matrices reflect the log-odds ratio between the evolutionary and random model and can therefore help in determining homology via the alignment score.

• The evolutionary and random models depend on generalized data sets used to derive them. This not an ideal solution.
Take-home messages - 2

• Apart from the PAM and BLOSUM series, a great number of further matrices have been developed.

• Matrices have been made based on DNA, protein structure, information content, etc.

• For local alignment, BLOSUM62 is often superior; for distant (global) alignments, BLOSUM50, GONNET, or (still) PAM250 work well.

• **Remember that gap penalties are always a problem:** unlike the matrices themselves, there is no formal way to calculate their values -- you can follow recommended settings, but these are based on trial and error and not on a formal framework.