There are two principle reasons for comparing and aligning protein sequences:

• To scan a database with a newly determined protein sequence and identify possible functions for the protein by analogy with well characterised proteins.

• To obtain an accurate alignment. This may be for protein modelling by comparison to proteins of known three-dimensional structure

Matrices: An example finding similar sentences

Query: The quick brown fox jumped over the lazy dog

\[ \begin{array}{cccccccc}
\text{The quick brown fox jumped over the lazy dog} & \quad & \quad & \quad & \quad & \quad & \quad & \quad \\
\end{array} \]

Search the database

\[ \begin{array}{cccccccc}
\text{The quick brown fox jumped over the sleeping dog} & \quad & \quad & \quad & \quad & \quad & \quad & \quad \\
\end{array} \]

\[ \begin{array}{cccccccc}
\text{The quick brown fox jumped over the barking dog} & \quad & \quad & \quad & \quad & \quad & \quad & \quad \\
\end{array} \]
### Matrices: Amino acid scoring schemes

<table>
<thead>
<tr>
<th></th>
<th>lazy</th>
<th>sleeping</th>
<th>bark</th>
</tr>
</thead>
<tbody>
<tr>
<td>lazy</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sleeping</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>bark</td>
<td>-5</td>
<td>-10</td>
<td>10</td>
</tr>
</tbody>
</table>

- **Identity scoring**: This is the simplest scoring scheme.
- **Genetic code scoring**: the minimum number of DNA/RNA base changes (0, 1, 2 or 3) that would be required to interconvert the codons for the two amino acids.
- **Chemical similarity scoring**: gives greater weight to the alignment of amino acids with similar physico-chemical properties.
One die: rolled twice
What is the chance of rolling 5 at least once? \(1-(5/6)^2=0.3\)
There are four states

<table>
<thead>
<tr>
<th>State</th>
<th>rol 1</th>
<th>rol 2</th>
<th>P (probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no 5</td>
<td>no 5</td>
<td>(5/6 \times 5/6)</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>no 5</td>
<td>(1/6 \times 5/6)</td>
</tr>
<tr>
<td>3</td>
<td>no 5</td>
<td>5</td>
<td>(5/6 \times 1/6)</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>(1/6 \times 1/6)</td>
</tr>
</tbody>
</table>

Statistics of aligning

When we align two sequences we do this because we want to find out if they are homologous.

One objective way to deal with would be to work with two models and for each model determine the probability that the model is true.

? Maybe we can device a model in which the two amino acids that are aligned have a biological relationship and compare that model with a model in which we have an alignment by chance.
Log odds matrices

\[ S_{ij} = \log \frac{q_{ij}}{p_i p_j} \]

P(bio-match model) vs P(chance meeting model)

If we take log of that we have a scale from + to -
Some calculations:

Suppose we have aligned to protein sequences which are 90% identical and we find an identity match between an amino acid A at position $i$

The probability of “substitution” of $A_i$ by $A_j$ is 0.9

The random model predicts a chance meeting of (you have 20 AA) $0.05 \times 0.05 = 0.0025$

$P_{\text{match}}/P_{\text{random}} = 324 \quad 2\log_{10}360 = 8.5$ (score)

“mismatch” $(0.1/19)/P_{\text{random}} = 2 \rightarrow$ score = 1

Alignment scoring

We want to assign a score to an alignment that gives a measure of the relative likelihood that the two sequences are related as opposed to being unrelated.

. . . .T S A . .
. . . .T T L . .

We can score the ratio of two probabilities $P_{\text{match}}$ (M) vs. $P_{\text{random}}$ (R)

$\log\left( \frac{P(x,y|M)}{P(x,y|R)} \right)$

1. The random model ->
Amino Acids | Codons | Observed Frequency in Vertebrates
---|---|---
Alanine | GCU, GCA, GCC, GCG | 7.4 %
Arginine | CGU, CGA, CGC, CGG, AGA, AGG | 4.2 %
Asparagine | AAU, AAC | 4.4 %
Aspartic Acid | GAU, GAC | 5.9 %
Cysteine | UGU, UGC | 3.3 %
Glutamic Acid | GAA, GAG | 5.8 %
Glutamine | CAA, CAG | 3.7 %
Glycine | GGU, GGA, GGC, GGG | 7.4 %
Histidine | CAU, CAC | 2.9 %
Isoleucine | AUU, AUA, AUC | 3.8 %
Leucine | CUU, CUA, CUC, CUG, UUA, UUG | 7.6 %
Lysine | AAA, AAG | 7.2 %
Methionine | AUG | 1.8 %
Phenylalanine | UUU, UUC | 4.0 %
Proline | CUA, CCA, CCC, CCG | 5.0 %
Serine | UCU, UCA, UCC, UCG, AGU, AGC | 8.1 %
Threonine | ACU, ACA, ACC, ACC, AGC | 6.2 %
Tryptophan | UGG | 1.4 %
Tyrosine | UAU, UAC | 3.3 %
Valine | GUU, GUA, GUC, GUG | 6.8 %
Stop Codons | UAA, UAG, UGA | ---

\[ P(x,y|R) \]

\[ P(x,y|R) \] is different for an alignment of two serines \((0.081)^2\) or two tryptophans \((0.013)^2\) or alanine and proline \((0.074 \times 0.05)\).

\[ P(x,y|M=) \] is a joint probability and in a biological context it is the probability that the two aligned amino acid sequences independently have been derived from a common ancestor \(z\).

In case of identity \(x=y\) and at a small evolutionary distance (two subspecies) it could mean that \(x=z\) and \(y=z\).
In an alignment we score the ratio of two probabilities \( \text{match (M) vs. random (R)} \) 
\( q_{ij} \) depends on the evolutionary distance 
The same identity pair scores lower at a greater evolutionary distance!

**Log odds matrices**

\[
S_{ij} = \log \frac{q_{ij}}{p_ip_j}
\]

Dayhoff matrices

- Dayhoff examined alignments of closely similar sequences where the likelihood of a particular mutation is the result of a set of successive mutations was low.
- Drawback:
  - Since relatively few families were considered, the resulting matrix of accepted point mutations included a large number of entries equal to 0 or 1.
• A complete picture of the mutation process including those amino acids which did not change was determined by calculating the average ratio of the number of changes a particular amino acid type underwent to the total number of amino acids of that type present in the database.

• This is the mutability frequency.

• This was combined with the point mutation data to give the mutation probability matrix where each element gives the probability of the amino acid in column mutating to the amino acid in row after a particular evolutionary time, for example after 2 PAM (Percentage of Acceptable point Mutations per years).

Successive PAMs can be constructed by multiplying this matrix.
The ratios are expressed as logarithms to the base 2 multiplied by 2 (so-called 1/2 bit values) or

\[ 2x \frac{\ln \text{ratio}}{\ln 2} = 2x \log_{2} \text{ratio}. \]

Use of the log-odds (ratio!) matrix provides a prediction of the reliability of a sequence alignment above chance alignment. Using such a matrix, an alignment score of 6 between two amino acids means that the chance of a correct alignment is \(8 \times (2^3)\) times greater than a chance alignment. The probability of alignment between two protein sequences above chance, called the sequence similarity score, is calculated by summing the scores for each aligned amino acid pair and subtracting a penalty score for gaps.

---

**Observed substitutions the Dayhof matrix**

Color codes: Amino acids are grouped according to the chemistry of the side group: C-sulphhydril, STPAG-small hydrophilic, NDEQ-acid, acid amide and hydrophilic, HRK-basic, MILV-small hydrophobic and FYW-aromatic.

|   | C | S | T | P | A | R | L | D | K | R | H | M | I | L | V | F | Y | N |
| C |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| S |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| T |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| P |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| A |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| R |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| L |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| K |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| R |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| H |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| M |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| I |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| L |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| V |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| F |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Y |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

PAM 256 = 80% deviation
The mutation probability matrix is specific for a particular evolutionary distance, but may be used to generate matrices for greater evolutionary distances by multiplying it repeatedly by itself. At the level of 2,000 PAM Schwartz and Dayhoff suggest that all the information present in the matrix has degenerated except that the matrix element for Cys-Cys is 10% higher than would be expected by chance.

However: errors in the original matrix are amplified with greater evolutionary distance.

The logarithm of each element is taken to allow probabilities to be summed over a series of amino acids rather than requiring multiplication. The resulting matrix is the "log-odds matrix" or "Dayhoff's matrix". It is often used at a distance of close to 256 PAM since this lies near to the limit of detection of distant relationships where approximately 80% of the amino acid positions are observed to have changed.

At the evolutionary distance of 256 PAMs one amino acid in five is unchanged but the frequency is different. That is 48% of the tryptophans, 41% of the cysteines, 20% of the histidines and only 7% of serines would be unchanged.
A scoring scheme can be context dependent

That beast has to go, she said pointing at the lazy dog
|    |    |    |    |    |    |    |    | ? |
That beast has to go, she said pointing at the barking dog

The quick brown fox jumped over the lazy dog
|    |    |    |    |    |    |    |
The quick brown fox jumped over the barking dog

The BLOSUM Matrix

First a database of multiple alignments without gaps for short regions of related sequences was derived. The matrix values are based on the observed amino acid substitutions in a large set of approximately 2000 conserved amino acid patterns, called blocks. These blocks have been found in a database of protein sequences representing over 500 groups of related proteins and act as signatures of these protein families.

Each column in the aligned sequences provided a set of possible amino acid substitutions. The types of substitutions were then scored for all aligned patterns in the database.

The idea is that more common substitutions represent a closer relationship between two amino acids in related proteins and thus receive a higher score in sequence alignment. Conversely, rare substitutions should have a low score.
Substitution frequencies for all pairs of amino acids were then calculated between the groups and this used to calculate a log odds BLOSUM (blocks substitution matrix) matrix.

Different matrices are obtained by varying the clustering threshold. Patterns which were 62% identical were grouped together to make the blosum-62 substitution matrix (see figure), and those which were 80% identical were used to make the blosum-80 matrix etc.

**Colour codes:** Amino acids are grouped according to the chemistry of the side group: C-sulfhydryl, STPAG-small hydrophilic, NDEQ-acid, acid amide and hydrophilic, HRK-basic, MILV-small hydrophobic and FYW-aromatic.
A PAM250 matrix where approximately 80% of the amino acid positions are observed to have changed would be the equivalent of a BLOSUM20 matrix!

So which of these matrices should we use?

Altschul

For alignments that do not include gaps he concluded, in broad agreement with Schwarz and Dayhoff, that a matrix of 200 PAMS was most appropriate when the sequences to be compared were thought to be related. However, when comparing sequences that were not known in advance to be related, for example when database scanning, a 120 PAM matrix was the best compromise.

When using a local alignment method Altschul suggests that three matrices should ideally be used: PAM40, PAM120 and PAM250. The lower PAM matrices (40-120) will tend to find short alignments of highly similar sequences, while higher PAM matrices (120-250) will find longer, weaker local alignments.
They conclude that overall the BLOSUM 62 matrix is the most effective.

However, all the substitution matrices investigated perform better than BLOSUM 62 for a proportion of the families. This suggests that no single matrix is the complete answer for all sequence comparisons. It is probably best to compliment the BLOSUM 62 matrix with comparisons using PET91 at 250 PAMS, and Overington structurally derived matrices.

It seems likely that as more protein three-dimensional structures are determined, substitution tables derived from structure comparison will give the most reliable matrices.