Amino acid substitution matrices

- Identity matrix
- Genetic code matrix
- Matrices based on AA properties
- Matrices based on empirical data
  - Dayhoff matrices (PAM120, page 83)
  - BLOSUM matrices (BLOSUM62, page 83)
  - Gonnet matrices (Gonnet 250)
  - JTT matrices
  - and more ...


PAM matrices (Dayhoff et al. 1978)

- Accepted point mutations (point accepted mutations, percent accepted mutations)
  → accepted by selection: no (or very weak) deleterious effect, maintaining the function
- Based on 1,572 changes in 71 groups of closely related proteins (34 protein families)
  → at least 85% identical
  → no ambiguity in alignments, no gap
  → most likely observed substitutions do not affect protein functions (accepted by selection, close to neutral)
  → successive (multiple) substitutions at one site are minimal (no hidden substitution)

See also the textbook pages 119-122

PAM matrices

- Numbers of accepted point mutations: \(f(a,b)\) are counted based on phylogenies
  → Assumption: substitutions are equally likely in each direction (e.g., G→A = A→G)

Phylogeny is reconstructed

Blue-sensitive opsin proteins

Amino acid substitution matrices

<table>
<thead>
<tr>
<th>Aliphatic</th>
<th>Non-polar</th>
<th>Tiny</th>
<th>Small</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polar</td>
<td>Charged</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aromatic</td>
<td>Proline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PAM matrices (Dayhoff et al. 1978)

- Accepted point mutations (point accepted mutations, percent accepted mutations)
  → accepted by selection: no (or very weak) deleterious effect, maintaining the function
- Based on 1,572 changes in 71 groups of closely related proteins (34 protein families)
  → at least 85% identical
  → no ambiguity in alignments, no gap
  → most likely observed substitutions do not affect protein functions (accepted by selection, close to neutral)
  → successive (multiple) substitutions at one site are minimal (no hidden substitution)

See also the textbook pages 119-122
PAM matrices

- Numbers of accepted point mutations: $f(a,b)$ are counted based on phylogenies
- Assumption: substitutions are equally likely in each direction (e.g., $G\rightarrow A = A\rightarrow G$)

Using the maximum parsimony principle, ancestral sequences can be inferred

Substitutions can be identified along the phylogeny

- Numbers of accepted point mutations: $f(a,b)$ are counted based on phylogenies
- Assumption: substitutions are equally likely in each direction (e.g., $G\rightarrow A = A\rightarrow G$)

Using the maximum parsimony principle, ancestral sequences can be inferred

### Relative mutability: $m(a)$

Probability that the amino acid $a$ will change in a given small evolutionary interval

$$m(a) = \frac{f(i,a)}{f(a)}$$

- **Amino acid:** $a$
- **Changes:** $\sum f(i,a)$
- **Freq. of occurrence:** $f(a)$
- **Relative mutability:** $m(a) = \frac{\sum f(i,a)}{f(a)}$

Substitutions are collected from trees with different lengths

Number of times amino acid $a$ is substituted by any other amino acid

$$m(a) = \frac{\sum_{a \neq b} (\text{Freq. of amino acid } a \times \text{Number of total substitutions}) x 100}{f(a) \times \text{(Total number of residues)}}$$

This denominator is called "the total exposure of the amino acid to mutation"
**PAM matrices**

**Relative mutability: \( m(a) \)**

Dayhoff et al. (1978)

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Relative Mutability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn</td>
<td>134</td>
</tr>
<tr>
<td>Ser</td>
<td>120</td>
</tr>
<tr>
<td>Thr</td>
<td>106</td>
</tr>
<tr>
<td>Glu</td>
<td>102</td>
</tr>
<tr>
<td>Ala</td>
<td>100</td>
</tr>
<tr>
<td>His</td>
<td>96</td>
</tr>
<tr>
<td>Lys</td>
<td>94</td>
</tr>
<tr>
<td>Gly</td>
<td>93</td>
</tr>
<tr>
<td>Val</td>
<td>74</td>
</tr>
</tbody>
</table>

*The value for Ala has been arbitrarily set at 100.*

**Mutation probability**

\[
M(a,b) = \lambda m(b) \times f(a,b) / \sum f(a,b), \ \text{where} \ a \neq b
\]

- \( m(b) \): relative mutability of amino acid \( b \)
- \( f(a,b) \): frequency of accepted point mutations between amino acids \( a \) and \( b \)
- \( \sum f(a,b) \): number of times the amino acid \( b \) is substituted by any other amino acid
- \( \lambda \): proportionality constant (normalization factor)

The probability of the amino acid \( b \) being replaced by the amino acid \( a \) after a given evolutionary time

\[
M(b,b) = 1 - \lambda m(b) : \text{unchange probability (the diagonal elements)}
\]

**Relatedness odds score**

\[
R(a,b) = \frac{M(a,b)}{f(a)}
\]

- \( R(a,b) \): the probability that amino acid \( b \) will change to \( a \) in a related sequence in a given interval
- \( f(a) \): the chance of a random occurrence of amino acid \( a \)

**Relatedness odds score matrix** is symmetrical

**Odds ratio**

- \( R(a,b) \): the probability that amino acid \( b \) was derived from \( a \) [nonrandom]
- \( P(a) \): random occurrence of \( a \)

**Mutation probability matrix**

Dayhoff et al. (1978)

Ala will not be changed with 98.67% probability

Ala will be replaced by Gly with 0.21% probability

**Mutation probability matrix is not symmetrical**
**PAM matrices**

- Relatedness odds score
  \[ R(a,b) = \frac{M(a,b)}{f(a)} \]
  where \( M(a,b) \) is the probability that amino acid \( b \) will change to \( a \) in a related sequence in a given interval.
  \( f(a) \) is the chance of a random occurrence of amino acid \( a \).

- Log odds score
  \[ S(a,b) = 10 \log_{10} \left( \frac{M(a,b)}{f(a)} \right) \]

- Odds ratio
  \[ \frac{R(a,b)}{R(b,a)} = \frac{M(a,b)}{f(a)} \times \frac{f(b)}{M(b,b)} \]
  where \( M(b,b) = 1 - \sum f(b)M(b,b) \), the probability of AA replaced by AA, after the evolutionary interval of PAM1 (when one mutation per 100 aa is found).

- Mutation probability matrix
  \[ M_n = (M_1)^n \]
  for PAMn.

**PAM matrices updated**

- JTT matrices
  by Jones, Taylor, and Thornton (1992)
  - 59,190 accepted point mutations for 16,300 proteins

- Gonnet matrices
  by Gonnet, Cohen, Benner (1992)
  - Based on exhaustive pairwise alignment from the protein database (~8,344,353 amino acids).
BLOSUM matrices (Henikoff and Henikoff 1992)

- Blocks substitution matrix
- Based on ~2,000 conserved amino acid patterns (or ungapped blocks), representing more than 500 families.
- Based on local, multiple alignment of all commonly-occurring motifs (blocks) in the protein sequence database.

- The Blocks Database
  http://blocks.fhcrc.org/ (for reference only, no longer updated)

BLOCK entry example

<table>
<thead>
<tr>
<th>Seq1</th>
<th>ACL</th>
<th>Seq2</th>
<th>ACV</th>
<th>Seq3</th>
<th>TCV</th>
<th>Seq4</th>
<th>MAI</th>
<th>Seq5</th>
<th>MAI</th>
</tr>
</thead>
</table>

Observed amino acid pairs:

- Predicted amino acid pairs:

  Observed frequencies of amino acid pairs:

  - Observed frequency of each amino acid in the pairs:
    \[ p_i = \frac{q_{ii} + \sum_j q_{ij}}{2} \]
    \[ p_A = 0.6 \]
    \[ p_C = 0.4 \]

  - Observed frequency of each amino acid in the pairs:
    \[ p_A = 0.2 \]
    \[ p_C = 0.8 \]

  - Expected frequencies of amino acid pairs:
    \[ e_{CC} = p_A^2 = 0.36 \]
    \[ e_{CA} = 2 \times p_A \times p_C = 0.32 \]

Log odds scores are calculated for each amino acid pairs:

\[ S_{ij} = \log \left( \frac{q_{ij}}{e_{ij}} \right) \]

In bit units: \[ S_{ij} = \log_2 (q_{ij}/e_{ij}) \]

Usually in half-bit units: \[ S_{ij} = 2 \log_2 (q_{ij}/e_{ij}) \]

*bit = binary digit (0 or 1)

BLOSUM matrix

- Observed and Expected frequencies of amino acid pairs are cumulatively counted from all columns of the BLOCKs

- Log odds scores are calculated for each amino acid pairs:

  \[ S_{ij} = \log \left( \frac{q_{ij}}{e_{ij}} \right) \]

  In bit units: \[ S_{ij} = \log_2 (q_{ij}/e_{ij}) \]

  Usually in half-bit units: \[ S_{ij} = 2 \log_2 (q_{ij}/e_{ij}) \]
BLOSUM matrices

- BLOSUMₙ represents the similarity threshold (e.g., BLOSUM62, BLOSUM45, BLOSUM80)
  - for any n, the corresponding BLOSUM matrix is generated mainly comparing sequences that are less than n% identical
e.g., BLOSUM62: Sequences with ≥62% identity are clustered and treated as a single sequence for counting.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins

BLOSUM and PAM matrices

- PAM matrices: based on a transition probability matrix for a Markov process
  - any Mᵣ matrix can be extrapolated based on PAM1 (M₁) matrix (e.g., M₆₂ = M₁ to the 62 power)
  - assume more distant changes are a reflection of the short-term changes designed to track the evolutionary origins of proteins
- BLOSUM matrices: not based on explicit evolutionary model
  - derived from all changes observed in the conserved blocks regardless of the overall degree of similarity
  - generated based on different similarity levels (BLOSUM50, BLOSUM62, etc.)
  - all BLOSUM matrices are generated based on observed data
  - designed to find conserved domains

Log Odds Matrix

- Log odds (Lod) score: PAM matrix
  \[ S(i,j) = 10 \log_{10} \left( \frac{M(i,j) / (f(i) \times f(j))}{m(j) \times f(i,j) / f(j)} \right) \]
- PAM matrix is symmetrical: \( S(i,j) = S(j,i) \)

Log Odds Matrix (PAM250)
Log Odds Matrix

- **Log odds (Lod) score:** BLOSUM matrix
  \[ S(i,j) = 2 \log_2 \left( \frac{q_{ij}}{e_{ij}} \right) \]
  - \( q_{ij} \): Observed frequency of (AA_i, AA_j) pairs
  - \( e_{ij} \): Expected frequencies of (AA_i, AA_j) pairs
  \[ e_{ii} = 2p_i^2 \text{ and } e_{ij} = 2p_i p_j \]
  - \( p_i \): Observed frequency of AA_i in the pairs
  \[ p_i = q_{ii} + \sum_{j \neq i} q_{ij} / 2 \]
- **BLOSUM matrix is symmetrical:**\( S(i,j) = S(j,i) \)

Log Odds Matrix (BLOSUM62)

- **Log odds (Lod) score:** general
  - also called log odds ratio or log likelihood ratio
  \[ S(i,j) = \frac{1}{\#} \log_2 \left( \frac{q_{ij}}{e_{ij}} \right) \]
  \[ S(i,j) = \frac{1}{\#} \log e(q_{ij} / p_i p_j) \]
  - \( q_{ij} \): the frequency of the amino acid pair, AA_i and AA_j
  - \( p_i, p_j \): the individual frequency of AA_i or AA_j
  - \# : a scaling factor
  \( 1/2 \) is used with BLOSUM

- **Likelihood ratio:**
  \[ \text{Likelihood ratio} = \log \left( \frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0} \right) \]
  - \( H_1 \): Hypothesis to be tested, \( H_0 \): Null hypothesis
  \ [- < \text{log(LR)} < +] \]