Pairwise sequence alignment
and alignment scoring

Analysis of Biological Sequences 140.638
Scoring rules/matrices

• Why are they important?
  • Choice of scoring rule can dramatically influence the sequence alignments obtained and, therefore, the analysis being done
  • Different scoring matrices have been developed for different situations; using the wrong one can make a big difference (choosing the wrong sequence as a potential functional ortholog, for example)
Scoring rules/matrices

• What do they mean?
  • Your goal is to figure out whether the two sequences have a common ancestor
  • Scoring matrices implicitly represent a particular theory of evolution
  • Elements of the matrices specify relationships between amino acid residues or nucleotides
Substitution Matrices

• We need scoring terms for each aligned residue pair

• Models: Random model (R): letter a occurs with frequency $q_a$

\[
\begin{array}{cccc}
    x_1 & x_2 & x_3 & x_4 \\
    \hline
    y_1 & y_2 & y_3 & y_4
\end{array}
\]

\[
P(x,y|R) = \prod_i q_{x_i} q_{y_i}
\]
Substitution Matrices—random model

$$P(x,y|R) = \prod_i q_{x_i} q_{y_i}$$

\[x = \text{ACCTGCC} \quad p(A) = 0.2\]
\[y = \text{ACGTCCA} \quad p(T) = 0.2\]
\[\text{ACCTGCC} \quad p(C) = 0.3\]
\[\text{| | | | |} \quad p(G) = 0.3\]
\[\text{ACGTCCA}\]

\[
P(x,y|R) = p(A)^2 p(C)^2 p(C) p(G) p(T)^2 p(G) p(C) p(C)^2 p(C) p(A) = 6.29 \times 10^{-9}
\]
Substitution Matrices—match model

• Models: Match model (M): aligned pairs of residues have joint probability $p_{ab}$

• $p_{ab} =$ probability that $a$ and $b$ came from common ancestor residue

\[
P(x, y | M) = \prod_{i} p_{x_i y_i}
\]
Substitution Matrices

Odds ratio:

\[
\frac{P(x, y|M)}{P(x, y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i q_{y_i}}}
\]

\[
= \prod_i \left( \frac{p_{x_i y_i}}{q_{x_i q_{y_i}}} \right)
\]
Substitution Matrices

\[ \prod_i \left( \frac{px_i y_i}{qx_i qy_i} \right) \]

Change to a sum by using logarithms . . .

Score = \[ \sum_i \log \left( \frac{px_i y_i}{qx_i qy_i} \right) = \sum_i s(a, b)_i \]

Where s(a,b) is just the score of aligning a residue of type a to a residue of type b
Substitution matrices

Log-odds ratio -> log likelihood ratio that the pair (a,b) is related vs unrelated (depends on scoring matrices)

The alignment score is the log likelihood that the sequences have common ancestry
Two major scoring matrices

- PAM = accepted point mutation
  - 71 trees with 1572 accepted mutations, sequences with >85% identity
  - “accepted” means new amino acid doesn’t disrupt the protein’s function too severely
  - PAM1 means average of 1% change over all amino acids
  - 1 PAM = 10my evolutionary distance
Two major scoring matrices

- **BLOSUM** = Blocks substitution matrices
  - Based on BLOCKS database (Henikoff & Henikoff, 1992) of over 2000 conserved amino acid patterns in over 500 proteins
PAM overview

- based on well-accepted phylogenetic trees

observations: one S/T change between close relatives, one P/Q change over distant branches, no change from C
BLOSUM overview

- based on alignments of known protein motifs, evolutionary relationship unknown

STTWC
SSTWC
STTPC
STTWC

observations: three T/S mismatches, three P/Q mismatches, no change from C
PAM matrices

• Each matrix describes changes expected for a given period of evolutionary time (measured by expected similarity of proteins)

• Count # of changes to each amino acid in the phylogenetic group and divide by the “exposure to mutation” of the residue

• Exposure to mutation = frequency of occurrence of amino acid * #amino acid changes in the group/100 sites
PAM matrices

• Amino acid changes are modeled by a Markov process, so each mutation is independent of previous mutations and of adjacent positions

• This means that we can calculate the matrices for more distantly related proteins by multiplying matrices for closely related proteins (PAM 250 = PAM1 multiplied by itself 250 times)

• PAM 250 = 250% change over 2500 my. ~20% similarity at this level; shown to be best for proteins of 14-27% similarity
PAM matrices

- PAM120: 40% similarity
- PAM80: 50% similarity
- PAM60: 60% similarity

- Simulations have confirmed these numbers

- Choosing the best PAM matrix: ungapped alignment score will be highest when the correct matrix is used.
PAM matrices—assumptions

- $P(X \rightarrow Y) = P(Y \rightarrow X)$
- $P(X \rightarrow Z \rightarrow Y)$ is low in a single PAM period
- Markov model/independence
- All sequences have similar amino acid composition
An example: obtaining the PAM250 score for Tyr <-> Phe (from Mount book)

Original PAM data: 1572 observed amino acid changes, 260 were between Phe and Tyr

Mutation probability score: observed changes vs expected changes

- Relative mutability = likelihood that the amino acid will change
- Pair exchange frequency = fraction of Phe->Tyr/all Phe mutations
- Normalize to a sum of 1% probability of any amino acid change, then take log odds -> PAM1
An example: obtaining the PAM250 score for Tyr <-> Phe (from Mount book)

Original PAM data: 1572 observed amino acid changes, 260 were between Phe and Tyr

\[
\frac{260}{1572} \times \frac{\text{# times F changes}}{\text{# times F stays F}} \times \frac{\text{# times F mutates to Y}}{\text{# times F changes}}
\]

Normalize to a sum of 1% probability of any amino acid change, then take log odds -> PAM1

Multiply PAM1 by itself 250 times to get PAM250
BLOSUM

• Henikoff & Henikoff used PROTOMAT program to create BLOCKS database from Prosite catalog

• PROTOMAT looks for $A_1-d_1-A_2-d_2-A_3$ where $A_1$, $A_2$, $A_3$ are conserved residues and $d_1, d_2 < 25$ residue intervening sequence
# BLOSUM

- These initial patterns are consolidated into larger patterns by PROTOMAT

- Sequences above a similarity threshold are clustered into families (% threshold -> BLOSUM #)

<table>
<thead>
<tr>
<th>GHHPL</th>
<th>YCSSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHHPA</td>
<td>consolidate</td>
</tr>
<tr>
<td>GHHPL</td>
<td>↓</td>
</tr>
<tr>
<td>GHHPT</td>
<td>GHHPAYCSTW</td>
</tr>
<tr>
<td></td>
<td>GHHPLYCSTW</td>
</tr>
<tr>
<td></td>
<td>GHHPLYCSLW</td>
</tr>
<tr>
<td></td>
<td>GHHPTYCLLW</td>
</tr>
<tr>
<td></td>
<td>YCLLW</td>
</tr>
</tbody>
</table>
BLOSUM construction

1. Count mutations

<table>
<thead>
<tr>
<th>Sequence</th>
<th>N&lt;sub&gt;AA&lt;/sub&gt;</th>
<th>N&lt;sub&gt;VV&lt;/sub&gt;</th>
<th>N&lt;sub&gt;PP&lt;/sub&gt;</th>
<th>N&lt;sub&gt;AV&lt;/sub&gt;</th>
<th>N&lt;sub&gt;AP&lt;/sub&gt;</th>
<th>N&lt;sub&gt;VP&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>VVAPV</td>
<td>0+1+6+0+0</td>
<td>0+1+0+0+3</td>
<td>1+0+0+3+0</td>
<td>1+4+0+0+3</td>
<td>2+0+0+3+0</td>
<td>2+0+0+0+0+0</td>
</tr>
<tr>
<td>AAAPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVAPV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAAAV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BLOSUM construction

2. Tallying mutation frequencies

We don’t know ancestry, so each mutation gets entered twice

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>V</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>14</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>V</td>
<td>8</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>P</td>
<td>5</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VVAPV</td>
</tr>
<tr>
<td>AAAPA</td>
</tr>
<tr>
<td>PVAPV</td>
</tr>
<tr>
<td>PAAAV</td>
</tr>
</tbody>
</table>
BLOSUM construction
3. Matrix of mutation probabilities

Create probabilities from mutation frequencies:
60 events observed (4 sequences means 6 possible pairs per column, multiply by 5 columns, and every pair counts twice)

<table>
<thead>
<tr>
<th></th>
<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>
<td></td>
</tr>
<tr>
<td>V</td>
<td>8/60</td>
<td>8/60</td>
<td>2/60</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>5/60</td>
<td>2/60</td>
<td>8/60</td>
<td></td>
</tr>
</tbody>
</table>
BLOSUM construction

4. Calculate abundance of each residue
(Marginal probability)

$q_i$ is the marginal probability, meaning the expected probability of occurrence of amino acid $i$

<table>
<thead>
<tr>
<th>$p_{ij}$</th>
<th>A</th>
<th>V</th>
<th>P</th>
<th>$q_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>14/60</td>
<td>8/60</td>
<td>5/60</td>
<td>27/60</td>
</tr>
<tr>
<td>V</td>
<td>8/60</td>
<td>8/60</td>
<td>2/60</td>
<td>18/60</td>
</tr>
<tr>
<td>P</td>
<td>5/60</td>
<td>2/60</td>
<td>8/60</td>
<td>15/60</td>
</tr>
</tbody>
</table>

if $q_i \times q_j = p_{ij}$, the two amino acids are independent.
BLOSUM construction
5. Obtaining a BLOSUM matrix

- **BLOSUM** is a log-likelihood matrix:

\[
S_{ij} = 2\log_2\left(\frac{p_{ij}}{(q_iq_j)}\right)
\]

<table>
<thead>
<tr>
<th>S_{ij}</th>
<th>A</th>
<th>V</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>-0.036</td>
<td>1.134</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-0.866</td>
<td>-2.34</td>
<td>2.19</td>
</tr>
</tbody>
</table>
• BLOSUM: based on short conserved sequences (blocks)
  • Based on a range of evolutionary periods
  • Each matrix constructed separately
  • Indirectly accounts for interdependence of residues
  • Range of sequences, range of replacements
  • Overcounts related mutations

• SCORING MATRIX

• PAM: evolutionary model
  • Based on extrapolation from a short evolutionary period
  • Errors in PAM1 are magnified through PAM250
  • Assumes Markov process
  • Many sequences depart from average composition
  • Rare replacements too infrequent to be represented accurately

• SUBSTITUTION MATRIX
Issues

• Both BLOSUM and PAM matrices are derived from small sets of sequences from biased databases
• Both types of matrices require aligned sequences for their construction
• Both types of matrices depend on global, ungapped alignments