Substitution matrices

Sequence analysis 2007

Lecture 4 - 10/12/07
Definition

• Two-dimensional matrix with score values describing the probability of one amino acid or nucleotide being replaced by another during sequence evolution.
Scoring matrices for nucleotide sequences

• Can be simple:
  • e.g. positive value for match and zero for mismatch.
  • frequencies of mutation are equal for all bases.

• 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**

**Diagrams of purines and pyrimidines**

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.

   ![Genetic sequence](image)
So?

Rule of thumb:

⇒ if ORF exists,
then align at protein level
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)
Scoring matrices for amino acid sequences
Example: Cysteines are very common in metal binding motifs
Now let’s think about alignments

- Lets 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.

\[
\frac{\Pr(x, y \mid M)}{\Pr(x, y \mid R)} \quad \leftrightarrow \quad \text{sequences have common ancestor}
\]
\[
\leftrightarrow \quad \text{sequences are aligned by chance}
\]

- 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 is given by:

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

- Target frequencies
  - The “evolutionary true” alignments allow us to get 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 frequency 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

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

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

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

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

- 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 freq. is equal to that expected by chance

- a negative score
  … that the freq. is less to that expected by chance
Alignment score

- The alignment score $S$ is given by the sum of all amino acid pair substitution scores:

$$S = \sum_i s(x_i, y_i) = \log \frac{\Pr(x, y | M)}{\Pr(x, y | R)}$$

- 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:

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

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

\[
s(a, b \mid t) = \log \frac{p_{ab}}{q_a q_b} = \log \frac{P(b \mid a, t)}{q_b}
\]
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.
But there is a whole series of matrices: **PAM10 ... PAM250**

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

\[
\begin{array}{c}
\text{X} \times \text{X} \times \text{X} = \text{PAM}\text{X}
\end{array}
\]

Multiply Matrices \(N\) times to make PAM “\(X\)”; 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 = 2.5 accepted mutations per amino acid
# PAM numbers vs. observed mutation 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

\[
\begin{array}{cccccccccccccccccc}
A & 2 \\
R & -2 & 6 \\
N & 0 & 0 & 2 \\
D & 0 & -1 & 2 & 4 \\
C & -2 & -4 & -4 & -5 & 12 \\
Q & 0 & 1 & 1 & 2 & -5 & 4 \\
E & 0 & -1 & 1 & 3 & -5 & 2 & 4 \\
G & 1 & -3 & 0 & 1 & -3 & -1 & 0 & 5 \\
H & -1 & 2 & 2 & 1 & -3 & 3 & 1 & -2 & 6 \\
I & -1 & -2 & -2 & -2 & -2 & -3 & -2 & 5 \\
L & -2 & -3 & -4 & -6 & -2 & -3 & -4 & -2 & 5 & 6 \\
K & -1 & 3 & 1 & 0 & -5 & 1 & 0 & -2 & 0 & -2 & -3 & 5 \\
M & -1 & 0 & -2 & -3 & -5 & -1 & -2 & -3 & 2 & 4 & 0 & 6 \\
F & -4 & -4 & -6 & -4 & -5 & -5 & -5 & -2 & 1 & 2 & -5 & 0 & 9 \\
P & 1 & 0 & -1 & -1 & -1 & 0 & -1 & 0 & -3 & -1 & -2 & -5 & 6 \\
S & 1 & 0 & 1 & 0 & 0 & -1 & 0 & 1 & -1 & -1 & -3 & 0 & -2 & -3 & 1 & 2 \\
T & 1 & -1 & 0 & 0 & -2 & -1 & 0 & 0 & -1 & 0 & -2 & 0 & -1 & -3 & 0 & 1 & 3 \\
Y & -3 & -4 & -2 & -4 & 0 & -4 & -4 & -5 & 0 & -1 & -1 & -4 & -2 & 7 & -5 & -3 & -3 & 0 & 10 \\
B & 0 & -1 & 2 & 3 & -4 & 1 & 2 & 0 & 1 & -2 & -3 & 1 & -2 & -5 & -1 & 0 & 0 & 5 & -3 & -2 & 2 \\
Z & 0 & 0 & 1 & 3 & -5 & 3 & 3 & -1 & 2 & -2 & -3 & 0 & -2 & -5 & 0 & 0 & -1 & -6 & -4 & -2 & 2 & 3
\end{array}
\]

- R exchange is too large (due to paucity of data)
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.

• 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.
- All matrices are directly calculated; no extrapolations are used.
- Again: 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.

• The database can be searched to classify protein and nucleotide sequences.
The Blocks database

Gapless alignment blocks
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 (all blocks have \( \geq 62\% \) sequence identity);
- No extrapolations are made in going to higher evolutionary distances
- High number - closely related sequences
  - Low number - distant sequences
- **BLOSUM62** is the most popular: best for general alignment.
The log-odds matrix for **BLOSUM62**

|     | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | X |
| A   | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| R   | -1| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N   | -2| 0 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D   | -2| 1 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| C   | 0 | -3| -3| 9 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Q   | -1| 1 | 0 | -3| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| E   | -1| 0 | 0 | -4| 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| G   | 0 | -2| 0 | 1 | -3| -2| 2 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| H   | -2| 0 | 1 | -3| 3 | 0 | 0 | -2| 8 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| I   | -1| -3| -3| -1| -3| -3| -4| -3|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 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| -2| -3| -3| -3| -3| 1 | 0 | 0 | -3| 0 | 6  |   |   |   |   |   |   |   |   |
| P   | -1| -2| -2| -1| -3| 1 | 1 | -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| -2| -2| -1| -1| -1| -2| -1| 1 | 5  |   |   |   |   |   |   |   |
| W   | -3| -3| -4| -3| -2| -3| -4| -4| -3| 1 | 4 | -3| -3| -2| 7  |   |   |   |   |   |   |   |
| Y   | -2| -2| -2| -3| -2| -2| -3| -2| -3| -2| -3| -4| -3| -2| 11  |   |   |   |   |   |   |   |
| V   | 0 | -3| -3| -3| -2| -2| -3| -3| -3| -3| -2| -2| -2| 0 | -3| 1 | 4  |   |   |   |   |
| X   | 0 | -1| -1| -2| -1| -1| -1| -1| -1| -1| -1| -1| -1| 0 | 0 | -2| -1| -1| -1|   |   |

- **Positive for chemically similar substitution**
- **Common amino acids have low weights**
- **Rare amino acids have high weights**
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.

\[ H = \sum s_{ij} \log_2 \left( \frac{s_{ij}}{p_i p_j} \right) \]

<table>
<thead>
<tr>
<th>NAME</th>
<th>CLUSTER</th>
<th>E</th>
<th>#blocks</th>
<th>#PAIRS</th>
<th>ENTROPY</th>
<th>EXPECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blosum30</td>
<td>30°</td>
<td></td>
<td>261</td>
<td>76,394</td>
<td>.1424</td>
<td>-.1074</td>
</tr>
<tr>
<td>Pan350</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.20</td>
<td></td>
</tr>
<tr>
<td>Pan340</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>Blosum35</td>
<td>35°</td>
<td></td>
<td>439</td>
<td>136,328</td>
<td>.2111</td>
<td>-.1560</td>
</tr>
<tr>
<td>Pan330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.22</td>
<td></td>
</tr>
<tr>
<td>Pan290</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>Blosum40</td>
<td>40°</td>
<td></td>
<td>672</td>
<td>227,179</td>
<td>.2851</td>
<td>-.2090</td>
</tr>
<tr>
<td>Pan280</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>Pan250a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.36</td>
<td>-.28</td>
</tr>
<tr>
<td>Blosum45</td>
<td>45°</td>
<td></td>
<td>900</td>
<td>357,558</td>
<td>.3795</td>
<td>-.2879</td>
</tr>
<tr>
<td>Pan240</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>Pan210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.48</td>
<td></td>
</tr>
<tr>
<td>Blosum50</td>
<td>50°</td>
<td></td>
<td>1161</td>
<td>557,053</td>
<td>.4808</td>
<td>-.3573</td>
</tr>
<tr>
<td>Pan200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.51</td>
<td></td>
</tr>
<tr>
<td>Pan190</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.55</td>
<td></td>
</tr>
<tr>
<td>Blosum55</td>
<td>55°</td>
<td></td>
<td>1331</td>
<td>753,984</td>
<td>.5637</td>
<td>-.4179</td>
</tr>
<tr>
<td>Pan180</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Pan170</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>Blosum60</td>
<td>60°</td>
<td></td>
<td>1529</td>
<td>1,087,140</td>
<td>.6603</td>
<td>-.4917</td>
</tr>
<tr>
<td>Blosum62</td>
<td>62°</td>
<td></td>
<td>1572</td>
<td>1,245,852</td>
<td>.6979</td>
<td>-.5809</td>
</tr>
</tbody>
</table>

Some considerations

• Did you notice that both PAM and BLOSUM matrices have rounded values?

• Apparent contradictions: A/L = -1
  K/E = +1
Evolution and Matrix “landscape”

- Recent evolution → identity matrix
- Ancient evolution → convergence to random model
Specialized matrices

- Several other aa exchange matrices have been constructed, for situations in which non-standard amino acid frequencies occur.

Specialized matrices

- **Transmembrane specific** substitution matrices:
  - **PHAT** (Ng P, Henikoff J, Henikoff S, Bioinformatics 2000;16(9):760-766)
    *Built from predicted hydrophobic and transmembrane regions of the blocks database*
  - **BATMAS** (Sutormin RA, Rakhmaninova AB, Gelfand S, Proteins 2003; 51(1):85-95)
    *Derived from predicted TM-kernels of bacterial proteins*
A note on reliability

- All these matrices are designed using standard 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.
• 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.
Pair-wise alignment quality versus sequence identity

- Vogt et al., JMB 249, 816-831, 1995
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 the generalized data 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.
Further reading

• Nature Biotechnology 2004, vol. 22 (8)

Where did the BLOSUM62 alignment score matrix come from?

Sean R Eddy

Many sequence alignment programs use the BLOSUM62 score matrix to score pairs of aligned residues. Where did BLOSUM62 come from?