Sequence Alignments

Dynamic programming approaches, scoring, and significance

Lucy Skrabanek
ICB, WMC
January 31, 2013
Sequence alignment

• Compare two (or more) sequences to:
  – Find regions of conservation (or variability)
  – Common motifs
  – Find similar sequences in a database
  – Assess likelihood of common evolutionary history

• Once we have constructed the best alignment between two sequences, we can assess how similar they are to each other.
Methods of sequence alignment

• Dot-matrix plots

• Dynamic programming algorithms (DPA)
  – Used to determine evolutionary homology
  – Needleman-Wunsch (global)
  – Smith-Waterman (local)

• Heuristic methods
  – Used to determine similarity to sequences in a database
  – BLAST

• Combination of DPA and heuristic
  – (FASTA)
Dot-matrix plot

Window: 15 Unity matrix Score = 8 9 10 >= 11

X: 292 1ALPHA_BLE LSKQAQAYRQMSLLRPPGCREAYPGDFVYLHSRLLERAAKLNSLLCEGSMTALPIV
Y: 303 1AF1_SCERE LSKASLAYRQLSLMLRRPPGCREAYPGDFVYLHSRLLERAAKLSEKECSGSLTALPVI
Dynamic programming

• Bellman, 1955

• Tabular design technique
  – Partition the problem into sub-problems
  – Solve sub-problems recursively
  – Combine solutions to solve the original problem
  – Applicable when the sub-problems are not independent
Dynamic programming

• **Sub-problems** are computed once and saved in a table or matrix.
  – Table is built from the bottom up

• The basic ideas behind dynamic programming are:
  – The **organization of work** in order to avoid repetition of work already done.
  – Breaking a large problem into smaller incremental steps, which are each solved optimally.

• **DPAs** are typically used when a problem has **many possible** solutions and an **optimal** one has to be found.
  – Hallmark of DPA: An optimal solution to a problem contains optimal solutions to sub-problems

January 31, 2013

Luce Skrabanek
Example alignment

ELVIS LIVES?
Scoring alignments

• Associate a score with every possible alignment
  – A higher score indicates a “better” alignment

• Need a scoring scheme that will give us
  a. Values for residue substitution
  b. Penalties for introducing gaps

• The score for each alignment is then the sum of all the residue substitution scores minus any penalties for gaps introduced
# BLOSUM62 substitution matrix

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

January 31, 2013

Luce Skrabanek
N-W mathematical formulation

Global alignment: used when the sequences are of a comparable length, and show similarity along their whole lengths

\[
F(i,j) = \text{MAX} \left\{ \begin{array}{l}
F(i-1, j-1) + s(x_i, y_j), \\
F(i-1, j) - d, \\
F(i, j-1) - d
\end{array} \right. 
\]

where \( F(i,j) \) is the value in cell \((i,j)\); \( s \) is the score for that match in the substitution matrix; \( d \) is the gap penalty
## Simple global alignment

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>L</th>
<th>V</th>
<th>I</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>0</td>
<td>-2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
# Traceback step

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>L</th>
<th>V</th>
<th>I</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L</strong></td>
<td>0</td>
<td>-2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>I</strong></td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td><strong>V</strong></td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td><strong>10</strong></td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

January 31, 2013
Luce Skrabanek
Resulting alignment

Not correct to just align the highest number of identical residues!

EL-VIS
- LIVES

Acidic

Score:
0+4+3+3-2+4
= 12

ELVI-S
- LIVES

Hydrophobic

Glutamic acid (E) [glutamate]

Isoleucine (I)

Valine (V)

January 31, 2013
Luce Skrabanek
Local alignment

• Differences from global alignments
  – Minimum value allowed = 0
    • Corresponds to starting a new alignment
  – The alignment can end anywhere in the matrix, not just the bottom right corner
    • To find the best alignment, find the best score and trace back from there
  – Expected score for a random match MUST be negative
S-W local alignment formula

Local alignment: used when the sequences are of very different lengths, or when they show only short patches of similarity

\[ F(i,j) = \text{MAX} \{ \]
\[ 0, \]
\[ F(i-1, j-1) + s(x_i, y_j), \]
\[ F(i-1, j) - d, \]
\[ F(i, j-1) - d \] \}

January 31, 2013

Luce Skrabane
<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>M</th>
<th>A</th>
<th>L</th>
<th>I</th>
<th>S</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

January 31, 2013

Luce Skrabanek
<table>
<thead>
<tr>
<th>S</th>
<th>M</th>
<th>A</th>
<th>L</th>
<th>I</th>
<th>S</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

January 31, 2013

Luce Skrabanek
Resulting alignments

ALLISH
AL–IGN

Score:
$4+4-2+4+0+1$
$=11$

ALLISH
AL–IGN

Score:
$4-2+4+4+0+1$
$=11$

January 31, 2013
Luce Skrabanek
Alignment with affine gap scores

• **Affine gap scores**
  – “Affine” means that the penalty for a gap is computed as a linear function of its length
  – Gap **opening** penalty
  – Less prohibitive gap **extension** penalty

• Have to keep track of multiple values for each pair of residue coefficients \( i \) and \( j \) in place of the single value \( F(i,j) \)
  – We keep two matrices, \( M \) and \( I \)

• **Time taken**: increases to \( O(NM^2) \)
Affine gap score formula

\[ M(i,j) = \text{MAX} \{ \]
\[ M(i-1, j-1) + s(x_i, y_j), \]
\[ I(i-1, j-1) + s(x_i, y_j) \}
\]

\[ I(i,j) = \text{MAX} \{ \]
\[ M(i, j-1) - d, \]
\[ I(i, j-1) - e, \]
\[ M(i-1, j) - d, \]
\[ I(i-1, j) - e \} \]
Improvements

• Gotoh - 1982
  – Simplified algorithm to improve affine gapped alignment performance (from $O=NM^2$ to $O=NM$)
    • Although space required: $3xMN$
  – Used linear relationship for gap weight:
    • $wx = g + rx$

• Myers and Miller - 1988
  – Divide-and-conquer
  – based on Hirschberg, 1975
Conclusions

• DPA mathematically proven to yield the optimal alignment between any two given sequences.

• Final alignment depends on \textit{scoring matrix} for similar residues and \textit{gap penalties}.
Gap penalty options

• Depends on task
  – Maximizing quality of a pairwise alignment
    • Lower gap penalties (biased towards allowing complete global alignment)
  – Maximizing sensitivity, searching a database
    • Higher gap penalties to maximize the difference between scores for homologous sequences and unrelated sequences

BLOSUM62 11, 1 / 5,3
BLOSUM80 10, 1 / 9,3
BLOSUM45 15,2

PAM30 9,1
PAM 70 10,1
PAM 250 14,2

Reese and Pearson, Bioinformatics, 2002
Vogt et al., JMB, 1995
BLAST - heuristic strategy

• Basic Local Alignment Search Tool
• Break the query and all database sequences into “words”
• Look for matches that have at least a given score, given a particular scoring matrix
  → Seed alignment
• Extend seed alignment in both directions to find an alignment with a score or an E-value above a threshold
  – Refinements: use DPAs for extension; 2-hit strategy
• The word size and threshold score determines the speed of the search
BLAST alignment strategy

1. Break sequences up into “words”

2. Find all matching “words” above a score threshold

3. Locate any perfect matches in the other set of “words”

4. Align and extend, with gaps if necessary

Use original “word”!

ELVIS

LVI = 4+4+4
LVV = 4+4+3
LII = 4+3+4
LVL = 4+4+2
LLI = 4+2+4
LIV = 4+3+3

LIVES

eLVI-s
-LIVes
PSI-BLAST (Position-Specific Iterated) BLAST

- Uses a PSSM (position-specific scoring matrix) derived from BLAST results
  - Generated by calculating position-specific scores for each position in the alignment
    - Highly conserved positions receive high scores
    - Weakly conserved positions receive low scores
  - PSSM used as the scoring matrix for the next BLAST search
- Iterative searching leads to increased sensitivity
Scoring matrices

• Strongly influence the outcome of the sequence comparison
• Implicitly represent a particular theory of evolution

Frommlet et al, Comp Stat and Data Analysis, 2006
Examples of scoring matrices

- Identity
- BLOSUM
- PAM
- Gonnet (also called GCB - Gonnet, Cohen and Benner)
- JTT (Jones Thornton Taylor)
- Distance matrix (minimum number of mutations needed to convert one residue to another)
  
  e.g., $I \rightarrow V = 1$ ($ATT \rightarrow GTT$)
- Physicochemical characteristics
  
  e.g., hydrophobicity scoring (George et al., 1990)
# DNA matrices

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

## Transition/transversion

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>-5</td>
<td>-5</td>
<td>-1</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>0</td>
<td>-1</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td>-1</td>
<td>0</td>
<td>-5</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>-5</td>
<td>-5</td>
<td>0</td>
</tr>
</tbody>
</table>

## WU-BLAST

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>-4</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
</tr>
<tr>
<td>G</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
<td>5</td>
</tr>
</tbody>
</table>

### Purines: A, G

### Pyrimidines: C, T

### Transition: A -> G

### Transversion: C -> G
PAM matrices

• PAM (Percent/Point Accepted Mutation)
  – One PAM unit corresponds to one amino acid change per 100 residues, or 0.01 changes/residue, or roughly 1% divergence
    • PAM30: 0.3 changes/residue
    • PAM250: 2.5 substitutions/residue
  – Each matrix gives the changes expected for a given period of evolutionary time
    • Frequency of replacement of one amino acid by another
  – Assumes that amino acid frequencies remain constant over time and that substitutions observed over short periods of evolutionary history can be extrapolated to longer distances
  – Uses a Markov model of evolution
Markov model of evolution

• Assumptions
  – Each amino acid site in a protein can change to any of the other amino acids with the probabilities given in the matrix
  – Changes that occur at each site are independent of one another
  – The probability of an amino acid change is the same regardless of its position in the protein
  – The probability of mutation is history-independent
PAM matrix construction

• Align sequences
• Infer the ancestral sequence
• Count the number of substitution events
• Estimate the propensity of an amino acid to be replaced
• Construct a mutation probability matrix
• Construct a relatedness odds matrix
• Construct a log odds matrix
PAM matrix construction

- Aligned sequences are ≥ 85% similar
  - To reduce the possibility that observed changes are due to two substitution events rather than one
  - To facilitate the reconstruction of phylogenetic trees and the inference of ancestral sequences
Ancestral sequences

Organism A: A W T V A A A V R T S I
Organism B: A Y T V A A A V R T S I
Organism C: A W T V A A A V L T S I (ancestral)

One of the possible relationships between organisms A, B and C:

Assume that each position has undergone a maximum of one substitution (high similarity between sequences).

Implies that the number of changes between sequences is proportional to the evolutionary distance between them.

January 31, 2013  Luce Skrabanek
Frequency of substitution

• Count the number of changes of each amino acid into every other amino acid
  – Compare each sequence with its ancestor, inferred from the phylogenetic trees
• In 71 groups of proteins, they counted 1,572 changes (Dayhoff, 1979)
  – Of the 190 possible exchanges among all amino acids, 35 were never observed
  – Most common exchange was Asp ↔ Glu

January 31, 2013
Luce Skrabanek
Accepted point mutations

Relative mutability

• Calculate the likelihood that an amino acid will be replaced
  – Estimate relative mutability (relative number of changes)

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed changes</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Frequency of occurrence</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Relative mutability</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Scale mutabilities

- Mutabilities were scaled to the number of replacements per occurrence of the given amino acid per 100 residues in each alignment.
- Normalizes data for variations in amino acid composition, mutation rate, sequence length.

- Count the total number of changes of each amino acid and divide by the total "exposure to mutation".
- "Exposure to mutation" is the frequency of occurrence of each amino acid multiplied by the total number of amino acid changes per 100 sites, summed for all alignments.

*(Ala has arbitrarily been set to 100)*
Mutation probability matrix

→ Mutation probability matrix for one PAM
  – Each element of this matrix gives the probability that the amino acid in column i will be replaced by the amino acid in row j after a given evolutionary interval
  – 1 PAM ≈ 1 mutation/100 residues/10 MY
  – 1 PAM ≈ 0.01 changes/aligned residue

\[
M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}}
\]

\[
M_{ii} = 1 - \lambda m_i
\]

where:
- \( M_{ij} \) = mutation probability
- \( m_j \) = relative mutability
- \( A_{ij} \) = number of times i replaced by j (from matrix of accepted point mutations)

The proportionality constant (\( \lambda \)) is chosen such that there is an average of one mutation every hundred residues.

January 31, 2013
Luce Skrabane
PAM1 mutation probability matrix

| Original Amino Acid | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| Ala                 | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 22 | 35 | 32 | 0 | 2 | 18 |
| Arg                 | 1 9913 | 1 | 0 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| Asn                 | 4 1 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 12 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| Asp                 | 6 0 42 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 1 |
| Cys                 | 1 1 0 0 9973 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| Gln                 | 3 9 4 5 0 9876 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 2 | 0 | 1 |
| Glu                 | 10 0 7 56 0 35 9865 | 4 | 2 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 |
| His                 | 21 1 12 11 1 3 7 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 5 |
| Ile                 | 1 2 18 3 1 20 1 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| Leu                 | 2 2 3 1 2 1 2 0 | 0 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| Lys                 | 3 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 9947 | 2 | 45 | 13 | 3 | 1 | 3 | 6 | 2 | 15 |
| Met                 | 2 37 25 6 0 12 | 7 | 3 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 |
| Phe                 | 1 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 9946 | 0 | 2 | 1 | 3 | 28 | 0 |
| Pro                 | 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2 |
| Ser                 | 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| Thr                 | 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 1 | 1 | 2 | 8 | 6 | 1 | 5 | 32 | 9871 | 0 | 2 | 9 |
| Trp                 | 0 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| Tyr                 | 1 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 1 | 0 | 1 | 12 | 0 | 1 | 2 | 9945 | 1 |
| Val                 | 13 | 2 | 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 9901 |

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, $M_{ij}$, gives the probability that the amino acid in column $j$ will be replaced by the amino acid in row $i$ after a given evolutionary interval, i.e., $M_{ij}$. 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.
PAM1, PAM10, PAM250

• To get different PAM matrices, the **mutation probability matrix** is multiplied by itself some integer number of times
  – PAM1 \(^{10}\) \(\rightarrow\) PAM10
  – PAM1 \(^{250}\) \(\rightarrow\) PAM250

• PAM10 applies to proteins that differ by 0.1 changes per aligned residue
  – 10 amino acid changes per 100 residues

• PAM250 reflects evolution of sequences with an average of 2.5 substitutions per position
  – 250% change in 2500 MY
  – Approximately 20% similarity
# PAM250 matrix mutation probability matrix

| ORIGINAL AMINO ACID | A | R | N | D | C | Q | E | G | F | P | S | T | W | Y | V |
|---------------------|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|
| A Ala               | 13| 6 | 9 | 9 | 5 | 8 | 9 | 12| 5 | 8 | 6 | 7 | 7 | 4 | 11| 11| 2 | 9 |
| R Arg               | 3 | 17| 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 3 | 7 |
| N Asn               | 4 | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 1 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 |
| D Asp               | 5 | 4 | 8 | 11| 1 | 7 | 10| 5 | 6 | 2 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 |
| C Cys               | 2 | 1 | 1 | 2 | 52| 1 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 3 | 2 | 3 |
| Q Gln               | 3 | 5 | 5 | 6 | 1 | 10| 7 | 3 | 7 | 2 | 3 | 5 | 2 | 1 | 4 | 3 | 3 | 1 |
| E Glu               | 5 | 4 | 7 | 11| 1 | 9 | 12| 5 | 6 | 2 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 |
| G Gly               | 12| 5 | 10| 1 | 4 | 7 | 9 | 27| 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11| 2 | 3 |
| H His               | 2 | 5 | 5 | 4 | 2 | 7 | 4 | 2 | 15| 2 | 2 | 3 | 2 | 2 | 4 | 3 | 3 | 1 |
| I Ile               | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10| 6 | 2 | 6 | 5 | 2 | 3 | 4 | 1 |
| L Leu               | 6 | 4 | 4 | 3 | 2 | 6 | 4 | 3 | 5 | 15| 34| 24| 9 | 13| 5 | 4 | 6 | 6 |
| K Lys               | 10| 18| 10| 8 | 2 | 10| 8 | 5 | 9 | 5 | 4 | 24| 9 | 2 | 6 | 8 | 8 | 4 |
| M Met               | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 |
| F Phe               | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32| 1 | 2 | 2 | 4 |
| P Pro               | 7 | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 3 | 3 | 4 | 3 | 2 | 20| 6 | 5 | 1 | 2 |
| S Ser               | 9 | 6 | 6 | 7 | 7 | 6 | 7 | 9 | 5 | 4 | 6 | 5 | 3 | 9 | 10| 9 | 4 | 4 |
| T Thr               | 8 | 5 | 6 | 6 | 4 | 5 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11| 2 |
| W Trp               | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 55| 1 | 0 |
| Y Tyr               | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15| 1 | 2 | 2 | 3 |
| V Val               | 7 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15| 10| 4 | 10| 5 | 5 | 5 | 7 | 2 | 4 |

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala will become Arg in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.
Relatedness odds matrix

- Calculate a relatedness odds matrix:
  - Divide each element of the mutation probability matrix by the frequency of occurrence of the replacing residue
- Gives the ratio that the substitution represents a real evolutionary event as opposed to random sequence variation

\[
R_{ij} = \frac{M_{ij}}{f_i}
\]

<table>
<thead>
<tr>
<th>Residue</th>
<th>Frequency of occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>0.096</td>
</tr>
<tr>
<td>Gly</td>
<td>0.090</td>
</tr>
<tr>
<td>Lys</td>
<td>0.085</td>
</tr>
<tr>
<td>Leu</td>
<td>0.085</td>
</tr>
<tr>
<td>Val</td>
<td>0.078</td>
</tr>
<tr>
<td>Thr</td>
<td>0.062</td>
</tr>
<tr>
<td>Ser</td>
<td>0.057</td>
</tr>
<tr>
<td>Asp</td>
<td>0.053</td>
</tr>
<tr>
<td>Glu</td>
<td>0.053</td>
</tr>
<tr>
<td>Phe</td>
<td>0.045</td>
</tr>
<tr>
<td>Asn</td>
<td>0.042</td>
</tr>
<tr>
<td>Pro</td>
<td>0.041</td>
</tr>
<tr>
<td>Ile</td>
<td>0.035</td>
</tr>
<tr>
<td>His</td>
<td>0.034</td>
</tr>
<tr>
<td>Arg</td>
<td>0.034</td>
</tr>
<tr>
<td>Gln</td>
<td>0.032</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.030</td>
</tr>
<tr>
<td>Cys</td>
<td>0.025</td>
</tr>
<tr>
<td>Met</td>
<td>0.012</td>
</tr>
<tr>
<td>Trp</td>
<td>0.012</td>
</tr>
</tbody>
</table>
Log odds matrix

• Calculate the log odds matrix
  – Take the log of each element
    • More convenient when using the matrix to score an alignment
      – score an alignment by summing the individual log odds scores for each aligned residue pair, rather than multiplying the odds scores
  • Assume that each aligned residue is statistically independent of the others
    – Biologically dubious but mathematically convenient

January 31, 2013
Luce Skrabanek
PAM 250 log odds matrix

Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of -10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.
Problems with PAM

• Assumptions:
  – Replacement at a site depends only on the amino acid at that site and the probability given by the matrix
  – Sequences that are being compared have average amino acid composition
  – All amino acids mutate at the same rate
Problems with PAM contd.

• Sources of error:
  – Many sequences do not have average amino acid composition
  – Rare replacements were observed too infrequently to resolve relative probabilities accurately
  – Errors in PAM1 are magnified in the extrapolation to PAM250
  – Replacement is not equally probable over the whole sequence
Matrix values are based on the observed amino acid substitutions in a large set of conserved amino acid blocks (≈ signatures of protein families)
  – Frequently occurring highly conserved sequences are collapsed into one consensus sequence to avoid bias

Patterns that show similarity above a threshold are grouped into a set
  – Find the frequencies of each residue pair in the dataset
  → 62% threshold = BLOSUM62 matrix
BLOSUM vs. PAM

- Not based on an explicit evolutionary model
- Considers all changes in a family
- Matrices to compare more distant sequences are calculated separately
- Based on substitutions of conserved blocks
- Designed to find conserved domains in proteins

- Markov-based model of evolution
- Based on prediction of first changes in each family
- Matrices to compare more distant sequences are derived
- Based on substitutions over the whole sequence
- Tracks evolutionary origins of proteins
BLOSUM vs. PAM

Frommlet et al, Comp Stat and Data Analysis, 2006
Z-score

Calculation of statistical significance

• Shuffle one of the sequences
  – with the same amino acid or nucleotide composition and local sequence features

• Align the shuffled sequences to the first sequence
  – Generate scores for each of these alignments

• Generate an extreme value distribution
  – Calculate mean score and standard deviation of distribution

• Get the Z-score

\[ Z = \frac{\text{RealScore} - \text{MeanScore}}{\text{StandardDeviation}} \]
Statistical significance: many other methods

- Many random sequences (with no resemblance to either of the original sequences) are generated and aligned with each other
- Random sequences of different lengths are generated; scores increase as the log of the length
- A sequence may be aligned with all the sequences in a database library, the related sequences are discarded and the remainder are used to calculate the statistics
  - Variations: shuffle the sequence or shuffle the library