Sequence comparison: Score matrices

http://faculty.washington.edu/jht/GS559_2013/

Genome 559: Introduction to Statistical and Computational Genomics
Prof. James H. Thomas
FYI - informal inductive proof of best alignment path

Consider the last step in the best alignment path to node $\alpha$ below. This path must come from one of the three nodes shown, where $X$, $Y$, and $Z$ are the cumulative scores of the best alignments up to those nodes. We can reach node $\alpha$ by three possible paths: an A-B match, a gap in sequence A or a gap in sequence B:

$$
\text{seq } A
\begin{array}{c}
X \\
\bullet
\end{array}
\begin{array}{c}
Y \\
\bullet
\end{array}
\begin{array}{c}
\text{match} \\
\text{gap}
\end{array}
\begin{array}{c}
\text{gap} \\
\rightarrow
\end{array}
\begin{array}{c}
Z \\
\bullet
\end{array}
\begin{array}{c}
\rightarrow
\end{array}
\begin{array}{c}
\alpha \\
\bullet
\end{array}
\text{seq } B
$$

The best-scoring path to $\alpha$ is the maximum of:

$X + \text{match}$

$Y + \text{gap}$

$Z + \text{gap}$

BUT the best paths to $X$, $Y$, and $Z$ are analogously the max of their three upstream possibilities, etc. Inductively QED.
Local alignment - review

\[
\begin{array}{cccc}
A & C & G & T \\
A & 2 & -7 & -5 & -7 \\
C & -7 & 2 & -7 & -5 \\
G & -5 & -7 & 2 & -7 \\
T & -7 & -5 & -7 & 2 \\
\end{array}
\]

\[d = -5\]

\[
\begin{array}{cccc}
A & A & A & G \\
0 & 0 & 0 & 0 \\
A & 0 & 2 & 2 & 0 \\
G & 0 & 0 & 0 & 4 \\
C & 0 & 0 & 0 & 0 \\
\end{array}
\]

(no arrow means no preceding alignment)
Local alignment

- Two differences from global alignment:
  - If a score is negative, replace with 0.
  - Traceback from the highest score in the matrix and continue until you reach 0.

- Global alignment algorithm: Needleman-Wunsch.
- Local alignment algorithm: Smith-Waterman.
dot plot of two DNA sequences

overlay of the global DP alignment path
Protein score matrices

• Quantitatively represent the degree of conservation of typical amino acid residues over evolutionary time.

• All possible amino acid changes are represented (matrix of size at least 20 x 20).

• Most commonly used are several different BLOSUM matrices derived for different degrees of evolutionary divergence.

• DNA score matrices are simpler (and conceptually similar).
**BLOSUM62 Score Matrix**

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

Hydrophobic

- glycine (G)
- alanine (A)
- valine (V)
- leucine (L)
- isoleucine (I)
- methionine (M)
- proline (P)
- tryptophan (W)
- phenylalanine (F)

Polar

- cysteine (C)
- serine (S)
- threonine (T)
- tyrosine (Y)
- asparagine (N)
- glutamine (Q)

Charged

- histidine (H)
- lysine (K)
- arginine (R)
- aspartate (D)
- glutamate (E)
### BLOSUM62 Score Matrix

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

Hydrophobic
- glycine G
- alanine A
- valine V
- leucine L
- isoleucine I
- methionine M
- proline P
- tryptophan W

Polar
- cysteine C
- serine S
- threonine T
- tyrosine Y
- asparagine N
- glutamine Q

Charged
- histidine H
- lysine K
- arginine R
- aspartate D
- glutamate E
Deriving BLOSUM scores

- Find sets of sequences whose alignment is thought to be correct (this is partly bootstrapped by alignment).

- Measure how often various amino acid pairs occur in the alignments.

- Normalize this to the expected frequency of such pairs randomly in the same set of alignments.

- Derive a log-odds score for aligned vs. random.
Example of alignment block (the BLO part of BLOSUM)

- Thousands of such blocks go into computing a single BLOSUM matrix.
- Represent full diversity of sequences.
- Results are summed over all columns of all blocks.
Pair frequency vs. expectation

Actual aligned pair frequency:

\[ q_{ij} = \frac{1}{T} \sum c_{ij} \]

where \( c_{ij} \) is the count of \( ij \) pairs and \( T \) is the total pair count.

Randomly expected pair frequency:

\[ e_{aa} = p_a p_a \]

\[ e_{ab} = p_a p_b + p_b p_a = 2 p_a p_b \]

where \( p_a \) and \( p_b \) are the overall probabilities (frequencies) of specific residues \( a \) and \( b \).
Log-odds score calculation (so adding scores == multiplying probabilities)

\[
    s_{ij} = \log_2 \frac{q_{ij}}{e_{ij}}
\]

For computational speed often rounded to nearest integer and (to reduce round-off error) they are often multiplied by 2 (or more) first, giving a “half-bit” score:

\[
    \text{matrixScore} = (\text{rounded}) \ 2 \log_2 \frac{q_{ij}}{e_{ij}}
\]

(computers can add integers faster than floats)
### BLOSUM62 matrix (half-bit scores)

(9 half-bits = 4.5 bits)

Frequency of \( C \) residue over all proteins: 0.0162 (you have to look this up)

Reverse calculation of aligned \( C-C \) pair frequency in BLOSUM data set:

\[
\frac{q_{cc}}{e_{cc}} = 2^{(4.5)} = 22.63 \quad \text{Thus} \quad q_{cc} = 22.63 \times 0.000262 = 0.00594
\]

\[
e_{cc} = 0.0162 \times 0.0162 = 0.000262
\]
Constructing Blocks

- Blocks are ungapped alignments of multiple sequences, usually 20 to 100 amino acids long.

- Cluster the members of each block according to their percent identity.

- Make pair counts and score matrix from a large collection of similarly clustered blocks.

- Each BLOSUM matrix is named for the percent identity cutoff in step 2 (e.g. BLOSUM70 for 70% identity).
Randomly Distributed Gaps

if \( p_g = k \) (probability of a gap at each position in the sequence)

then \( P(g_1) = k, P(g_2) = k^2, \ldots, P(g_n) = k^n \)

[Note - the slope of the line on a log-linear plot will vary according to the frequency of gaps, but it will always be linear]
Distribution of real alignment gap lengths in a large set of X-ray structure-aligned proteins

Nowhere near linear - hence the use of affine gap penalties (there ideally would be several levels of decreasing affine penalties)
What you should know

• How a score matrix is derived
• What the scores mean probabilistically
• Why gap penalties should be affine