Pairwise Alignment and Database Searching

Anders Gorm Pedersen
Molecular Evolution Group
Center for Biological Sequence Analysis
Sequences are related

- Darwin: all organisms are related through descent with modification
- => Sequences are related through descent with modification
- => Similar molecules have similar functions in different organisms
Sequences are related, II

Phylogenetic tree of globin-type proteins found in humans
Why compare sequences?

- Determination of evolutionary relationships
- Prediction of protein function and structure (database searches).

Protein 1: binds oxygen

Sequence similarity

Protein 2: binds oxygen?
Dotplots: visual sequence comparison

1. Place two sequences along axes of plot
2. Place dot at grid points where two sequences have identical residues
2. Diagonals correspond to conserved regions
Pairwise alignments

43.2% identity;  
Global alignment score: 374

10 20 30 40 50
alpha V-LSPADKTNVKAAWGKVGAGHAGGAEALERMFLSFPTTKTYFPFH-DLS-----HGSA
beta VHLTPEEKSAVTALWGVK--NVDEVGGALRGRLVYPWTQRFFESFGDLSTPDAVMGNP

10 20 30 40 50

60 70 80 90 100 110
alpha QVKGHGKVKVADALTNAHVDMPNALSALSDLHAHKLRVDPVNFKLSSHCLLVTAAHL
beta KVKAHGKVKLGAFSGLAHLDNLKGTFALTSLHCDKLHVDPENFRLLGVLVCVLAAHF

120 130 140
alpha PAEFTPAVHASLDFKLFASVSTVLTSKYR
beta GKEFTPPVQAAYQKVAGVANALAHKYH

120 130 140
Pairwise alignment

100.000% identity in 3 aa overlap

SPA
:::
SPA

Percent identity is not a good measure of alignment quality
Pairwise alignments: alignment score

43.2% identity;  

Global alignment score: 374

alpha  V-LSPADKTNVKAAGKVGHAVGAYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
beta   VHLTPEEKSAVTALWGKV--NVDEVGGAEALGRLLVVYPWTQRFESFGDLSTPDAVMGNP

alpha  QVKGHGKKVADALTNAVHAVDDMPNALSALSDLHAHKLRLVDPVNFKLLSHCLLVTLAHL
beta   KVKAHGKKVGLGAFSDGLAHLDNLKGTFATLSHELHCDELHLVDPENFRLGLGVRVCLLHAF

alpha  PAEFTPAVHASLDKFLASVSTVLTSKYR
beta   GKEFTPPVQAAYQKVAVANALAHKYH
Alignment scores: match vs. mismatch

Simple scoring scheme (too simple in fact…):

Matching amino acids: 5
Mismatch: 0

Scoring example:

K A W S A D V
: : : : :
K D W S A E V
\[5+0+5+5+5+0+5 = 25\]
Pairwise alignments: conservative substitutions

43.2% identity;  
Global alignment score: 374

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>V-LSPADKTNVKAAWGKVGAHAGEYGAEA</td>
<td>ALERMFLSFP</td>
<td>TTKTYFP</td>
<td>HF-DLS-</td>
<td>----HGSA</td>
</tr>
<tr>
<td>beta</td>
<td>VHLTPEEKSAVTALW</td>
<td>GKV--NVDE</td>
<td>VGEA</td>
<td>LGRLLV</td>
<td>VYPWTQ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>110</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>QVKGHGKKVADALTNAV</td>
<td>AHVD</td>
<td>MPNALSAL</td>
<td>SDLHAH</td>
<td>KLVRVDP</td>
<td>VNFKLLL</td>
<td>SHCLL</td>
<td>VT</td>
<td>LAAHL</td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>KVKAHGKKVLGAF</td>
<td>SDGLAHL</td>
<td>DKGTATL</td>
<td>SELHCDK</td>
<td>LHVD</td>
<td>PENFRL</td>
<td>LGNV</td>
<td>LVCVL</td>
<td>AHHF</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>120</th>
<th>130</th>
<th>140</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>PAEFTPAVHASL</td>
<td>DKFLASVST</td>
<td>VLT</td>
<td>SKYR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>GKEFTPPVQ</td>
<td>AAYQKV</td>
<td>VAGVAN</td>
<td>ALAHKYH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Amino acid properties

Serine (S) and Threonine (T) have similar physicochemical properties

Aspartic acid (D) and Glutamic acid (E) have similar properties

=> Substitution of S/T or E/D occurs relatively often during evolution

=> Substitution of S/T or E/D should result in scores that are only moderately lower than identities
## Protein substitution matrices

|     | A   | R   | N   | D   | C   | Q   | E   | G   | H   | I   | L   | K   | M   | F   | P   | S   | T   | W   | Y   | V   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A   | 5   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| R   | -2  | 7   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| N   | -1  | -1  | 7   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| D   | -2  | -2  | 2   | 8   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| C   | -1  | -4  | -2  | -4  | 13  |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Q   | -1  | 1   | 0   | 0   | -3  | 7   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| E   | -1  | 0   | 0   | 2   | -3  | 2   | 6   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| G   | 0   | -3  | 0   | -1  | -3  | -2  | -3  | 8   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| H   | -2  | 0   | 1   | -1  | -3  | 1   | 0   | -2  | 10  |     |     |     |     |     |     |     |     |     |     |     |     |     |
| I   | -1  | -4  | -3  | -4  | -2  | -3  | -4  | -4  | 5   |     |     |     |     |     |     |     |     |     |     |     |     |     |
| L   | -2  | -3  | -4  | -4  | -2  | -2  | -3  | -4  | -3  | 2   | 5   |     |     |     |     |     |     |     |     |     |     |     |
| K   | -1  | 3   | 0   | -1  | -3  | 2   | 1   | -2  | 0   | -3  | -3  | 6   |     |     |     |     |     |     |     |     |     |     |
| M   | -1  | -2  | -2  | -4  | -2  | 0   | -2  | -3  | -1  | 2   | 3   | -2  | 7   |     |     |     |     |     |     |     |     |     |     |
| F   | -3  | -3  | -4  | -5  | -2  | -4  | -3  | -4  | -1  | 0   | 1   | -4  | 0   | 8   |     |     |     |     |     |     |     |     |     |     |
| P   | -1  | -3  | -2  | -1  | -4  | -1  | -1  | -2  | -2  | -3  | -4  | -1  | -3  | -4  | 10  |     |     |     |     |     |     |     |     |     |
| S   | 1   | -1  | 1   | 0   | -1  | 0   | -1  | -3  | -3  | 0   | -2  | -3  | -1  | 5   |     |     |     |     |     |     |     |     |     |
| T   | 0   | -1  | 0   | -1  | -1  | -1  | -2  | -2  | -1  | -1  | -1  | -2  | -1  | 2   | 5   |     |     |     |     |     |     |     |     |     |
| W   | -3  | -3  | -4  | -5  | -5  | -1  | -3  | -3  | -3  | -2  | -3  | -1  | 1   | -4  | -4  | -3  | 15  |     |     |     |     |     |     |     |     |
| Y   | -2  | -1  | -2  | -3  | -3  | -1  | -2  | -3  | 2   | -1  | -1  | -2  | 0   | 4   | -3  | -2  | -2  | 2   | 8   |     |     |     |     |
| V   | 0   | -3  | -3  | -4  | -1  | -3  | -3  | -4  | -4  | 4   | 1   | -3  | 1   | -1  | -3  | -2  | 0   | -3  | -1  | 5   |     |     |     |     |

### BLOSUM50 matrix:

- Positive scores on diagonal (identities)
- Similar residues get higher (positive) scores
- Dissimilar residues get smaller (negative) scores
### Pairwise alignments: insertions/deletions

43.2% identity; Global alignment score: 374

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>V-LSPADKTNVKAAGVGAHEYGAEALERMFLS-FPTTKTYFPHF-DLS----HGSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVYPWTQRFFESFGDLSTPDAVMGNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>alpha</td>
<td>QVKGHGKKVADALTNAVAHAVDDMPNALSALSDLHAKLRVDPVNFKLSSLHCLLLVTLLAHL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>KVKAHGKKVLGAFAVGLHLDNLKGTFATLSHELCHDKLHVDPNGFPRLGNMLVCLAHHF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>130</td>
<td>140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>PAEFTPVHASLDKFLASVSTVLTLYR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>GKEFTPPVQAAYQKVAGVANALAHKYH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alignment scores: insertions/deletions

Affine gap penalties:
Multiple insertions/deletions may be one evolutionary event =>
Separate penalties for gap opening and gap elongation
Compute 4 alignment scores: two different alignments using two different alignment matrices (and the same gap penalty system)

Score 1: Alignment 1 + BLOSUM-50 matrix + gaps
Score 2: Alignment 1 + BLOSUM-Trp matrix + gaps
Score 3: Alignment 2 + BLOSUM-50 matrix + gaps
Score 4: Alignment 2 + BLOSUM-Trp matrix + gaps
Handout: summary of results

<table>
<thead>
<tr>
<th></th>
<th>Alignment 1</th>
<th>Alignment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOSUM-50</td>
<td>38</td>
<td>51</td>
</tr>
<tr>
<td>BLOSUM-Trp</td>
<td>118</td>
<td>91</td>
</tr>
</tbody>
</table>
Protein substitution matrices

|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| R | -2 | 7 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N | -1 | -1 | 7 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D | -2 | -2 | 2 | 8 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| C | -1 | -4 | -2 | -4 | 13 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Q | -1 | 1 | 0 | 0 | -3 | 7 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| E | -1 | 0 | 0 | 2 | -3 | 2 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| G | 0 | -3 | 0 | -1 | -3 | -2 | -3 | 8 |   |   |   |   |   |   |   |   |   |   |   |   |
| H | -2 | 0 | 1 | -1 | -3 | 1 | 0 | -2 | 10 |   |   |   |   |   |   |   |   |   |   |   |
| I | -1 | -4 | -3 | -4 | -2 | -3 | -4 | -4 | -4 | 5 |   |   |   |   |   |   |   |   |   |   |
| L | -2 | -3 | -4 | -4 | -2 | -2 | -3 | -4 | -3 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |
| K | -1 | 3 | 0 | -1 | -3 | 2 | 1 | -2 | 0 | -3 | -3 | 6 |   |   |   |   |   |   |   |   |   |   |
| M | -1 | -2 | -2 | -4 | -2 | 0 | -2 | -3 | -1 | 2 | 3 | -2 | 7 |   |   |   |   |   |   |   |   |   |   |
| F | -3 | -3 | -4 | -5 | -2 | -4 | -3 | -4 | -1 | 0 | 1 | -4 | 0 | 8 |   |   |   |   |   |   |   |   |   |   |
| P | -1 | -3 | -2 | -1 | -4 | -1 | -1 | -2 | -2 | -3 | -4 | -1 | -3 | -4 | 10 |   |   |   |   |   |   |   |   |   |   |
| S | 1 | -1 | 1 | 0 | -1 | 0 | -1 | 0 | -1 | -3 | -3 | 0 | -2 | -3 | -1 | 5 |   |   |   |   |   |   |   |   |   |   |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |
| W | -3 | -3 | -4 | -5 | -5 | -1 | -3 | -3 | -3 | -3 | -2 | -3 | -1 | 1 | -4 | -4 | -3 | 15 |   |   |   |   |   |   |   |   |   |   |
| Y | -2 | -1 | -2 | -3 | -3 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | 0 | 4 | -3 | -2 | -2 | 2 | 8 |   |   |   |   |   |   |   |   |   |   |
| V | 0 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -4 | 4 | 1 | -3 | 1 | -1 | -3 | -2 | 0 | -3 | -1 | 5 |   |   |   |   |   |   |   |   |   |   |

**BLOSUM50 matrix:**

• Positive scores on diagonal (identities)

• Similar residues get higher (positive) scores

• Dissimilar residues get smaller (negative) scores
Protein substitution matrices: different types

- **Identity matrix**
  (match vs. mismatch)

- **Genetic code matrix**
  (how similar are the codons?)

- **Chemical properties matrix**
  (use knowledge of physicochemical properties to design matrix)

- **Empirical matrices**
  (based on observed pair-frequencies in hand-made alignments)
  - PAM series
  - BLOSUM series
  - Gonnet
Estimation of the BLOSUM 50 matrix

- For each alignment in the BLOCKS database the sequences are grouped into clusters with at least 50% identical residues (for BLOSUM 50).

- All pairs of sequences are compared, and the observed pair frequencies are noted (e.g., A aligned with A makes up 1.5% of all pairs. A aligned with C makes up 0.01% of all pairs, etc.).

- Expected pair frequencies are computed from single amino acid frequencies. (e.g., \( f_{A,C} = f_A \times f_C = 7\% \times 3\% = 0.21\%).

- For each amino acid pair the substitution scores are essentially computed as:

\[
S_{A,C} = \log \frac{\text{Pair-freq(obs)}}{\text{Pair-freq(expected)}}
\]
Pairwise alignment

Optimal alignment:

alignment having the highest possible score given a substitution matrix and a set of gap penalties
Pairwise alignment: the problem

The number of possible pairwise alignments increases explosively with the length of the sequences:

Two protein sequences of length 100 amino acids can be aligned in approximately $10^{60}$ different ways.

Time needed to test all possibilities is same order of magnitude as the entire lifetime of the universe.
Pairwise alignment: the solution

"Dynamic programming"
(the Needleman-Wunsch algorithm)
Alignment depicted as path in matrix

TCGCA
TC-CA

TCGCA
T-CCA
Alignment depicted as path in matrix

Meaning of point in matrix: all residues up to this point have been aligned (but there are many different possible paths).

Position labeled “x”: TC aligned with TC

```
TC
TC
-TC
T-C
--TC
TC--
```
Dynamic programming: computation of scores

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.
Dynamic programming: computation of scores

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

\[
\text{score}(x,y) = \max \left\{ \text{score}(x,y-1) - \text{gap-penalty} \right\}
\]
Dynamic programming: computation of scores

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

\[
\text{score}(x,y) = \max \begin{cases} 
\text{score}(x,y-1) - \text{gap-penalty} \\
\text{score}(x-1,y-1) + \text{substitution-score}(x,y) 
\end{cases}
\]
Dynamic programming: computation of scores

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

score(x,y) = \max\ \begin{cases} 
\text{score}(x,y-1) - \text{gap-penalty} \\
\text{score}(x-1,y-1) + \text{substitution-score}(x,y) \\
\text{score}(x-1,y) - \text{gap-penalty} 
\end{cases}
Dynamic programming: computation of scores

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

Each new score is found by choosing the maximum of three possibilities. For each square in matrix: keep track of where best score came from.

Fill in scores one row at a time, starting in upper left corner of matrix, ending in lower right corner.

\[
score(x,y) = \max\left\{\begin{array}{l}
score(x,y-1) - \text{gap-penalty} \\
score(x-1,y-1) + \text{substitution-score}(x,y) \\
score(x-1,y) - \text{gap-penalty}
\end{array}\right.
\]
Dynamic programming: example

\[
a[i,j] = \max \begin{cases} a[i,j-1] - 2 \\ a[i-1,j-1] + p(i,j) \\ a[i-1,j] - 2 \end{cases}
\]
Dynamic programming: example

\[ a[i,j] = \max \begin{cases} a[i,j-1] - 2 \\ a[i-1,j-1] + p(i,j) \\ a[i-1,j] - 2 \end{cases} \]
Dynamic programming: example

\[
a[i, j] = \max\begin{cases}
a[i, j-1] - 2 \\
a[i-1, j-1] + p(i, j) \\
a[i-1, j] - 2
\end{cases}
\]
Dynamic programming: example

\[
a[i,j] = \max \begin{cases} 
  a[i,j-1] - 2 \\
  a[i-1,j-1] + p(i,j) \\
  a[i-1,j] - 2 
\end{cases}
\]
Dynamic programming: example
Dynamic programming: example

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & -2 & -4 & -6 & -8 & -10 \\
1 & -2 & 1 & -1 & -3 & -5 & -7 \\
2 & -4 & -1 & 2 & 0 & -2 & -4 \\
3 & -6 & -3 & 0 & 1 & 1 & -1 \\
4 & -8 & -5 & -2 & -1 & 0 & 2 \\
\end{array}
\]

\[
T \quad C \quad G \quad C \quad A \\
: \quad : \quad : \quad : \quad : \\
T \quad C \quad - \quad C \quad A \\
1+1-2+1+1 = 2
\]
Global versus local alignments

**Global alignment**: align full length of both sequences. (The “Needleman-Wunsch” algorithm).

**Local alignment**: find best partial alignment of two sequences (the “Smith-Waterman” algorithm).
Local alignment overview

- The recursive formula is changed by adding a fourth possibility: zero. This means local alignment scores are never negative.

\[
\text{score}(x, y) = \max \begin{cases} 
\text{score}(x, y-1) - \text{gap-penalty} \\
\text{score}(x-1, y) - \text{gap-penalty} \\
\text{score}(x-1, y-1) + \text{substitution-score}(x, y) \\
0
\end{cases}
\]

- Trace-back is started at the highest value rather than in lower right corner
- Trace-back is stopped as soon as a zero is encountered
Local alignment: example

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>E</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>W</th>
<th>G</th>
<th>H</th>
<th>E</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<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>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>18</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>8</td>
<td>21</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>20</td>
<td>27</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>6</td>
<td>13</td>
<td>18</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>16</td>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>

**AWGHE**

**AW-HE**
Substitution matrices and sequence similarity

- Substitution matrices come as series of matrices calculated for different degrees of sequence similarity (different evolutionary distances).

- "Hard" matrices are designed for similar sequences
  - Hard matrices are designated by high numbers in the BLOSUM series (e.g., BLOSUM80)
  - Hard matrices yield short, highly conserved alignments

- "Soft" matrices are designed for less similar sequences
  - Soft matrices have low BLOSUM values (45)
  - Soft matrices yield longer, less well conserved alignments
Alignments: things to keep in mind

“Optimal alignment” means “having the highest possible score, given substitution matrix and set of gap penalties”.

This is NOT necessarily the biologically most meaningful alignment.

Specifically, the underlying assumptions are often wrong: substitutions are not equally frequent at all positions, affine gap penalties do not model insertion/deletion well, etc.

Pairwise alignment programs **always** produce an alignment - even when it does not make sense to align sequences.