Needleman-Wunsch Algorithm

- Key idea: build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences.
- Optimal align of $S[1], ..., S[i]$ vs $T[1], ..., T[j]$: 
  
$$ V(i,j) = \max \begin{cases} 
V(i-1,j-1) + \sigma(S[i],T[j]) \\
V(i-1,j) + \sigma(S[i], -) \\
V(i,j-1) + \sigma(-, T[j]) 
\end{cases} $$

for all $1 \leq i \leq n, 1 \leq j \leq m.$
### Align by Dynamic Programming

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>←T</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td></td>
<td>c</td>
<td>a</td>
<td>d</td>
<td>b</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>c</td>
<td>-4</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>d</td>
<td>-5</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>b</td>
<td>-6</td>
<td>-3</td>
<td>-3</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

### Scoring Rules/Matrices

- **How should \( \sigma \) be defined?**
  - \( \sigma(A,G), \sigma(A,-), \sigma(A,-), \) etc?

- **Why are they important?**
  - The choice of a scoring rule can strongly influence the outcome of sequence analysis.

- **What do they mean?**
  - Scoring matrices implicitly represent a particular theory of evolution.
  - Elements of the matrices specify the similarity of one residue to another.

*Refers to an amino acid.*
Outline: Scoring Alignments

- Probabilistic meaning

- Scoring matrices
  - PAM: scoring based on evolutionary statistics
  - BLOSUM: tuning to evolutionary conservation

- Gaps revisited

---

Probabilistic Interpretation

\[
\begin{align*}
X: & \quad TCCAGGTG - GAT \\
 & \quad | \quad | \quad | \quad | \quad | \quad | \quad |
Y: & \quad TGCAAGTGC\textcolor{red}{{-T}}
\end{align*}
\]

Chance or true homology?

Sharing a common ancestor
Likelihood Ratio

X: TCCAGGTG–GAT
   ||||| |||
Y: TGCAAGTGCG–T

\[
\frac{\Pr(\text{Data} \mid \text{Homology})}{\Pr(\text{Data} \mid \text{Chance})}
\]

Pr( Data | Chance )
Given an alignment between TCCAGG and TGCAAG,

\[
\frac{\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{\textperiodcentered})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})}{\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})\Pr(\text{-})}
\]
Pr( Data | Homology )

Given an alignment between TCCAGG and TGCAAG,

\[
\begin{array}{cccccc}
T & C & C & A & G & G \\
\text{Pr}_1 & \text{Pr}_2 & \text{Pr}_3 & \text{Pr}_4 & \text{Pr}_5 & \text{Pr}_6 \\
T & G & C & A & A & G \\
\end{array}
\]

Likelihood Ratio

\[
\frac{\Pr(\text{Data|homology})}{\Pr( \text{Data| Chance})} = \prod_i \frac{\Pr(x_i,y_i)}{\Pr(x_i)\Pr(y_i)}
\]
Score: Log Likelihood Ratio

- The most commonly used alignment score of aligning two sequences is the log likelihood ratio of the alignment under two models:
  - Common ancestry
  - By chance

\[
Score = \log \left( \prod_i \frac{Pr(x_i, y_i)}{Pr(x_i)Pr(y_i)} \right) = \\
= \sum_i \log \left( \frac{Pr(x_i, y_i)}{Pr(x_i)Pr(y_i)} \right) = \sum_i s(x_i, y_i)
\]

The \( S \) in a Scoring Matrix
(as log likelihood ratio)

How do we acquire the probabilities \( Pr(a), Pr(a,b) \)?
Making a Scoring Matrix

- Scoring matrices $S$ are created based on biological evidence.
  - Alignments can be thought of as two sequences that differ due to mutations.
  - Some of these mutations have little effect on the protein’s function, therefore some penalties will be less harsh than others.

\begin{align*}
  \text{TCCAGGTG} & \text{-GAT} \\
  | & | & | & | & | \\
  \text{TGCAAGTGCG} & \text{-T}
\end{align*}

Scoring Matrix: Example

- Notice that although R (arginine) and K (Lysine) are different amino acids, they have a positive score.

\begin{align*}
\begin{array}{cccc}
  A & R & N & K \\
  A & 5 & -2 & -1 & -1 \\
  R & - & 7 & -1 & 3 \\
  N & - & - & 7 & 0 \\
  K & - & - & - & 6 \\
\end{array}
\end{align*}

- Why? They are both positively charged amino acids $\rightarrow$ will not greatly change function of protein.

\begin{align*}
\text{AKRANR} \quad \text{KAAANK} \\
\text{-1 + (-1) + (-2) + 5 + 7 + 3 = 11}
\end{align*}
Conservation

- Amino acid changes that tend to preserve the physical/chemical properties of the original residue
  - Polar to polar
    - aspartate (D) $\rightarrow$ glutamate (E)
  - Nonpolar to nonpolar
    - alanine (A) $\rightarrow$ valine (V)
  - Similarly behaving residues
    - leucine (L) to isoleucine (I)

- More prone to mutate in the evolutionary process.

Edit Operations Over Time

$$\text{TCCAGGTG$-$GAT}$$

$$\text{tgcaagtgcg$-$t}$$
We need a probabilistic model for the evolutionary changes of the sequence.
Most Widely Used Scoring Matrices

- Amino acid substitution matrices
  - PAM
  - BLOSUM

- DNA substitution matrices
  - Warning: when the sequences of interest code for protein, it is almost always better to compare the protein translations than to compare the DNA sequences directly.
  - DNA is less conserved than protein sequences
    - After only a small amount of evolutionary change, the DNA sequences, when compared using simple nucleotide substitution scores, contain less information with which to deduce homology than do the encoded protein sequences
  - Less effective to compare coding regions at nucleotide level

PAM

- Point Accepted Mutation*
- 1 PAM = PAM\(_1\) = 1% average change of all amino acid positions
  - After 100 PAMs of evolution, not every residue will have changed
    - some residues may have mutated several times
    - some residues may have returned to their original state
    - some residues may not changed at all

PAM Matrices: Training Data

- Take aligned set of closely related proteins
  - 71 groups of proteins that were at least 85% similar

- Each group of sequences were organized into a phylogenetic tree
  - Creates a model of the order in which substitutions occurred

- Count the number of changes of each amino acid into every other amino acid
  - Each substitution is considered to be an “accepted mutation” - an amino acid change “accepted” by natural selection

PAM: Point Accepted Mutation

- $A_{ij}$: number of times amino acid $j$ mutates to amino acid $i$.
  - A mutation could go in both directions, therefore the tally of mutation $i$-$j$ enters both $A_{ij}$ and $A_{ji}$ entries, while the tally of conservation $i$-$i$ enters $A_{ii}$ entry twice.

\[
\begin{array}{cccccc}
A & B & C & D & G & H \\
\hline
A & 8 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
B & 0 & 8 & A_1 & 1 & 1 & 0 & 0 & 0 & 0 \\
C & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
D & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
E & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
F & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
G & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
J & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]
Mutability of Residue j

- $m_j$ is the probability that amino acid $j$ will change in a given evolutionary interval.
  - It depends on how similar the sequences used to tally $A_{ij}$ are.

\[
m_j = 1 - \frac{A_{ij}}{\sum_{j=1,20} A_{ij}} = \frac{\sum_{i=1,20} A_{ij}}{\sum_{i=1,20} A_{ij}}
\]

**Relative mutability of amino acids**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (Asn)</td>
<td>134</td>
</tr>
<tr>
<td>H (His)</td>
<td>66</td>
</tr>
<tr>
<td>S (Ser)</td>
<td>120</td>
</tr>
<tr>
<td>R (Arg)</td>
<td>65</td>
</tr>
<tr>
<td>D (Asp)</td>
<td>106</td>
</tr>
<tr>
<td>K (Lys)</td>
<td>56</td>
</tr>
<tr>
<td>P (Pro)</td>
<td>56</td>
</tr>
<tr>
<td>A (Ala)</td>
<td>100</td>
</tr>
<tr>
<td>G (Gly)</td>
<td>49</td>
</tr>
<tr>
<td>T (Thr)</td>
<td>97</td>
</tr>
<tr>
<td>Y (Tyr)</td>
<td>41</td>
</tr>
<tr>
<td>I (Ile)</td>
<td>96</td>
</tr>
<tr>
<td>F (Phe)</td>
<td>41</td>
</tr>
<tr>
<td>M (Met)</td>
<td>94</td>
</tr>
<tr>
<td>L (Leu)</td>
<td>40</td>
</tr>
<tr>
<td>Q (Gln)</td>
<td>93</td>
</tr>
<tr>
<td>C (Cys)</td>
<td>20</td>
</tr>
<tr>
<td>V (Val)</td>
<td>74</td>
</tr>
<tr>
<td>W (Trp)</td>
<td>18</td>
</tr>
</tbody>
</table>

Total Mutation Rate

- $P_j$: probability of occurrence of amino acid $j$

\[
P_j = \frac{\sum_{i=1,20} A_{ij}}{\sum_{i=1,20} \sum_{j=1,20} A_{ij}}
\]

- Total mutation rate of all amino acids

\[
\sum_{j=1,20} P_j m_j = \text{m is the probability that amino acid } j \text{ will change in a given evolutionary interval.}
\]

- Normalize total mutation rate to 1%
  - $\lambda$ is a scaling constant to make sure that the total mutation is 1%

\[
\lambda \cdot \sum_{j=1,20} P_j m_j = 1\% \quad \Rightarrow \text{solve for } \lambda
\]

- This defines an evolutionary period: the period in which the 1% of all sequences are mutated
### Normalized Mutation Probability Matrix

- Normalize mutation probability matrix such that the total mutation rate is 1%

\[ M_0 \ (i \neq j) : \text{Probability of amino acid } j \text{ changing into } i \text{ in the evolutionary period} \]

\[ M_0 = \frac{A_{ij}}{\sum_{i < j} A_{ij}} \]

\[ M_0 : \text{Probability of amino acid } j \text{ not changing in PAM-1} \]

\[ M_0 = 1 - \sum_{i < j} M_{ij} = 1 - \lambda m_j \]

- dayhoff, m. o.; schwartz, r. m.; orcutt, b. c. (1978). "a model of evolutionary change in proteins". atlas of protein sequence and structure 5 (3): 345–352.

### Mutation Probability Matrix (transposed) M*10000

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| R | 4 | 1 | 9922 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| N | 0 | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| D | 0 | 6 | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 3 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| L | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| W | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

In Two PAM1 Periods

- $M^{(1)}$: PAM-1 mutation probability matrix
- $M^{(2)}$: PAM-2 mutation probability matrix
  - Mutations that happen in twice the evolution period of that for a PAM1

- $\{A \rightarrow R\} = \{A \rightarrow A \text{ and } A \rightarrow R\}$ or
  - $\{A \rightarrow N \text{ and } N \rightarrow R\}$ or
  - $\{A \rightarrow D \text{ and } D \rightarrow R\}$ or
  - ... or
  - $\{A \rightarrow V \text{ and } V \rightarrow R\}$ or


$\Pr(A \rightarrow R \text{ in 2 periods}) =$

- $\Pr(A \rightarrow A \text{ in 1st period}) \times \Pr(A \rightarrow R \text{ in 2nd period}) +$
- $\Pr(A \rightarrow N \text{ in 1st period}) \times \Pr(N \rightarrow R \text{ in 2nd period}) +$
- $\Pr(A \rightarrow D \text{ in 1st period}) \times \Pr(D \rightarrow R \text{ in 2nd period}) +$
  ...

$P^{(2)}_{ij} = P_{AA} \cdot P_{AR} + P_{AN} \cdot P_{NR} + P_{AD} \cdot P_{DR} + ...$

$\begin{bmatrix}
P_{AA} & P_{AR} & P_{AN} & \ldots \\
P_{RA} & P_{RR} & P_{RN} & \ldots \\
P_{NA} & P_{NR} & P_{NN} & \ldots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\times
\begin{bmatrix}
P_{AA} & P_{AR} & P_{AN} & \ldots \\
P_{RA} & P_{RR} & P_{RN} & \ldots \\
P_{NA} & P_{NR} & P_{NN} & \ldots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}$

- PAM-k Mutation Prob. Matrix

\[ M^{(2)} = M^{(1)} \times M^{(1)} \]
\[ M^{(K)} = \{M^{(1)}\}^K \]

PAM-k Log-Likelihood Matrix

- Log likelihood ratio score

\[ S_{ij} = 10 \log_{10} \left( \frac{\Pr(a_i, a_j)}{\Pr(a_i)\Pr(a_j)} \right) \]

\[ S_{ij} = 10 \log_{10} \frac{(M^K)_{ij}}{P_i} \]

\[ P_i = \frac{\sum_{j=1,20} A_{ij}}{\sum_{i=1,20} \sum_{j=1,20} A_{ij}} \]

\( S \) is a symmetric matrix
**BLOSUM: Henikoff & Henikoff 92**

- **BLOSUM: Block Substitution Matrices**
- **Motivation:** PAM use of matrix power can result in large errors
- **Key idea:** consider conserved patterns (blocks) of a large sample of proteins
  - Classify protein families (over 500 families)
  - Family has characteristic patterns (signatures) that are conserved
  - The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments.
- \( P(a,b) = \text{probability of (a,b) substitution; } P(a) = \text{probability of "a"} \)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bpi Bovine</td>
<td>npGivaRItqkgLdyacqqvgtltqkele</td>
</tr>
<tr>
<td>Bpi Human</td>
<td>npGvvvRIsqkgLdyasqgiqvaalQkelk</td>
</tr>
<tr>
<td>Cept Human</td>
<td>eaGivcRItkpaLLLvlhntvktvkiQtafqa</td>
</tr>
<tr>
<td>Lbp Human</td>
<td>npGlvavRItdkgLqyaqegllalQsell</td>
</tr>
<tr>
<td>Lbp Rabbit</td>
<td>npGlitRItdkgLeyaareqllalQrkl1</td>
</tr>
</tbody>
</table>
Scoring Matrices (e.g., BLOSUM)

- BLOSUMx is based on patterns that are x% similar
- The level of x% can provide different performance in identifying similarity
- BLOSUM62 provides good scoring (used as default)

Constructing BLOSUM $r$

- To avoid bias in favor of a certain protein, first eliminate sequences that are more than $r\%$ identical
- The elimination is done by either
  - removing sequences from the block, or
  - finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster.
- BLOSUM $r$ is the matrix built from blocks with no more the $r\%$ of similarity
  - E.g., BLOSUM62 is the matrix built using sequences with no more than 62% similarity.
  - Note: BLOSUM 62 is the default matrix for protein BLAST
Collecting substitution statistics

1. Count amino acids pairs in each column; e.g.,
   - 6 AA pairs, 4 AB pairs, 4 AC, 1 BC, 0 BB, 0 CC.
   - Total = 6+4+4+1=15
2. Normalize results to obtain probabilities ($p_X$'s and $p_{XY}$'s)
3. Compute log likelihood ratio score matrix from probabilities:
   $s(X,Y) = \log \left( \frac{p_{XY}}{p_X p_Y} \right)$

Comparison

- PAM is based on an evolutionary model using phylogenetic trees
- BLOSUM assumes no evolutionary model, but rather conserved “blocks” of proteins