CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS11.html
Global Alignment: An example

V: G A A T T C A G T T A
W: G G A T C G A

Given
\[ \delta[I, J] = \text{Score of Matching the } I^{th} \text{ character of sequence V & the } J^{th} \text{ character of sequence W} \]

Compute
\[ S[I, J] = \text{Score of Matching First } I \text{ characters of sequence V & First } J \text{ characters of sequence W} \]

Match/Mismatch score
Recurrence Relation
\[ \begin{align*}
S[I, J] &= \text{MAXIMUM} \{ \\
S[I-1, J-1] + \delta(V[I], W[J]), \\
S[I-1, J] + \delta(V[I], \text{——}), \\
S[I, J-1] + \delta(\text{——}, W[J]) \} \\
\end{align*} \]
What happens with last character(s)?

1. Last characters MATCH

2. Last characters MISMATCH

3. Last character of W aligned with GAP

4. Last character of V aligned with GAP
How to fill in the matrix?

1. Add Match or Mismatch Score
2. Add gap penalty for gap in seq 1
3. Add gap penalty for gap in seq 2
Global Alignment: An example

\[ S[I, J] = \text{MAXIMUM} \{ \]
\[ S[I-1, J-1] + \delta(V[I], W[J]), \]
\[ S[I-1, J] + \delta(V[I], -), \]
\[ S[I, J-1] + \delta(-, W[J]) \}\]

V: G A A T T C A G T T A
W: G G A T C G A

Global Alignment:

Match score = 1; Mismatch = Gap = -1
Traceback

V: G A A T T T C A G T T A

W: G G A - T C - G - - A
Alternative Traceback

V: G - A A T T C A G T T A
  |   |   |   |   |   |   |
W: G G - A - T C - G - - A

V: G A A T T C A G T T A
  |   |   |   |   |   |   |
W: G G A - T C - G - - A
### Improved Traceback

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</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>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>×1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>×1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>×1</td>
<td>↑1</td>
<td>↑1</td>
<td>↑1</td>
<td>↑1</td>
<td>↑1</td>
<td>×2</td>
<td>←2</td>
<td>←2</td>
<td>←2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>↑1</td>
<td>↑1</td>
<td>×2</td>
<td>←2</td>
<td>←2</td>
<td>×2</td>
<td>↑2</td>
<td>↑2</td>
<td>↑2</td>
<td>×3</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>↑1</td>
<td>←2</td>
<td>↑2</td>
<td>×3</td>
<td>×3</td>
<td>←3</td>
<td>←3</td>
<td>×3</td>
<td>×3</td>
<td>↑3</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>↑1</td>
<td>↑2</td>
<td>↑2</td>
<td>↑3</td>
<td>↑3</td>
<td>×4</td>
<td>←4</td>
<td>←4</td>
<td>←4</td>
<td>←4</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>↑1</td>
<td>↑2</td>
<td>↑2</td>
<td>↑3</td>
<td>↑3</td>
<td>↑4</td>
<td>↑4</td>
<td>×5</td>
<td>←5</td>
<td>←5</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>↑1</td>
<td>↑2</td>
<td>×3</td>
<td>↑3</td>
<td>↑3</td>
<td>×5</td>
<td>↑5</td>
<td>↑5</td>
<td>↑5</td>
<td>×6</td>
</tr>
</tbody>
</table>
### Improved Traceback

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>G</strong></td>
<td>0</td>
<td>×1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>×1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
<td>←1</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>0</td>
<td>×1</td>
<td>↑1</td>
<td>↑1</td>
<td>↑1</td>
<td>↑1</td>
<td>↑1</td>
<td>×2</td>
<td>←2</td>
<td>←2</td>
<td>←2</td>
<td>←2</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>0</td>
<td>↑1</td>
<td>↑1</td>
<td>×2</td>
<td>←2</td>
<td>←2</td>
<td>←2</td>
<td>×2</td>
<td>↑2</td>
<td>↑2</td>
<td>↑2</td>
<td>×3</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>0</td>
<td>↑1</td>
<td>←2</td>
<td>↑2</td>
<td>×3</td>
<td>×3</td>
<td>←3</td>
<td>←3</td>
<td>×3</td>
<td>×3</td>
<td>×3</td>
<td>↑3</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>0</td>
<td>↑1</td>
<td>↑2</td>
<td>↑2</td>
<td>↑3</td>
<td>↑3</td>
<td>×4</td>
<td>←4</td>
<td>←4</td>
<td>←4</td>
<td>←4</td>
<td>←4</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>0</td>
<td>↑1</td>
<td>↑2</td>
<td>↑2</td>
<td>↑3</td>
<td>↑3</td>
<td>↑4</td>
<td>↑4</td>
<td>×5</td>
<td>←5</td>
<td>←5</td>
<td>←5</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>0</td>
<td>↑1</td>
<td>↑2</td>
<td>×3</td>
<td>↑3</td>
<td>↑3</td>
<td>↑4</td>
<td>×5</td>
<td>↑5</td>
<td>↑5</td>
<td>↑5</td>
<td>×6</td>
</tr>
</tbody>
</table>
# Improved Traceback

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>V:</td>
<td>G</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W:</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>A</td>
<td>-</td>
<td>T</td>
<td>C</td>
<td>-</td>
<td>G</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V: G A - A T T C A G T T A</th>
</tr>
</thead>
<tbody>
<tr>
<td>W: G - G A - T C - G - - A</td>
</tr>
</tbody>
</table>
Subproblems

- Optimally align $V[1..I]$ and $W[1..J]$ for every possible values of $I$ and $J$.
- Having optimally aligned
  - $V[1..I-1]$ and $W[1..J-1]
  - $V[1..I]$ and $W[1..J-1]
  - $V[1..I-1]$ and $W[1, J]$

  it is possible to optimally align $V[1..I]$ and $W[1..J]$

- $O(mn)$,
  - where $m = \text{length of } V$,
  - and $n = \text{length of } W$. 
Generalizations of Similarity Function

- Mismatch Penalty = $\alpha$
- Spac$es$ (Insertions/Deletions, \textit{Indels}) = $\beta$
- Affine Gap Penalties:
  - (Gap open, Gap extension) = $(\gamma, \delta)$
- Weighted Mismatch = $\Phi(a,b)$
- Weighted Matches = $\Omega(a)$
## Alternative Scoring Schemes

**Match +1**

**Mismatch −2**

**Gap (-2, -1)**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-8</td>
<td>-9</td>
<td>-10</td>
<td>-11</td>
<td>-12</td>
</tr>
<tr>
<td>G</td>
<td>-2</td>
<td>× 1</td>
<td>← -1</td>
<td>← -2</td>
<td>← -3</td>
<td>← -4</td>
<td>← -5</td>
<td>← -6</td>
<td>← -7</td>
<td>← -8</td>
<td>← -9</td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>↑-1</td>
<td>× -1</td>
<td>← -3</td>
<td>← -4</td>
<td>← -5</td>
<td>← -6</td>
<td>← -7</td>
<td>× -5</td>
<td>← -7</td>
<td>← -8</td>
</tr>
<tr>
<td>A</td>
<td>-4</td>
<td>↑-2</td>
<td>× 0</td>
<td>× 0</td>
<td>← -2</td>
<td>← -3</td>
<td>← -4</td>
<td>← -5</td>
<td>← -6</td>
<td>← -7</td>
<td>× -8</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>↑-3</td>
<td>↑-2</td>
<td>↑-2</td>
<td>× 1</td>
<td>← -1</td>
<td>← -2</td>
<td>← -3</td>
<td>← -4</td>
<td>← -5</td>
<td>← -6</td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td>↑-4</td>
<td>↑-3</td>
<td>↑-3</td>
<td>↑-1</td>
<td>× -1</td>
<td>× 0</td>
<td>← -2</td>
<td>← -3</td>
<td>← -4</td>
<td>← -5</td>
</tr>
<tr>
<td>G</td>
<td>-7</td>
<td>↑-5</td>
<td>↑-4</td>
<td>↑-4</td>
<td>↑-2</td>
<td>↑-3</td>
<td>↑-2</td>
<td>× -2</td>
<td>× -1</td>
<td>← -3</td>
<td>← -4</td>
</tr>
<tr>
<td>A</td>
<td>-8</td>
<td>↑-6</td>
<td>↑-5</td>
<td>↑-5</td>
<td>↑-3</td>
<td>↑-4</td>
<td>↑-3</td>
<td>× -1</td>
<td>↑-3</td>
<td>× -3</td>
<td>× -5</td>
</tr>
</tbody>
</table>

**V:** G A A T T C A G T T A

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>−</th>
<th>A</th>
</tr>
</thead>
</table>

**W:** G G A T − C − G − − A

1/25/11

CAP5510/CGS5166
Local Sequence Alignment

- **Example:** comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- **Smith-Waterman Algorithm**
Recurrence Relations (Global vs Local Alignments)

- \( S[I, J] = \text{MAXIMUM} \{ \)
  - \( S[I-1, J-1] + \delta(V[I], W[J]) \),
  - \( S[I-1, J] + \delta(V[I], \_\_\_) \),
  - \( S[I, J-1] + \delta(\_\_, W[J]) \) \}

- Global Alignment

- \( S[I, J] = \text{MAXIMUM} \{ 0, \)
  - \( S[I-1, J-1] + \delta(V[I], W[J]) \),
  - \( S[I-1, J] + \delta(V[I], \_\_\_) \),
  - \( S[I, J-1] + \delta(\_\_, W[J]) \) \}

Global Alignment

Local Alignment
Local Alignment: Example

**Match +1**

**Mismatch −1**

**Gap (-1, -1)**
Properties of Smith-Waterman Algorithm

- How to find all regions of “high similarity”?
  - Find all entries above a threshold score and traceback.

- What if: Matches = 1 & Mismatches/spaces = 0?
  - Longest Common Subsequence Problem

- What if: Matches = 1 & Mismatches/spaces = -\(\infty\)?
  - Longest Common Substring Problem

- What if the average entry is positive?
  - Global Alignment
Calculation of an alignment score

\[ S = \sum \text{(identities, mismatches)} - \sum \text{(gap penalties)} \]

\[ \text{Score} = \text{Max}(S) \]
How to score mismatches?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>L</th>
<th>K</th>
<th>M</th>
<th>F</th>
<th>P</th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>Y</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-4</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>2</td>
<td>-2</td>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>3</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>4</td>
</tr>
</tbody>
</table>

Blosum62 scoring matrix

Slide: Courtesy J. Pevsner
How to score mismatches?
For each amino acid pair $a, b$

- For each BLOCK
  - Align all proteins in the BLOCK
  - Eliminate proteins that are more than $n\%$ identical
  - Count $F(a)$, $F(b)$, $F(a,b)$
  - Compute Log-odds Ratio

\[
\log \left( \frac{F(a,b)}{F(a)F(b)} \right)
\]
### Scoring Matrix to Use

- **PAM 40**: Short alignments with high similarity (70-90%)
- **PAM 160**: Members of a protein family (50-60%)
- **PAM 250**: Longer alignments (divergent sequences) (~30%)
- **BLOSUM90**: Short alignments with high similarity (70-90%)
- **BLOSUM80**: Members of a protein family (50-60%)
- **BLOSUM62**: Finding all potential hits (30-40%)
- **BLOSUM30**: Longer alignments (divergent sequences) (<30%)
Rat versus mouse globin

Rat versus bacterial globin

More conserved

Less conserved

Slide: Courtesy J. Pevsner
BLAST: Steps

- Choose your sequence
- Choose your tool
- Choose your database
- Select parameters, if needed
- Interpret your results
BLAST Variants

- **Nucleotide BLAST**
  - **Standard blastn**
  - **MEGABLAST** (Compare large sets, Near-exact searches)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering)

- **Protein BLAST**
  - **Standard blastp**
  - **PSI-BLAST** (Position Specific Iterated BLAST)
  - **PHI-BLAST** (Pattern Hit Initiated BLAST; reg expr. Or Motif search)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering, PAM-30)

- **Translating BLAST**
  - **Blastx**: Search nucleotide sequence in protein database (6 reading frames)
  - **Tblastn**: Search protein sequence in nucleotide dB
  - **Tblastx**: Search nucleotide seq (6 frames) in nucleotide DB (6 frames)
BLAST Cont’d

- **RPS BLAST**
  - Compare protein sequence against Conserved Domain DB; Helps in predicting rough structure and function

- **Pairwise BLAST**
  - blastp (2 Proteins), blastn (2 nucleotides), tblastn (protein-nucleotide w/ 6 frames), blastx (nucleotide-protein), tblastx (nucleotide w/6 frames-nucleotide w/ 6 frames)

- **Specialized BLAST**
  - Human & Other finished/unfinished genomes
  - *P. falciparum*: Search ESTs, STSs, GSSs, HTGs
  - VecScreen: screen for contamination while sequencing
  - IgBLAST: Immunoglobin sequence database
Databases used by BLAST

- **Protein**
  - nr (everything), swissprot, pdb, alu, individual genomes

- **Nucleotide**
  - nr, dbest, dbsts, htgs (unfinished genomic sequences), gss, pdb, vector, mito, alu, epd

- **Misc**
### BLAST Parameters and Output

- **Type of sequence**, nucleotide/protein
- **Word size**, $w$
- **Gap penalties**, $p_1$ and $p_2$
- **Neighborhood Threshold Score**, $T$
- **Score Threshold**, $S$
- **E-value Cutoff**, $E$
- **Number of hits to display**, $H$
- **Database to search**, $D$
- **Scoring Matrix**, $M$
- **Score** $s$ and **E-value** $e$

  E-value $e$ is the expected number of sequences that would have an alignment score greater than the current score $s$. 
**FIGURE 11.7** The initiation of a BLAST search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words “in the neighborhood” of the original query word. Any occurrences of these neighborhood words in sequences within the target database then are investigated. See text for details.
Find BLAST from the home page of NCBI and select protein BLAST…
Choose align two or more sequences...
Enter the two sequences (as accession numbers or in the fasta format) and click BLAST.

Optionally select “Algorithm parameters” and note the matrix option.
Pairwise alignment result of human beta globin and myoglobin

**Myoglobin RefSeq**

<table>
<thead>
<tr>
<th>Query 4</th>
<th>LTPEEKSAVTALWGKVNVDVG--GEALGRLLVVYPWTQRFSEFGDLSPTDAVMGNPKV 61</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct 3</td>
<td>LSDGEWQLVLNVWGVKEADIPGCHGQVLTFLKPGTPFTLEKFDFKHKLFHDSEMKASEDL 62</td>
</tr>
<tr>
<td>Query 62</td>
<td>KAHGKKVLGAFSDLGLAHLDNKLGTFAUTSELHCDKLVHVDPEFNLGLGVNLVCVLVACHFPK 121</td>
</tr>
<tr>
<td>Sbjct 63</td>
<td>KKHGATVTLGAGILKKGGHEAEIKPQLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG 122</td>
</tr>
<tr>
<td>Query 122</td>
<td>EFTPPVQAAYQKVAVGVALSHKAY 146</td>
</tr>
<tr>
<td>Sbjct 123</td>
<td>DFGADAQGAMNKEALELFKDMASY 147</td>
</tr>
</tbody>
</table>

**Information about this alignment:**
- Score: 47.4 bits (144)
- Expect: 8e-11
- Method: Compositional matrix adjust
- Identities: 37/145 (25%)
- Positives: 57/145 (39%)
- Gaps: 2/145 (1%)

**Query = HBB**

**Subject = MB**

**Middle row displays identities; + sign for similar matches**

**Slide: Courtesy J. Pevsner**
Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores

Score = 18.1 bits \( \boxed{35} \), Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

<table>
<thead>
<tr>
<th></th>
<th>VTALWGVNVD--EVGGEALGRLL</th>
<th>V +WGKV D G E L RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sbjct</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>11</th>
<th>5</th>
<th>6</th>
<th>4</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>sum of matches: +60</th>
</tr>
</thead>
<tbody>
<tr>
<td>match</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mismatch</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-4</td>
<td>0</td>
<td>-4</td>
<td>sum of mismatches: -13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-11</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>gap open</td>
<td>-11</td>
<td>-1</td>
</tr>
<tr>
<td>gap extend</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sum of gap penalties: -12

total raw score: 60 - 13 - 12 = 35

Slide: Courtesy J. Pevsner
Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores

Score = 18.1 bits (35), Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

<table>
<thead>
<tr>
<th>Query</th>
<th>12</th>
<th>VTALWGKVNVD--EVGGEALGRLL</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>11</td>
<td>VLNVWGKVEADIPGHQEVLIRLF</td>
<td>34</td>
</tr>
</tbody>
</table>

|  | 4   | 11  | 5   | 6   | 6   | 5   | 5   | sum of matches: +60 |
|  | 6   | 4   |     |     |     |     |     |                  |

|  | -1  | 1   | 0   | -2  | -2  | -4  | 0   | sum of mismatches: -13 |
|  | -2  | 0   |     | -3  | 0   |     |     |                  |

gap open: -11  gap extend: -1

sum of gap penalties: -12

total raw score: 60 - 13 - 12 = 35

V matching V earns +4
T matching L earns -1

These scores come from a “scoring matrix”!

Slide: Courtesy J. Pevsner
Rules of Thumb

- Most sequences with significant similarity over their entire lengths are homologous.
- Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- A homologous to B & B to C ⇒ A homologous to C.
- Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.
Results of searches using different scoring systems may be compared directly using normalized scores.

If \( S \) is the (raw) score for a local alignment, the normalized score \( S' \) (in bits) is given by

\[
S' = \frac{\lambda - \ln(K)}{\ln(2)}
\]

The parameters depend on the scoring system.

Statistically significant normalized score,

\[
S' > \log\left(\frac{N}{E}\right)
\]

where \( E \)-value = \( E \), and \( N \) = size of search space.
Multiple Alignments

- **Global**
  - ClustalW, ClustalX
  - MSA
  - T-Coffee

- **Local**
  - BLOCKS
  - eMOTIF
  - GIBBS
  - HMMER
  - MACAW
  - MEME

- **Other**
  - Profile Analysis from msa (UCSD)
  - SAM HMM (from msa)
MSA of glyceraldehyde 3-phosphate dehydrogenases: example of high conservation

fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTPG SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKvlisap PKGDEPVKQL VYGVHVDEYD GE.DVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAI TATQKTVGDG SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVGDG SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVGDG SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAI TATQKTVGDG SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVGDG SHKDWRGGRG ASQNIIPSST
archaeon  KVLDEEFGLI AGQLTTTVHAY TGSQNLMGDP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKV I PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKA
bacterium GAAKAVGKV I PELNGKLTGM AFRVPTPNVS VVDLTVRLLEK AATYEQIAAA
yeast     GAAKAVGKVI PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNA
Multiple Alignments: CLUSTALW

* identical
: conserved substitutions
. semi-conserved substitutions

<table>
<thead>
<tr>
<th>gi</th>
<th>Sequence</th>
<th>Alignment</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>2213819</td>
<td>CDN-ELKSEAIEHLCASEFALR---------MKIKEVKKENGDKK</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>12656123</td>
<td>----ELKSEAIEHLCASEFALR--------MKIKEVKKENG-</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>7512442</td>
<td>CKNKNDDDNDIMETLCNDFALK---------IKVKEITYINRTK</td>
<td>211</td>
<td></td>
</tr>
<tr>
<td>1344282</td>
<td>QDECKFDYVEVETYSSGAFLGRVFCGAEPPPHLVSSHELAVLFRTDH</td>
<td>400</td>
<td></td>
</tr>
</tbody>
</table>

Red: AVFPMLW (Small & hydrophobic)
Blue: DE (Acidic)
Magenta: RHK (Basic)
Green: STYHCNGQ (Hydroxyl, Amine, Basic)
Gray: Others

Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physico-chemical properties thought to be important in the determination of protein structure.
Multiple Alignment

A. Estimate the amino acid frequencies in the motif columns of all but one sequence. Also obtain background.

```
xxxMxxxxx  xxxMxxxxx
xxxxxMxx   xxxxxxxMxx
xxxxxxMxx  xxxxxxxxMxx
xMxxxxxx   xMxxxxxxxx
xxxxxxxxx  xxxxxxxxxxx
xxxxxMxxxx  MxxxxxxMxxx
xxxxxMxxx  xxxxxxxMxxx
xMxxxxxx   xMxxxxxxxx
xxxxxxxxx  xxxxxxxxxM
```

Random start positions chosen

Location of motif in each sequence provides first estimate of motif composition
How to Score Multiple Alignments?

- **Sum of Pairs Score (SP)**
  - Optimal alignment: $O(d^N)$ [Dynamic Prog]
  - Approximate Algorithm: **Approx Ratio 2**
    - Locate Center: $O(d^2N^2)$
    - Locate Consensus: $O(d^2N^2)$

*Consensus char*: char with min distance sum
*Consensus string*: string of consensus char
*Center*: input string with min distance sum
Multiple Alignment Methods

- **Phylogenetic Tree Alignment** *(NP-Complete)*
  - Given tree, task is to label leaves with strings

- **Iterative Method(s)**
  - Build a MST using the distance function

- **Clustering Methods**
  - Hierarchical Clustering
  - K-Means Clustering
Multiple Alignment Methods (Cont’d)

- **Gibbs Sampling Method**

- **Hidden Markov Model**
Multiple Sequence Alignments (MSA)

- **Choice of Scoring Function**
  - Global vs local
  - Gap penalties
  - Substitution matrices
  - Incorporating other information
  - Statistical Significance

- **Computational Issues**
  - Exact/heuristic/approximate algorithms for optimal MSA
  - Progressive/Iterative/DP
  - Iterative: Stochastic/Non-stochastic/Consistency-based

- **Evaluating MSAs**
  - Choice of good test sets or benchmarks (BAliBASE)
  - How to decide thresholds for good/bad alignments
Progressive MSA: CLUSTALW

This example shows how a progressive alignment strategy can be misled. In the initial alignment of sequences 1 and 2, ClustalW has a choice between aligning CAT with CAT and making an internal gap or making a mismatch between C and F and having a terminal gap. Since terminal gaps are much cheaper than internals, the ClustalW scoring schemes prefers the former. In the next stage, when the extra sequence is added, it turns out that properly aligning the two CATs in the previous stage would have led to a better scoring sums-of-pairs multiple alignment.

C. Notredame, Pharmacogenomics, 3(1), 2002.
Software for MSA

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Website</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>Exact</td>
<td><a href="http://www.ibc.wustl.edu/ibc/msa.html">http://www.ibc.wustl.edu/ibc/msa.html</a></td>
<td>[28]</td>
</tr>
<tr>
<td>OMA</td>
<td>Iterative DCA</td>
<td><a href="http://bibiserv.techfak.uni-bielefeld.de/oma">http://bibiserv.techfak.uni-bielefeld.de/oma</a></td>
<td>[61]</td>
</tr>
<tr>
<td>ComAlign</td>
<td>Consistency-based</td>
<td><a href="http://www.daimi.au.dk/~ocaprani">http://www.daimi.au.dk/~ocaprani</a></td>
<td>[75]</td>
</tr>
<tr>
<td>Praline</td>
<td>Iterative/progressive</td>
<td><a href="mailto:jhering@nimr.mrc.ac.uk">jhering@nimr.mrc.ac.uk</a></td>
<td>[48]</td>
</tr>
<tr>
<td>HMMER</td>
<td>Iterative/Stochastic/HMM</td>
<td><a href="http://hmmer.wustl.edu/">http://hmmer.wustl.edu/</a></td>
<td>[68]</td>
</tr>
<tr>
<td>GA</td>
<td>Iterative/Stochastic/GA</td>
<td><a href="mailto:czhang@watnew.uwaterloo.ca">czhang@watnew.uwaterloo.ca</a></td>
<td>[52]</td>
</tr>
</tbody>
</table>

MSA: Conclusions

- Very important
  - Phylogenetic analyses
  - Identify members of a family
  - Protein structure prediction

- No perfect methods

- Popular
  - Progressive methods: **CLUSTALW**
  - Recent interesting ones: Prrp, SAGA, DiAlign, T-Coffee

  - **CLUSTALW** works reasonably well, in general
  - **DiAlign** is better for sequences with long insertions & deletions (indels)
  - **T-Coffee** is best available method