BSC 4934: Q’BIC Capstone Workshop

Giri Narasimhan
ECS 254A; Phone: x3748
giri@cis.fiu.edu
http://www.cis.fiu.edu/~giri/teach/BSC4934_Su09.html
24 June through 7 July, 2009
Types of Sequence Alignments - 1

- **Global Alignment**: similarity over entire length

- **Local Alignment**: no overall similarity, but some segment(s) is/are similar
Types of Sequence Alignments - 2

- **Semi-global Alignment**: end segments may not be similar

- **Multiple Alignment**: similarity between sets of sequences
Sequence Alignment

- **Global:**

- **Local:**
  - Smith-Waterman (1981)
  - Useful when commonality is small and global alignment is meaningless. Often unaligned portions “mask” short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.

- **Dynamic Programming (DP) based.**
Why gaps?

- **Example**: Finding the gene site for a given (eukaryotic) cDNA requires “gaps”.
- **What is cDNA?** cDNA = Copy DNA

---

**Diagram:**

- DNA
  - Transcription
  - mRNA
  - Translation
- cDNA
  - Reverse Transcription
  - Protein
How to score mismatches?

![BLOSUM 62 Matrix](image)
BLAST & FASTA

- FASTA
  [Lipman, Pearson ’85, ’88]
- Basic Local Alignment Search Tool
  [Altschul, Gish, Miller, Myers, Lipman ’90]
<table>
<thead>
<tr>
<th>BLAST Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Program(s) to search all sequence databases</td>
</tr>
<tr>
<td>- Tremendous Speed/Less Sensitive</td>
</tr>
<tr>
<td>- Statistical Significance reported</td>
</tr>
<tr>
<td>- WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTc13 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)</td>
</tr>
</tbody>
</table>
Lipman et al.: speeded up finding “runs” of “hot spots”.

Eugene Myers ’94: “Sublinear algorithm for approximate keyword matching”.

Karlin, Altschul, Dembo ’90, ’91: “Statistical Significance of Matches”
Example: Aligning HIV sequences.
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: Immunoglobulin sequence database
BLAST Credits

- Stephen Altschul
- Jonathan Epstein
- David Lipman
- Tom Madden
- Scott McGinnis
- Jim Ostell
- Alex Schaffer
- Sergei Shavirin
- Heidi Sofia
- Jinghui Zhang
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 \( s \).
## Scoring Matrix to Use

<table>
<thead>
<tr>
<th>Scoring Matrix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 40</td>
<td>Short alignments with high similarity (70-90%)</td>
</tr>
<tr>
<td>PAM 160</td>
<td>Members of a protein family (50-60%)</td>
</tr>
<tr>
<td>PAM 250</td>
<td>Longer alignments (divergent sequences) (~30%)</td>
</tr>
<tr>
<td>BLOSUM90</td>
<td>Short alignments with high similarity (70-90%)</td>
</tr>
<tr>
<td>BLOSUM80</td>
<td>Members of a protein family (50-60%)</td>
</tr>
<tr>
<td>BLOSUM62</td>
<td>Finding all potential hits (30-40%)</td>
</tr>
<tr>
<td>BLOSUM30</td>
<td>Longer alignments (divergent sequences) (&lt;30%)</td>
</tr>
</tbody>
</table>
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.
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

\[ S' = (\lambda S - \ln K)/\ln 2 \]

The parameter \( \lambda \) scales for the scoring system, while K scales for the search space size.

**Statistically significant normalized score**,

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

where E-value = E, and N = size of search space.

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.
Types of Sequence Alignments

- **Global**
  - HIV Strain 1
  - HIV Strain 2

- **Local**

- **Semi-Global**

- **Multiple**
  - Strain 1
  - Strain 2
  - Strain 3
  - Strain 4

06/29/09  Q'BIC Bioinformatics
Global Alignment:
An example

\[ \text{V: G A A T T C A G T T A} \]
\[ \text{W: G G A T C G A} \]

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

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

**Recurrence Relation**
\[ S[I, J] = \text{MAXIMUM} \{ \\
S[I-1, J-1] + \delta(V[I], W[J]), \\
S[I-1, J] + \delta(V[I], \text{-}), \\
S[I, J-1] + \delta(\text{-}, W[J]) \} \]
Global Alignment: An example

\[ S[I, J] = \text{MAXIMUM} \left\{ \begin{array}{c}
S[I-1, J-1] + \delta(V[I], W[J]), \\
S[I-1, J] + \delta(V[I], \_), \\
S[I, J-1] + \delta(\_, W[J]) \end{array} \right\} \]
### Traceback

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

#### W: G G A – T C – G – – A

---

**06/29/09**

Q'BIC Bioinformatics
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

Previous
## 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>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></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></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>×3</td>
<td></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></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></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>×5</td>
<td>←5</td>
<td>←5</td>
<td></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>↑4</td>
<td>×5</td>
<td>↑5</td>
<td>↑5</td>
<td></td>
</tr>
</tbody>
</table>

06/29/09  Q'BIC Bioinformatics  26
## 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>G</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>0</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></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>2</td>
<td>2</td>
<td></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>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>3</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>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

06/29/09   Q'BIC Bioinformatics   27
## Improved Traceback

### V: G A - A T T C A G T T A

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</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>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### W: G - G A - T C - G - A

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</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>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table:

<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>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>
<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>↑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>↑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>
<td></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>
<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>
<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>↑4</td>
<td>×5</td>
<td>↑5</td>
<td>↑5</td>
<td>↑5</td>
<td>×6</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 = $\alpha$
- Spaces (Insertions/Deletions, InDels) = $\beta$
- Affine Gap Penalties:
  - (Gap open, Gap extension) = ($\gamma, \delta$)
- Weighted Mismatch = $\Phi(a,b)$
- Weighted Matches = $\Omega(a)$
## Alternative Scoring Schemes

<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>× -7</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>

**Match +1**  
**Mismatch −2**  
**Gap (−2, −1)**

<table>
<thead>
<tr>
<th>V:</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W:</th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>–</th>
<th>C</th>
<th>–</th>
<th>G</th>
<th>–</th>
<th>–</th>
<th>A</th>
</tr>
</thead>
</table>

06/29/09  
Q'BIC Bioinformatics
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] = MAXIMUM** {
  
  \[S[I-1, J-1] + \delta(V[I], W[J]),\]
  
  \[S[I-1, J] + \delta(V[I], \_\_),\]
  
  \[S[I, J-1] + \delta(\_\_, W[J])\] }

- **S[I, J] = 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

<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>-</td>
<td>G</td>
<td>A</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>W:</td>
<td>G</td>
<td>G</td>
<td>-</td>
<td>A</td>
<td>T</td>
<td>-</td>
<td>C</td>
<td>-</td>
<td>G</td>
<td>-</td>
<td>--</td>
</tr>
</tbody>
</table>
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
How to score mismatches?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>9</td>
<td>-3</td>
<td>-4</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>D</td>
<td>-2</td>
<td>-3</td>
<td>6</td>
<td>2</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>E</td>
<td>-1</td>
<td>-4</td>
<td>2</td>
<td>5</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>6</td>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*BLOSUM 62*
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

$$\log \left( \frac{F(a,b)}{F(a)F(b)} \right)$$
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)
Multiple Alignments: CLUSTALW

* identical
: conserved substitutions
. semi-conserved substitutions

<table>
<thead>
<tr>
<th>gi</th>
<th>Alignment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2213819</td>
<td>gi</td>
<td>2213819 CDN-ELKSEAIIEHLCASEFALR-------------MKIKEVKKENGDKK</td>
</tr>
<tr>
<td>12656123</td>
<td>gi</td>
<td>12656123 -----ELKSEAIIEHLCASEFALR-------------MKIKEVKKENG</td>
</tr>
<tr>
<td>7512442</td>
<td>gi</td>
<td>7512442 CRNKNDDDNDIMETLCKNDFALK-------------IKVKEITYINRDTK</td>
</tr>
<tr>
<td>1344282</td>
<td>gi</td>
<td>1344282 QDECKFDYVEVYETSSSSGAFSLLGRFCGAEPPLVSSHHHELAVLFRTDH</td>
</tr>
</tbody>
</table>

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

- **Family alignment for the ITAM domain** (Immunoreceptor tyrosine-based activation motif)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD3D_MOUSE/1-2</td>
<td>EQLYQPLRDR EDTQ-YSRLG GN</td>
</tr>
<tr>
<td>Q90768/1-21</td>
<td>DQLYQPLGER NDGQ-YSQLA TA</td>
</tr>
<tr>
<td>CD3G_SHEEP/1-2</td>
<td>DQLYQPKER EDDQ-YSHLR KK</td>
</tr>
<tr>
<td>P79951/1-21</td>
<td>NDLYQPLGQR SEDT-YSHLN SR</td>
</tr>
<tr>
<td>FCEG_CAVPO/1-2</td>
<td>DGIYTGLSTR NQET-YETLK HE</td>
</tr>
<tr>
<td>CD3Z_HUMAN/3-0</td>
<td>DGLYQGLSTA TKDT-YDALH MQ</td>
</tr>
<tr>
<td>C79A_BOVIN/1-2</td>
<td>ENLYEGLNLD DCSM-YEDIS RG</td>
</tr>
<tr>
<td>C79B_MOUSE/1-2</td>
<td>DHTYEGLNID QTAT-YDIV TL</td>
</tr>
<tr>
<td>CD3H_MOUSE/1-2</td>
<td>NQLYNELNLG RREE-YDVLE KK</td>
</tr>
<tr>
<td>CD3Z_SHEEP/1-2</td>
<td>NPVYNELNGV RREE-YAVLD RR</td>
</tr>
<tr>
<td>CD3E_HUMAN/1-2</td>
<td>NPDYEPIRKG QRDLYSGLN QR</td>
</tr>
<tr>
<td>CD3H_MOUSE/2-0</td>
<td>EGYNALQKD KMAEAYSEIG TK</td>
</tr>
<tr>
<td>Consensus/60%</td>
<td>-.LYpsLspc pcsp.YspLs pp</td>
</tr>
</tbody>
</table>
Multiple Alignment

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

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.

Software for MSA

Table 1. Some recent and less recent available methods for MSAs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Website/Contact</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> [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> [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> [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> [48]</td>
</tr>
<tr>
<td>HMMER</td>
<td>Iterative/Stochastic/HMM</td>
<td><a href="http://hmmer.wustl.edu/">http://hmmer.wustl.edu/</a> [68]</td>
</tr>
<tr>
<td>GA</td>
<td>Iterative/Stochastic/GA</td>
<td><a href="mailto:czhang@watnow.uwaterloo.ca">czhang@watnow.uwaterloo.ca</a> [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

- Review of Methods [C. Notredame, Pharmacogenomics, 3(1), 2002]
  - CLUSTALW works reasonably well, in general
  - DiAlign is better for sequences with long insertions & deletions (indels)
  - T-Coffee is best available method