Types of Sequence Alignments

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

- **Local**

- **Semi-Global**

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

1/25/07 CAP5510
### 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>
</tbody>
</table>

- **G**
  - -2  \(\times 1\) \(\leftarrow -1\) \(\leftarrow -2\) \(\leftarrow -3\) \(\leftarrow -4\) \(\leftarrow -5\) \(\leftarrow -6\) \(\leftarrow -7\) \(\leftarrow -8\) \(\leftarrow -9\) \(\leftarrow -10\)

- **A**
  - -4  \(\uparrow 2\) \(\times 0\) \(\times 0\) \(\leftarrow -2\) \(\leftarrow -3\) \(\leftarrow -4\) \(\leftarrow -5\) \(\leftarrow -6\) \(\leftarrow -7\) \(\leftarrow -8\) \(\times -7\)

- **T**
  - -5  \(\uparrow 3\) \(\uparrow 2\) \(\uparrow 2\) \(\times 1\) \(\leftarrow -1\) \(\leftarrow -2\) \(\leftarrow -3\) \(\leftarrow -4\) \(\leftarrow -5\) \(\leftarrow -6\) \(\leftarrow -7\)

- **C**
  - -6  \(\uparrow 4\) \(\uparrow 3\) \(\uparrow 3\) \(\uparrow 1\) \(\times -1\) \(\times 0\) \(\leftarrow -2\) \(\leftarrow -3\) \(\leftarrow -4\) \(\leftarrow -5\) \(\leftarrow -6\)

- **G**
  - -7  \(\uparrow 5\) \(\uparrow 4\) \(\uparrow 4\) \(\uparrow 2\) \(\uparrow 3\) \(\uparrow 2\) \(\times -2\) \(\times -1\) \(\leftarrow -3\) \(\leftarrow -4\) \(\leftarrow -5\)

- **A**
  - -8  \(\uparrow 6\) \(\uparrow 5\) \(\uparrow 5\) \(\uparrow 3\) \(\uparrow 4\) \(\uparrow 3\) \(\times -1\) \(\uparrow 3\) \(\times -3\) \(\times -5\) \(\times -3\)

**Match +1**

**Mismatch –2**

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

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

**W:** G G A T - C - G - - A

1/25/07 CAP5510
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)

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

- $S[I, J] = \text{MAXIMUM} \{ 0, 
  S[I-1, J-1] + \delta(V[I], W[J]), 
  S[I-1, J] + \delta(V[I], --), 
  S[I, J-1] + \delta(--, W[J]) \}$
## 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>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>G</td>
<td>0</td>
<td>×1</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>×2</td>
<td>×1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>×1</td>
<td>0</td>
<td>0</td>
<td>×1</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>↑0</td>
<td>×1</td>
<td>×2</td>
<td>←1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>×1</td>
<td>×1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>↑0</td>
<td>×0</td>
<td>×2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<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>×1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>×1</td>
<td>×1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>×1</td>
<td>0</td>
<td>0</td>
<td>×1</td>
</tr>
</tbody>
</table>

**Match +1**

**Mismatch −1**

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

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

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

1/25/07  CAP5510  6
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?

![Matrix for scoring mismatches]
BLOSUM n Substitution Matrices

- 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

\[
\log\left( \frac{F(a,b)}{F(a)F(b)} \right)
\]
BLAST & FASTA

- FASTA
  [Lipman, Pearson ‘85, '88]
- Basic Local Alignment Search Tool
  [Altschul, Gish, Miller, Myers, Lipman ‘90]
BLAST Overview

- Program(s) to search all sequence databases
- Tremendous Speed/Less Sensitive
- Statistical Significance reported
- WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTcl3 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)
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”
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**
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.