Types of Sequence Alignments

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

- **Local**
  - Strain 1
  - Strain 2
  - Strain 3
  - Strain 4

- **Semi-Global**

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

1/20/05

CAP5510/CGS5166 (Lec 4)
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
\[ S[I, J] = \text{Score of Matching First } I \text{ characters of sequence V & First } J \text{ characters of sequence W} \]

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

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

V: G A A 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></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>↑4</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>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>\times 1</td>
<td>\leftarrow 1</td>
<td>\leftarrow 1</td>
<td>\leftarrow 1</td>
<td>\leftarrow 1</td>
<td>\leftarrow 1</td>
<td>\times 1</td>
<td>\leftarrow 1</td>
<td>\leftarrow 1</td>
<td>\leftarrow 1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>\times 1</td>
<td>\uparrow 1</td>
<td>\uparrow 1</td>
<td>\uparrow 1</td>
<td>\uparrow 1</td>
<td>\uparrow 1</td>
<td>\times 2</td>
<td>\leftarrow 2</td>
<td>\leftarrow 2</td>
<td>\leftarrow 2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>\uparrow 1</td>
<td>\uparrow 1</td>
<td>\times 2</td>
<td>\leftarrow 2</td>
<td>\leftarrow 2</td>
<td>\times 2</td>
<td>\uparrow 2</td>
<td>\uparrow 2</td>
<td>\uparrow 2</td>
<td>\times 3</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>\uparrow 1</td>
<td>\leftarrow 2</td>
<td>\uparrow 2</td>
<td>\times 3</td>
<td>\times 3</td>
<td>\leftarrow 3</td>
<td>\leftarrow 3</td>
<td>\times 3</td>
<td>\times 3</td>
<td>\uparrow 3</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>\uparrow 1</td>
<td>\uparrow 2</td>
<td>\uparrow 2</td>
<td>\uparrow 3</td>
<td>\uparrow 3</td>
<td>\times 4</td>
<td>\leftarrow 4</td>
<td>\leftarrow 4</td>
<td>\leftarrow 4</td>
<td>\leftarrow 4</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>\uparrow 1</td>
<td>\uparrow 2</td>
<td>\uparrow 2</td>
<td>\uparrow 3</td>
<td>\uparrow 3</td>
<td>\uparrow 4</td>
<td>\uparrow 4</td>
<td>\times 5</td>
<td>\leftarrow 5</td>
<td>\leftarrow 5</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>\uparrow 1</td>
<td>\uparrow 2</td>
<td>\times 3</td>
<td>\uparrow 3</td>
<td>\uparrow 3</td>
<td>\uparrow 4</td>
<td>\times 5</td>
<td>\uparrow 5</td>
<td>\uparrow 5</td>
<td>\times 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>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>
<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>
<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>
<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>
<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>↑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></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>

V: G A - A T T C A G T T A
W: G - G A - T C - G - - A
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:
  \[(\text{Gap open, Gap extension}) = (\gamma, \delta)\]
- 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>←-2</td>
<td>←-3</td>
<td>←-4</td>
<td>←-5</td>
<td>←-6</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>

**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
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)

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

- $S[I, J] = \text{MAXIMUM} \{ 0,$
  
  $S[I-1, J-1] + \delta(V[I], W[J]),$
  
  $S[I-1, J] + \delta(V[I], \rightarrow),$
  
  $S[I, J-1] + \delta(\rightarrow, W[J]) \} \quad \text{Local Alignment}$
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>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>G</td>
<td>0</td>
<td>×1</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>×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>×1</td>
<td>×1</td>
<td>0</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
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?

\[
\begin{array}{cccccccc}
A & C & D & E & F & G & H \\
A & 4 & 0 & -2 & -1 & -2 & 0 & -2 \\
C & 0 & 9 & -3 & -4 & -2 & -3 & -3 \\
D & -2 & -3 & 6 & 2 & -3 & -1 & -1 \\
E & -1 & -4 & 2 & 5 & -3 & -2 & 0 \\
F & -2 & -2 & -3 & -3 & 6 & -3 & -1 \\
G & 0 & -3 & -1 & -2 & -3 & -3 & 0 \\
H & -2 & -3 & -1 & 0 & -2 & -3 & -3 & 0 \\
\end{array}
\]

\textit{BLOSUM 62}
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

\[
\log\left( \frac{F(a,b)}{F(a)F(b)} \right)
\]
String Matching Problem

Pattern $P$ → Set of Locations $L$

Text $T$ →
(Approximate) String Matching

**Input:** Text \( T \), Pattern \( P \)

**Question(s):**
- Does \( P \) occur in \( T \)?
- Find one occurrence of \( P \) in \( T \).
- Find all occurrences of \( P \) in \( T \).
- Count # of occurrences of \( P \) in \( T \).
- Find longest substring of \( P \) in \( T \).
- Find closest substring of \( P \) in \( T \).
- Locate direct repeats of \( P \) in \( T \).

*Many More variants*

**Applications:**
- Is \( P \) already in the database \( T \)?
- Locate \( P \) in \( T \).
- Can \( P \) be used as a primer for \( T \)?
- Is \( P \) homologous to anything in \( T \)?
- Has \( P \) been contaminated by \( T \)?
- Is \( \text{prefix}(P) = \text{suffix}(T) \)?
- Locate tandem repeats of \( P \) in \( T \).
**Input:** Text $T$; Pattern $P$

**Output:** All occurrences of $P$ in $T$.

**Methods:**
- Naïve Method
- Rabin-Karp Method
- FSA-based method
- Knuth-Morris-Pratt algorithm
- Boyer-Moore
- Suffix Tree method
- Shift-And method
Naive Strategy
Finite State Automaton

Finite State Automaton

ATAQAANANASPVANAGVERANANANESISITALVDANANANANANAS
### State Transition Diagram

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>S</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>AN</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANA</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>ANAN</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANANA</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>ANANAS</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note: The table represents transitions in a state diagram.*
**Input:** Text $T$; Pattern $P$

**Output:** All occurrences of $P$ in $T$.

**Sliding Window Strategy:**

Initialize window on $T$;

While (window within $T$) do
  Scan: if (window = $P$) then report it;
  Shift: shift window to right (by ?? positions)

endwhile;
Tries

Storing:
BIG
BIGGER
BILL
GOOD
GOSH

In this figure, the strings either start with B or G. Therefore, the root of the trie is connected to 3 edges called B, G and $.

LEAVES ARE GREEN.
THE SYMBOL "$" TERMINATES EACH WORD.
Suffix Tries & Compact Suffix Tries

Store all suffixes of GOOGOL$
Suffix Tries to Suffix Trees

COMPACT TRIE OF SUFFIXES OF THE TEXT: $GOOGOL$

SUFFIX TREE

Key:  G O O G O L $  
      1 2 3 4 5 6 7
Suffix Trees

- **Linear**-time construction!
- String Matching, Substring matching, substring common to k of n strings
- All-pairs prefix-suffix problem
- Repeats & Tandem repeats
- Approximate string matching