Given a query sequence, $Q$, of length $m$ and database, $D$, of length $n$, find related database sequences

... find related database sequences
... find sequences with MSPs with score at least $S_T$

Terminology

**Segment Pair (SP):** ungapped local alignment.

**Maximal Segment Pair (MSP):** an ungapped local alignment that cannot be improved by making it bigger or smaller.

**High-Scoring Segment Pair (HSP):** an MSP with score at least $S_T$, where $S_T$ is a user-defined minimum.

**Word:** String of length $w$. Typically, $w < 10$.

**Hit:** An ungapped alignment of a word in $Q$ and a word in $D$ with score at least $T$.

Ungapped BLAST

*Altschul et al, 90*

0. Select $S$ based on desired $E$
1. Construct $L$: a list of words of length $w$ with score $\geq T$
   
   Choose $w = 3$, $T = 13$ empirically (1997)

2. Scan database $D$ for hits – instances of words in $L$
3. Extend hits to find MSPs with score $> S_T$

   Stop extension if ungapped score drops below $X$
Sensitivity and Specificity

False positives and false negatives depend on the overlap of the distributions of:
- Chance MSPs
- MSPs in related sequences

Blast90 parameters

False Positives: Unrelated sequences with score ≥ $S_T$
False Negatives:
1. Related sequences with score < $S_T$
2. Related sequences that do not contain a hit

Increased running time:
Hits that are not in a high scoring segment pair.

Select $w$ and $T$ to get the best balance between false negatives and efficient running time.

### Table 1

<table>
<thead>
<tr>
<th>$w$</th>
<th>$T$</th>
<th>Probability of a hit</th>
<th>Related MSPs</th>
<th>Decreased $e$ of MSP hits by blast where $S$ equals</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>11</td>
<td>0.029</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>0.027</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>0.024</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>0.020</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>0.018</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>0.016</td>
<td>0.18</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>0.013</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>0.010</td>
<td>0.185</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>0.007</td>
<td>0.15</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Roughly the same FN rate
Problems with Blast 90

Two many unnecessary extensions:

1. Construct \( L \): a list of words of length \( w \) with score \( \geq T \)
   Choose \( w = 3, \ T = 13 \) empirically (1997)
2. Scan database \( D \) for hits – instances of words in \( L \)
3. Extend hits to find MSPs with score \( > S_T \)
   Stop extension if ungapped score drops below \( X \)

Problems with BLAST 90

Only finds ungapped alignments.
A possible solution: find several HSP’s and merge them:

- However, if \( P(S\&S') \) is significant but neither \( P(S) \), nor \( P(S') \)
  is significant alone, we need to find hits in both \( S \) and \( S' \) to
  find this alignment.
- We could increase probability that both \( S \) and \( S' \) are found
  by decreasing word threshold to \( T=11 \).
- But, this will increase the number of hits found in step two
  and the number of unnecessary extensions in step 3.

1. Gapped Extensions

- Find MSPs using ungapped extensions
  Stop if ungapped alignment score drops below \( X \)
- If MSP score exceeds threshold for gapped extension, \( S_T \)
  Perform a gapped extension.
  Stop if gapped alignment score drops below \( Xg \)
- Report match if gapped alignment score exceeds threshold, \( S_T \)

\[ S_{g}: \text{gapped extension threshold} \quad X: \text{ungapped extension cutoff} \]
\[ X_g: \text{gapped extension cutoff} \]
2. Two-Hit BLAST

- Reduce threshold $T$ to obtain more hits
- Only trigger an ungapped extension if there are two hits on the same diagonal within distance $A$

Two Hit BLAST

- Reduce threshold $T$ to obtain more hits
- Only trigger an ungapped extension if there are two hits on the same diagonal within distance $A$
**Putting it all together**

1. Find hits of length $w$ with similarity threshold $T$.
2. If there are two word pairs on same diagonal separated by a distance of at most $A$, perform an *ungapped* extension using cutoff, $X_1$.
3. If MSP score $S_T > S_g$, perform a gapped extension with dynamic programming with cutoff $X_2$.
4. If gapped extension score $S_2 > S_T$, report match.

**Comparing BLAST parameters**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>BLAST 90</th>
<th>BLAST 97</th>
</tr>
</thead>
<tbody>
<tr>
<td>matrix</td>
<td>$S[i,j]$</td>
<td>$S[i,j]$</td>
</tr>
<tr>
<td>threshold</td>
<td>$S_T$</td>
<td>$S_g$, $S_T$</td>
</tr>
<tr>
<td>word size</td>
<td>$w=4$</td>
<td>$w=3$</td>
</tr>
<tr>
<td>word threshold</td>
<td>$T=17$</td>
<td>$T=11$</td>
</tr>
<tr>
<td>max distance btw hits</td>
<td>$A=40$</td>
<td></td>
</tr>
<tr>
<td>extension cutoff</td>
<td>$X$</td>
<td>$X_1$, $X_2$</td>
</tr>
</tbody>
</table>

**Length adjustment**

Given a query sequence of length $m$ and database of length $n$, to a first approximation, the size of the search space is $mn$.

However, the database is not a single sequence but a set of concatenated sequences. An alignment cannot bridge the boundary between two database sequences. Furthermore, an alignment cannot start to close to the end of the query sequence. Both of these reduce the actual size of the search space.

*On average*, an alignment must start within the gray box to accrue a score of at least $S_T$ before reaching the end of the sequence.