CS 466
Introduction to Bioinformatics

Instructor: Jian Peng
Teaching Assistant: Wesley Qian
1. Break query sequence into words

Break query into words:

MEAA
EAA
AAV
AVK
VKE
KEE
EEI
EIS
ISV
...

MEAVKEEI SVDEAVDKNI
2. Find database hits

- Find exact matches to query words
- Can be done in efficiently
  - Hashing
  - Alternatively AC finite state machine

ATC $\rightarrow$ h $\rightarrow$ address1 $\xrightarrow{\text{retrieve}}$ \{1, 6, 100, 2000, 5454, ..., \}

AAA $\rightarrow$ h $\rightarrow$ address2 $\xrightarrow{\text{retrieve}}$ \{15, 21, 30, 785, 3434, ..., \}

TTC $\rightarrow$ h $\rightarrow$ address3 $\xrightarrow{\text{retrieve}}$ \{5, 164, 220, 502, 943, ..., \}
2. Find database hits
3. Extend hits

1. Find “seeds” (initial matches) of a fixed length (e.g. 11)
2. Try extending an alignment from each seed
How to handle possible mismatches in words?

**Query word, W=3 for proteins 
(W=11 for nucleotides)**

<table>
<thead>
<tr>
<th>Word</th>
<th>Score (BL-62)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSK</td>
<td>15</td>
</tr>
<tr>
<td>GAK</td>
<td>12</td>
</tr>
<tr>
<td>GNK</td>
<td>12</td>
</tr>
<tr>
<td>GTK</td>
<td>12</td>
</tr>
<tr>
<td>GSK</td>
<td>12</td>
</tr>
<tr>
<td>GDK</td>
<td>11</td>
</tr>
<tr>
<td>GQK</td>
<td>11</td>
</tr>
<tr>
<td>GEK</td>
<td>11</td>
</tr>
<tr>
<td>GGK</td>
<td>11</td>
</tr>
<tr>
<td>GKK</td>
<td>11</td>
</tr>
<tr>
<td>GSQ</td>
<td>11</td>
</tr>
<tr>
<td>GSE</td>
<td>11</td>
</tr>
</tbody>
</table>

**Neighbor words**
How to handle possible mismatches in words?

First step:
For each position $p$ of the query, find the list of words of length $w$ scoring more than $T$ when paired with the word starting at $p$:

List of words of length $w$, scoring more than $T$ with the $p$–word.
Second step:

For each words list, identify all exact matches with DB sequences:
How to handle possible mismatches in words?

Third step:
For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.

HSP = High Scoring Segment Pair
Parameters

• Word length $w$: 3 for protein, 11 for DNA/RNA
• Thresholds $T$ and $S$:
  • BLAST minimizes time spent on database sequences whose similarity with the query has little chance of exceeding this cutoff $S$.
  • Main strategy: seek only segment pairs (one from database, one query) that contain a word pair with score $\geq T$
  • Intuition: If the sequence pair has to score above $S$, its most well matched word (of some predetermined small length) must score above $T$
  • Lower $T$ => Fewer false negatives
  • Lower $T$ => More pairs to analyze
Parameters

- Larger values of \( w \) increases the number of neighborhood words, but decreases the number of chance matches in the database.
  - Increasing \( w \) decreases sensitivity.

- Larger values of \( T \) decrease the overall execution time, but increase the chance of missing a MSP having score \( \geq S \).
  - Increases \( T \) decreases the sensitivity.

- Larger values of \( S \) increase the specificity. The value of \( S \) is affected by changes in the expectation value parameter.
### Table 3. Sensitivity and Specificity of Single Perfect Nucleotide K-mer Matches as a Search Criterion

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81%</td>
<td>0.974</td>
<td>0.915</td>
<td>0.833</td>
<td>0.726</td>
<td>0.607</td>
<td>0.486</td>
<td>0.373</td>
<td>0.314</td>
</tr>
<tr>
<td>83%</td>
<td>0.988</td>
<td>0.953</td>
<td>0.897</td>
<td>0.815</td>
<td>0.711</td>
<td>0.595</td>
<td>0.478</td>
<td>0.415</td>
</tr>
<tr>
<td>85%</td>
<td>0.996</td>
<td>0.978</td>
<td>0.945</td>
<td>0.888</td>
<td>0.808</td>
<td>0.707</td>
<td>0.594</td>
<td>0.532</td>
</tr>
<tr>
<td>87%</td>
<td>0.999</td>
<td>0.992</td>
<td>0.975</td>
<td>0.942</td>
<td>0.888</td>
<td>0.811</td>
<td>0.714</td>
<td>0.659</td>
</tr>
<tr>
<td>89%</td>
<td>1.000</td>
<td>0.998</td>
<td>0.991</td>
<td>0.976</td>
<td>0.946</td>
<td>0.897</td>
<td>0.824</td>
<td>0.782</td>
</tr>
<tr>
<td>91%</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
<td>0.993</td>
<td>0.981</td>
<td>0.956</td>
<td>0.912</td>
<td>0.886</td>
</tr>
<tr>
<td>93%</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.987</td>
<td>0.968</td>
<td>0.957</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
<td>0.994</td>
<td>0.991</td>
<td></td>
</tr>
<tr>
<td>97%</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td><strong>B.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>1.3e+07</td>
<td>2.9e+06</td>
<td>635783</td>
<td>143051</td>
<td>32512</td>
<td>7451</td>
<td>1719</td>
<td>399</td>
</tr>
</tbody>
</table>

(A) Columns are for K sizes of 7–14. Rows represent various percentage identities between the homologous sequences. The table entries show the fraction of homologies detected as calculated from equation 3 assuming a homologous region of 100 bases. The larger the value of K, the fewer homologies are detected.

(B) K represents the size of the perfect match. F shows how many perfect matches of this size expected to occur by chance according to equation 4 in a genome of 3 billion bases using a query of 500 bases.
Choosing word length \( w \)

### Table 4. Sensitivity and Specificity of Single Perfect Amino Acid K-mer Matches as a Search Criterion

<table>
<thead>
<tr>
<th>K</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>71%</td>
<td>0.992</td>
<td>0.904</td>
<td>0.697</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>73%</td>
<td>0.996</td>
<td>0.931</td>
<td>0.752</td>
<td>0.560</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>0.998</td>
<td>0.952</td>
<td>0.803</td>
<td>0.625</td>
</tr>
<tr>
<td></td>
<td>77%</td>
<td>0.999</td>
<td>0.969</td>
<td>0.850</td>
<td>0.689</td>
</tr>
<tr>
<td></td>
<td>79%</td>
<td>0.999</td>
<td>0.981</td>
<td>0.890</td>
<td>0.752</td>
</tr>
<tr>
<td></td>
<td>81%</td>
<td>1.000</td>
<td>0.989</td>
<td>0.924</td>
<td>0.810</td>
</tr>
<tr>
<td></td>
<td>83%</td>
<td>1.000</td>
<td>0.994</td>
<td>0.950</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>85%</td>
<td>1.000</td>
<td>0.997</td>
<td>0.970</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>87%</td>
<td>1.000</td>
<td>0.999</td>
<td>0.984</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>89%</td>
<td>1.000</td>
<td>1.000</td>
<td>0.993</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td>91%</td>
<td>1.000</td>
<td>1.000</td>
<td>0.997</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>93%</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.995</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. K</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.2e+07</td>
<td>1.6e+06</td>
<td>62625</td>
<td>2609</td>
<td>112</td>
</tr>
</tbody>
</table>

(A) Columns are for K sizes of 3–7. Rows represent various percentage identities between the homologous sequences. The table entries show the fraction of homologies detected as calculated from equation 3 assuming a homologous region of 33 amino acids. (B) K represents the size of the perfect match. F shows how many perfect matches of this size are expected to occur by chance according to equation 4 in a translated genome of 3 billion bases using a query of 167 amino acids (corresponding to 500 bases).
Choosing threshold $S$

- BLAST may not find all segment pairs above threshold $S$
- Bounds on the error: not hard bounds, but statistical bounds
  - “Highly likely” to find the MSP
Choosing threshold $S$

- BLAST may not find all segment pairs above threshold $S$
- Bounds on the error: not hard bounds, but statistical bounds
  - “Highly likely” to find the MSP

- Is the score high enough to provide evidence of homology?
- Are the scores of alignments of random sequences higher than this score?
- What are is the expected number of alignments between random sequences with score greater than this score?
Choosing threshold \( S \)

- BLAST may not find all segment pairs above threshold \( S \)
- Bounds on the error: not hard bounds, but statistical bounds
  - “Highly likely” to find the MSP

Suppose the MSP has been calculated by BLAST (and suppose this is the true MSP)
- Suppose this observed MSP with a score \( S \).
- What are the chances that the MSP score for two unrelated sequences would be \( \geq S \)?
- If the chances are very low, then we can be confident that the two sequences must not have been unrelated
Statistics: Question

- Given two random sequences of lengths m and n
- What is the probability that they will produce an MSP score of $\geq S$?
Given a binary 0/1 sequence and a query string of \( k \) consecutive ones

- **Probability in a sequence of length \( k \):** \( \frac{1}{2^k} \)
- **Probability in a sequence of length \( k+1 \):**
  - \( 1 - (1 - \frac{1}{2^k})^2 \)
- **How about the probability in a sequence of length \( k+n \)?**
  - \( 1 - (1 - \frac{1}{2^k})^{n+1} \)
- **The longer the sequence, the more likely you are going to get \( k \) ones by chance!**
Statistics: more intuition

The probability will depend on:

• How long is are the sequences (the longer the easier to get a local score above threshold by chance)
• Scoring matrix
• Distribution of amino acids in each sequence
Statistics: Intuition

Frequency of aa occurring in nature

- Ala 0.1
- Val 0.3
- Trp 0.01
...

Random sequence 1

Random sequence 2

Real sequence 1

Real sequence 2

SCORE

SCORE
Approach

⇒ Evaluate the probability that a score between random or unrelated sequences will reach the score found between two real sequences of interest:

If that probability is very low, the alignment score between the real sequences is significant.

If \( \text{SCORE} > \text{SCORE} \) \( \Rightarrow \) the alignment between the real sequences is significant.
How to compute the probability?
Simulation

1. Generate many random sequence pairs

   Frequency of aa occurring in nature
   
   Ala 0.1
   Val 0.3
   Trp 0.01
   ...

   Random sequence 1
   
   Random sequence 2

   SCORE

2. Compute the distribution of the SCOREs