This is a sampling of questions from material presented in class. At the mid-term, you are allowed to bring an A4 size page (both sides) worth of class notes, but otherwise, it will be a closed book exam. Some of these questions are more open ended than what you would find in the mid-term.

1. Draw a finite automaton for the following partial PROSITE pattern.

\[ C-x(1,3)-[LFY]-x(1)-[FYW] \]

2. Does the sequence \( s = AACAAYYYNR \), have a substring that matches this PROSITE pattern from the previous question. Augment your automaton to allow substring queries. Now compute \( N(j) \) for all positions \( j \) in \( s \). Use the sets \( N(j) \) to say why a substring of \( s \) contains the pattern.

3. Consider a Position specific scoring matrix PSSM \( P \). For all amino-acids \( a \), and all positions \( 1 \leq i \leq m \), \( P[a,i] \) gives the score for \( a \) being in the \( i \)-th position. Assume a gap penalty \( g \). Given a query sequence \( s \), design an algorithm that checks if a substring of \( s \) aligns with the profile with a score greater than \( T \).

4. The default word size in BLAST is 11 for DNA, and 3 for protein sequences. Explain with an analysis, why the default word size is different.

5. 40% of the human genome is composed of Repeat sequence. Specifically, there are a few short sequences that occur repeatedly in many copies. What are the implications of this for BLAST searches, in terms of performance, and P-value?

6. Which of the two secondary structures is not stable by itself: alpha helix, or beta strand? How does it form a stable structure?

7. A BLAST search of a query sequence of length 100 against a database of size \( 10^7 \)bp results in an alignment with bit-score = 278. What is the E-value of this alignment? What is the P-value?

8. Consider a query of size 200bp and a database of size \( 10^6 \) base pairs. Explain, with some calculations, how much faster BLAST would be with a word size of 8. How does this change with a word size of 15. If BLAST is always faster with word-size 15, why do we want to use a smaller word-size?

9. Consider the DNA position specific score matrix \( P \) shown below. Suppose we are interested in searching for sequences that score at least 10 in a gapless alignment against PSSM \( P \). Give a list \( L \) of keywords of size 3, such that any sequence which scores at least 10 MUST contain at least one of the keywords in \( L \). Note that this is not a frequency matrix (Profile). We have pre-computed the scores, so that the score for \( A \) in the 1st position is 5, and so on.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>5</th>
<th>-1</th>
<th>-3</th>
<th>-2</th>
<th>5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-1</td>
<td>5</td>
<td>-1</td>
<td>3</td>
<td>2</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-2</td>
<td>2</td>
<td>4</td>
<td>-1</td>
<td>-4</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

10. Consider the sequences \( s = GGAATCATTACCA \), and \( t = AACCGATTCTGG \), a match score of 1, a mis-match score of \(-1\), and linear gap cost of \(-1\) (No gap opening penalty), compute the best local and global alignments of \( s \) and \( t \).

11. When you run BLAST, portions of the sequence are often masked out by X symbols. What does that mean?

12. tblastx and blastp both align two sequences at the protein level. How are they different?

13. Which tool in PROSITE allows you to search with PROSITE patterns. Which tool can you use to search with profiles?