Homework 1: Midterm preparation

Due: May 9, in class.

Problem 1. Reverse Complement.

What are the reverse complements of the following DNA sequences?

1. ATGCAT
2. TGTGTATCAATCTG
3. ACTGATCTAAATGCTAA
4. AATTGGCCAATTGGCC

Problem 2. Genetic Code.

Using the genetic code table (note: the table will be provided to you during the midterm exam as well), translate the following nucleotide sequences into the alphabet of amino acids, reporting the translations in each frame (use "#" for the stop codon).

1. ATGTCCAACCATCAAGTA
2. AAATGAGCCCATGTGATAAA
3. TTGTGATGTTGTACCTATAGTA
Problem 3. GC-percent.

We have a 50-character long DNA fragment, part of which is known to belong to a coding region. The coding region starts with the (only) occurrence of the ATG codon in the fragment and continues through the end of the fragment.

Compute the overall GC-percent of the entire fragment, as well as the GC-percent of the coding and non-coding regions. Do you think the latter two GC-percent numbers are significantly different (for your convenience, we mark positions 1, 11, 21, 31 and 41 in the string below)?

TCACCTTATCAATTCATCGAGTTGTCATGATGGGGACATCCGC

1 1 1 1 1

Problem 4. Codon Usage Bias.

Find the histogram of occurrences of Arginine (Arg, R) and the histogram of occurrences of Phenylalanine (Phe, F), in the following DNA fragment in the first frame (i.e., the first codon starts at the first nucleotide). Ignore all other amino acids (spaces are for your convenience).

AGT TTT CGT CGA CGT TTA TTC CGG CGA TTT GGT CGA TTC CGT TTT CGT CCCTTT CGT TAG

Problem 5. Codon Usage Bias.

Compute the optimal codon frequency for the DNA fragment from Problem 4 given the following optimal codons for Phenylalanine andArginine respectively (ignore occurrences of codons for other amino acids).

1. TTT and CGT.
2. TTC and CGA.
3. TTT and CGG.

Problem 6. Codon Usage Bias

For the DNA sequence in Problem 4 and for each codon for Phenylalanine and Arginine compute its RCSU.

Problem 7. Codon Usage Bias

Compute the quantities $S_{\text{Arg}}$ and $S_{\text{Phe}}$ for Arginine and Phenylalanine for the DNA sequence in Problem 4 (from the prep work for computing the Effective number of codons).

Problem 8. String Matching (KMP)

Compute the prefix function $\pi()$ (represent it as an array of values) for the following strings:

Show how KMP algorithm will look for the occurrences of the string $P = \text{ATGATCAT}$ in the string $S = \text{ATCATCATATGATCATGATCATTTG}$

Note: Here is a sample of how to do it. Let $P = \text{CTCGC}$ and $S = \text{ACTGACTCGCTA}$.

We show the work as follows.

The prefix function $\pi_P[i]$ is:

<table>
<thead>
<tr>
<th>Position</th>
<th>Character</th>
<th>$\pi_P[i]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1</td>
</tr>
</tbody>
</table>

The sequence of comparisons will look as follows (x shows mismatch, | shows match):

\[
\begin{align*}
\text{ACTGACTCGCTA} \\
x \\
\text{CTCGC} \\
\text{ACTGACTCGCTA} \\
| | | \downarrow \text{matched @ position 6} \\
\text{CTCGC} \\
\text{ACTGACTCGCTA} \\
| | | | | | | -> end of S reached. \\
\text{CTCGC}
\end{align*}
\]

Problem 10. Suffix Trees.

Construct suffix trees for the following strings:

S1: ACTCTA
S2: GCCGCT
S3: CATCAGCATC
Problem 11. Repeat search.

Modify the suffix tree for the string `GACTACGACT` to prepare it for the search for maximal repeated sequences. Find all maximal repeated sequences of length greater than 3.

Problem 12. Palindrome search.

Construct the generalized suffix tree that allows for detection of palindromes in the string `ACTCGGATTATCCAAA`. Search for odd-length palindromes of length five and length seven in this string. Report any palindromes found. Show all work.


You are given a DNA fragment of length 50,000 base pairs. The following information about the coding sequences in this file is also made available to you:

<table>
<thead>
<tr>
<th>Gene</th>
<th>Codon</th>
<th>Strand</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td>Codon1</td>
<td>+</td>
<td>3245</td>
<td>5431</td>
</tr>
<tr>
<td>Gene1</td>
<td>Codon2</td>
<td>+</td>
<td>6702</td>
<td>7342</td>
</tr>
<tr>
<td>Gene1</td>
<td>Codon3</td>
<td>+</td>
<td>7753</td>
<td>8320</td>
</tr>
<tr>
<td>Gene2</td>
<td>Codon1</td>
<td>-</td>
<td>12443</td>
<td>11405</td>
</tr>
<tr>
<td>Gene2</td>
<td>Codon2</td>
<td>-</td>
<td>10940</td>
<td>9802</td>
</tr>
<tr>
<td>Gene3</td>
<td>Codon1</td>
<td>+</td>
<td>15323</td>
<td>18940</td>
</tr>
<tr>
<td>Gene3</td>
<td>Codon2</td>
<td>+</td>
<td>19206</td>
<td>22150</td>
</tr>
<tr>
<td>Gene4</td>
<td>Codon1</td>
<td>+</td>
<td>26340</td>
<td>31230</td>
</tr>
<tr>
<td>Gene5</td>
<td>Codon1</td>
<td>-</td>
<td>43227</td>
<td>42932</td>
</tr>
<tr>
<td>Gene5</td>
<td>Codon2</td>
<td>-</td>
<td>41921</td>
<td>39848</td>
</tr>
</tbody>
</table>

Based on this information compute:

1. Average gene size
2. Average CDS size
3. Average exon size
4. Average intron size
5. Nucleotides to genes ratio
6. Gene nucleotide fraction
7. Coding nucleotide fraction
8. Exon density
9. Gene density
10. Relative gene coverage
11. Relative exon coverage

**Problem 14. Gene Density.**

Let us split the DNA fragment from **Problem 13** into two fragments: **Fragment 1** going from position 1 to position 25000 and **Fragment 2** going from position 25001 to position 50000. A gene is considered to be present in a fragment if at least one of its exons is found in the fragment.

Compute the gene density measures specified in **Problem 13** for **Fragment 1** and **Fragment 2**. Compare the gene density in two fragments. Provide some descriptive analysis of the two fragments with respect to their gene density. Explain, which gene density measures (and combinations) contribute to your analysis (and in what way).