Homework 1: Midterm preparation

Due: May 9, in class.

Problem 1. Reverse Complement.

What are the reverse complements of the following DNA sequences?

1. `GATCTGA`

2. `TGTGTATCCCTGTAGC`

3. `ACTGAAGTTAACTGTACG`

4. `AATTGGCCAATTGGCCCATGAG`

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. `ATGTCCAACCAGTTGATGCT`

2. `ATGCAATGAGCCCATGTGAT`

3. `TTGTGTGTACCTATAGAACTC`

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)?

```
TCACTTATCAATGTTCATTCCAGTTGGTCATGATGGGAGGTACATCCGC

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 CCC TTT 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 for Phenylalanine and Arginine 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_{Arg} and S_{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:

```
S1:   AGTTGTC

S2:   ATATCATCATCTATA

S3:   ACTACGTTTACTACTACAC

S4:   GGGTGGGCGGGTGCGCGG

S5:   ATAGATACATAGATACATAG
```

Problem 9. String Matching (KMP).

Show how KMP algorithm will look for the occurrences of the string `P= CACTAC` in the string `S= CACTAGCACTAAACACTACTACGAGTAG`

Note: Here is a sample of how to do it. Let `P = CTCGC` and `S = ACTGACTCGCTA`. We show the work as follows.

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

Position	1	2	3	4	5
Character	C	T	C	G	C
$\pi_P[i]$	0	0	1	0	1

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

```
ACTGACTCGCTA
x
CTCGC

ACTGACTCGCTA
 ||x
 CTCGC

ACTGACTCGCTA
     ||||| ---> matched @ position 6
     CTCGC

ACTGACTCGCTA
        --> end of S reached.
        CTCGC
```

Problem 10. Suffix Trees.

Construct suffix trees for the following strings:

```
S1:  ACTCTA

S2:  GCCGCT

S3:  CATCAGCATC
```

S4: ACACACAC

S5: GACTACGACT

Problem 11. Ukkonen's Algorithm.

Construct implicit suffix trees for the following strings. Show all suffix links in them.

S1: ACTCTA

S2: GCCGCT

S3: CATCAGCATC

S4: ACACACAC

S5: ATCACTTCATCACTA

Problem 12. Repeat search.

Construct 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 13. Generalized suffix trees.

Construct a generalized suffix tree for the following set of strings:

$S_1 = ATGAGTC$
$S_2 = GAGTAG$
$S_3 = CACGTC$
$S_4 = AACATAG$

Find all substrings of length 3 or longer that are present in at least two different strings. Show the nodes in the tree that represent these substrings.

Problem 13. 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.

Problem 14. Gene Density.

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:

Gene	Codon	Strand	Start	End
Gene1	Codon1	+	3245	5431
Gene1	Codon2	+	6702	7342
Gene1	Codon3	+	7753	8320
Gene2	Codon1	-	12443	11405
Gene2	Codon2	-	10940	9802
Gene3	Codon1	+	15323	18940
Gene3	Codon2	+	19206	22150
Gene4	Codon1	+	26340	31230
Gene4	Codon2	+	31843	32671
Gene5	Codon1	-	43227	42932
Gene5	Codon2	-	41921	39848

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 15. 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).