Biocomputing:

HW I

**Important Notes**

1. What to submit:
   1. To Canvas: This document in PDF or Word format with answers typed. Only one group member should upload.
   2. **For groups**: Submit group member and self-evaluation form: **Each group member must submit this form** to give score for themselves and their fellow group members about their contribution to the project. The form is available at https://forms.office.com/r/kKTUx62gm7
2. Answer questions by typing your answers below each question. Please do not delete the questions.
3. Remember to write your name(s) to the header above,
4. Cite your resources for questions such as 1 and 2.
5. Note academic integrity policy reminder below.

**Part 1. Molecular Biology**

1. (10 pts.) In 2-3 sentences explain each term below using your own words.
   1. Polymorphisms
   2. Crossing over
   3. Peptide bond
   4. Alternative splicing
   5. Allele
2. (15 pts.) In recent years, several non-coding RNAs that have key roles in normal physiology and disease state emerged. Among them are microRNAs and circularRNAs. Read about these RNAs in recent literature and give succinct answers to the questions below for both circularRNAs and microRNAs. Answers must be in your own words.
   1. Explain their biogenesis (how are they produced in the cells)?
   2. What are some of their functions?
   3. What diseases are they known to be associated with?
   4. What is competing endogenous RNA network? How does it relate with microRNAs and circularRNAs?
3. (10 pts.) Remember that transcription happens by following the template strand in 3’ to 5’ direction. Below is a non-template DNA sequence of a gene given (in 3’ to 5’ orientation).

3′ CTTTTTTTCAAACCTCTGTTGCGACCGGAAAAGGTCTCCGCTGGAGACGTA

* 1. Write DNA sequence of the template sequence (which would be reverse complement of the non-template sequence above). Make sure to specify the orientation.
  2. Write the corresponding mRNA based on the template sequence. Make sure the specify the orientation.
  3. Write the corresponding protein sequence using genetic code table
  4. List five silent mutations in this sequence such that even if they occur, they would cause no change in the corresponding protein sequence

**Part 2. Finding DnaA boxes in Salmonella enterica**

1. (20 pts) Implement problem 1N (d-neighborhood)
2. (20 pts) Implement 1J, Frequent Words with Mismatches and Reverse Complements (by adapting the approach on page 48 of the textbook, 1.15 on Stepik.org)
3. (25 pts) Follow the steps below to find DnaA boxes in Salmonella enterica (genome sequence is provided on Canvas.)  
   * 1. Run the skew method to compute #G - #C across entire genome
     2. Visualize the skew diagram
     3. Report approximate position of oriC based on skew diagram
     4. Search for most frequent approximate patterns in oriC sequence that you found. Make sure your oriC sequence should have about 100 nucleotides.
        + 1. Try pattern length k = 6 and 9
          2. Try Hamming distance d = 1 and 2
     5. In order to check if the frequent words that you find are unique to the oriC region, search these frequent words in the entire genome and examine where they occur. If they occur frequently elsewhere in the genome that’s possibly a false signal.
     6. Report the final list of most frequent pattern(s)

**Important Reminder on Academic Integrity Policy**

**This assignment must be done in a group up to 2 people. All solutions/writing must be in your own words. Make sure to cite your resources for questions that require any research.**

Academic honesty policy is strictly enforced in this class. Violators of this policy will be reported to the Office of the Provost and the Office for Student Success. There are no exceptions. Please refer to the course syllabus for the academic honesty policy for more details. If you have any questions about this policy, please contact the instructor. **All submissions will be scanned by Turnitin.**