Evening Session Problem Sets

Bio-centric Assignment 1.

The following synthetic protein has two domains from two distinct origins. Predict what is the function of each of the domains, allowed functions are:

a. DNA-binding domain/Transcription Factor domain
b. RNA-binding domain/Ribonucleoprotein domain
c. Fluorescent protein domain
d. Enzyme domain
e. Transmembrane domain

Nucleotide Sequence:

```
>BRACU_MNS-Biotech-CSE
ATGAAATTCAAGTTACATGTGAATTCTGCCAGGCAATACAAGGACCTGTGGAATATGAGTGATGACAAAC
CCTTTCTATGTACTGCGCCTGGATGTGGCCAGCGTTTTACCAACGAGGATCATTTGGCTGTCCATAAACA
TAAACATGAGATGACACTGAAATTTGGTCCAGCACGTAATGACAGTGTCATTGTGGCTGATCAGACCCCA
ACACCAACAAGATTCTTGAAAAACTGTGAAGAAGTGGGTTTGTTTAATGAGTTGGCGAGTCCATTTGAGA
ATGAATTCAAGAAAGCTTCAGAAGATGACATTAAAAAAATGCCTCTAGATTTATCCCCTCTTGCAACACC
TATCATAAGAAGCAAAATTGAGGAGCCTTCTGTTGTAGAAACAACTCACCAGGATAGTCCTTTACCTCAC
CCAGAGTCTACTACCAGTGATGAGAAGGAAGTACCATTGGCACAAACTGCACAGCCCACATCAGCTATTG
TTCGTCCAGCATCATTACAGGTTCCCAATGTGCTGCTTACAAGTTCTGACTCAAGTGTAATTATTCAGCA
GGCAGTACCTTCACCAACCTCAAGTACTGTAATCACCCAGGCACCATCCTCTAACAGGCCAATTGTCCCT
GTACCAGGCCCATTTCCTCTTCTGTTACATCTTCCTAATGGACAAACCATGCCTGTTGCTATTCCTGCAT
CAATTACAAGTTCTAATGTGCATGTTCCAGCTGCAGTCCCACTCGTTCGACCAGTCACCATGGTGCCTAG
TGTTCCAGGAATCCCAGGTCCTTCCTCTCCCCAACCAGTACAGTCAGAAGCAAAAATGAGATTAAAAGCT
GCTTTGACCCAGCAACATCCTCCAGTTACCAATGGTGATACTGTCAAAGGTCATGGTAGCGGATTGGTTA
GGACTCAGTCAGAGGAATCTCGACCGCAGTCATTACAACAGCCAGCCACATCCACTACAGAAACTCCGGC
TTCTCCAGCTCACACAACTCCACAGACCCAAAGTACAAGTGGTCGTCGGAGAAGAGCAGCTAACGAAGAT
CCTGATGAAAAAAGGAGAAAGTTTTTAGAGCGAAATAGAGCAGCAGCTTCAAGATGCCGACAAAAAAGGA
AAGTCTGGGTTCAGTCTTTAGAGAAGAAAGCTGAAGACTTGAGTTCATTAAATGGTCAGCTGCAGAGTGA
AGTCACCCTGCTGAGAAATGAAGTGGCACAGCTGAAACAGCTTCTTCTGGCTCATAAAGATTGCCCTGTA
ACCGCCATGCAGAAGAAATCTGGCTATCATACTGCTGATAAAGATGATAGTTCAGAAGACATTTCAGTGC
CGAGTAGTCCACATACAGAAGCTATACAGCATAGTTCGGTCAGCACATCCAATGGAGTCAGTTCAACCTC
CAAGGCAGAAGCTGTAGCCACTTCAGTCCTCACCCAGATGGCGGACCAGAGTACAGAGCCTGCTCTTTCA
CCGATAGCGCAGATACACATACTGGAGGGGAGGAGCGACGAGCAGAAGGAGACGCTGATAAGGGAGGTGA
GCGAGGCGATAAGCAGGAGCCTGGACGCGCCGCTGACGAGCGTGAGGGTGATAATAACGGAGATGGCGAA
GGGGCACTTCGGGATAGGGGGGGAGCTGGCGAGCAAGGTGAGGAGG
```

Answers:

I. Annotation of the N-terminal domain: [Choose one (a-e)] - 50 points
II. Annotation of the C-terminal domain: [Choose one (a-e)] - 50 points
III. Explain your answer. - 100 points

Concepts tested: Sequence similarity search (bastn, blastx), central dogma, gene annotation basics, protein domains, protein function

CS-centric assignment 1.

CS-centric assignment: Problems will be assembly based. A DNA is a long molecule that can be sequenced in one shot. One way of sequencing DNA is - *split the DNA into shorter segments randomly and sequence them. These random sequences are then assembled into a larger DNA sequence to get to the pre-segmentation sequence. This process is called "DNA sequence assembly."*

For example, if we assume our DNA sequence is "`ATGAGGAATTT`" that is randomly segmented into 3 segments, where individual segment's sequences are:

```
> Segment_1
ATGAG
>Segment_2
AGGAA
>Segment_3
AATTT
```

Notice that the last 2-bases of Segment 1 overlaps with the first 2-bases of the Segment 2. Similarly, the last 2 bases of Segment 2 overlap with the first 2-bases of Segment 3. If we assume that all these segments are from the same contiguous DNA, then we can assemble these segments into the following sequence (where overlaps are highlighted)

```
>Output
ATGAGGAATTT
```

This process is called DNA assembly (or genome assembly) and the software/tool that achieves this goal is called *DNA assembler*.

Write our own DNA assembler using your preferred language. [*i.e.,* write a script that can perform DNA assembly based on the sequence overlaps.]

Correct assembly: 50 points

Logic/ Script soundness: 150 points

Concepts tested: DNA assembly principles, Algorithm for sequence assembly

Collaborative assignment 1.

A DNA sequence is converted into a protein sequence for functions. Three bases of DNA correspond to individual amino acids (also known as codons). A codon table maps nucleotides to corresponding amino acids. The standard codon table is provided below:

```
codons = {
       'A': ['GCT', 'GCC', 'GCA', 'GCG'],
       'C': ['TGT', 'TGC'],
       'D': ['GAT', 'GAC'],
       'E': ['GAA', 'GAG'],
       'F': ['TTT', 'TTC'],
       'G': ['GGT', 'GGC', 'GGA', 'GGG'],
       'H': ['CAT', 'CAC'],
       'I': ['ATT', 'ATC', 'ATA'],
       'K': ['AAA', 'AAG'],
       'L': ['TTA', 'TTG', 'CTT', 'CTC', 'CTA', 'CTG'],
       'M': ['ATG'],
       'N': ['AAT', 'AAC'],
       'P': ['CCT', 'CCC', 'CCA', 'CCG'],
       'Q': ['CAA', 'CAG'],
       'R': ['CGT', 'CGC', 'CGA', 'CGG', 'AGA', 'AGG'],
       'S': ['TCT', 'TCC', 'TCA', 'TCG', 'AGT', 'AGC'],
       'T': ['ACT', 'ACC', 'ACA', 'ACG'],
       'V': ['GTT', 'GTC', 'GTA', 'GTG'],
       'W': ['TGG'],
       'Y': ['TAT', 'TAC'],
       '*': ['TAA', 'TAG', 'TGA']
}
```

Conversion of a DNA sequence to protein sequence starts whenever the first ATG is found and ends whenever a stop codon is encountered (TAA or TAG or TGA; represented as * in the codon table). For example,nucleotide sequence will be converted into the amino acid sequence provided below:

```
AATGTGTGATGAATAA
- M- C- D- E- *
```

Note that the first ATG here is from the second nucleotide position and therefore it is in +2 frame. There can be 3 different reading frames in the shown DNA strand (if ATG is in the first position then +1 frame, if ATG is in the second position its +2 frame, and if ATG is in the second position, then its +3 frame). Since the two strands of DNA are complementary, there are three additional frames in the complement strand (*i.e.,* -1,-2 & -3-frames).

Convert the following DNA sequence into the longest amino acid sequence. Note that protein sequence will start whenever the first ATG is encountered and will end whenever the first stop codon (*) is encountered. Also, each reading frame will give different amino acid sequences. Find the longest contiguous amino acid chain (starts in ATG and ends in any of the stop codons). Return the longest amino acid.

```
>BRACU-SDS-DNA
attttcatatatgtatatatttattatgtcacaaatagctacatactggataagccagaaagatgaggaaacatgtttgca
tctcacactagtgcagagattctgaaaaagaccccacttggaataccaaaccacacattagattgttctgttcccaattgt
gtgccaaagtgcactctgaactgttttggtaaagccgaccgtggagtcatatgaggctgaataacttgggagaatgtaagt
ctgcaaaataaacctaggactggattgatcctcaggccacttggcaggtgaatgtctcgggagtgaatatgagacaagctt
cctgaaaaggcttatatgacttaaagaactttttgtttaagtgtttggtcccaaataaactattaagatatataaagtaat
tcactgctcaaaaattaccgtcagataaatattaagggaagaaacactttaaaggagaatttgtatccacataaatcctta
aattcatggtcacaactttcctaatctcaaattgttctaggtcactgctatcagtctagttctgtcctctatttcttaatt
taataccaatactttaaatttttaaaatatgctgctttaggtagatttaaaatatatcctaaccgccactttctaacctt
cctgaaatatactgttcaataacaaatcctttccttgcttaaacataacttaaataaaacaagaaacttaccatttctctc
tagtgtaagatgacacatgcagcttagcactctgtagtagtctctcttcaattaggtaaaaaaggaagtttctatggtaac
catccttatagtcactcattcctccttccttcctcaccagcgcccaatatatatcacattactcatgaatacatataaact
gctagcatgtctagaatattctgtcttgaaattatactttctactagtgacttttagaaatgactatgccccagaataatt
aaaaagagaaaccaaataatttttaaattaatttatacagagtttttcaccatcaagtttattttcagcacccatggggag
ttaatagatcttccctgaaagagaaaaacctttcctaaggaaaaatcctagaattcatatatttggcaacatttaaaggag
gccatgaaatttttatcagtctagagaattgatctcctttaaggtgactctggtagttccaacgatgtttaaaggcatttc
ctgtaaaaaaaaaaaaaaaagagagcgagagaaaaagagagagagagagagagaatttctgatgattaaaaaaaaaaagt
gaggaatgctgagttaaacaaagttaaccacattctctcagagccttgaatgtgctaatatgtgctaatgtgtcttatggc
tctctaaggagggtgtagtcaaaatcatcttctacactgcttagttcccgggaacccacttttttttttttttattcatg
gtcgaatattatttattgtcagaaaggtacagcattcacaccaatatcagacaaaatagattttaactaaaaaattatttc
gagacaaaaataacaatatatgttaataaaaggctcaattaaaaatgtataacaattataaacacatacacatcaaacaac
agttccccaaaatacgtaaagcaaacattgacaggattgaagggagaaatagaccactctacaatagtagttggggtcttc
aacaccccactttcaataattaatcacagccacttaaggaaccagtgcataccaaaacacacttctggcagtgttgcttca
gggaattccgagtgtagggttaatgttccattttctctacctcttctaaaattcctccttcttcaaatggctattaataac
tttatattttgtcctttgtgtcagttcttaattcatgttgcttataagcatgatttttattattaaagtgagatgggata
aaaataaaatatttttgcaatgagataacgttttattttaattctcaccatttatatacaaacacaagtgaataaaacaca
tcgcaaaatggtaaaatttcatatttagtatttataggtgcatagtttcatgctcacatatttttaagtattatatatatt
aacaaatttcacaatacgtcattattcttagacagtatcattaaaagacacctaaaaatcttataatatatgatagcaaat
cactaacaacttctgaacaacagcaacaaaaaaaatagtgaggatttagaaataagtggtagtcacttaggtgtttttaatt
tgttttaacatcgtagattgaagccacaaaatccacagcacacaaagaccctgctaccatgtattcacttcagtgaaaggg
aagcaccgaaatgctgagtgggggcaggtacagatacttcaatcactgctgatggaagacttcgagatacactgtaaaaac
tttgagaaatgtcatgactgggcacattggaactgagcacttgtacaggatggaacatctgtcagcagatctcaagaaact
gggagcaaagatatttcttaggcttcaggttcgtagtcttgatacccttcctcagaaggcatttcataagcctcattgtca
ggatccacaggcatatcttccagaattccttcctgtggggctccttcttcattcttgcccaactggtcctttttgacaaag
ccagtggctgctgcaatgctccctgctccctccactgtcttctgggctactgctgtcacacccgtcaccactgctcctcca
acatttgtcacttgctctttggtcttctcagccactgttgccacaccatgcaccactccctccttggttttggagcctaca
tagagaacaccctcttttgtctttcctgctgcttctgccacaccctgtttggttttctcagcagcagccacaactccctcc
ttggcctttgaaagtcctttcatgaatacatccatggctaatgaattcctttacaccacactgtcgt
```

Any translation: 50 points

Correct frame translation: 150 points

Concepts tested: Central dogma, codon, Open reading frames.