Name: _____
Score: _____ / _____

Exam 2

Dear All,

Here is your exam for Bioinformatics (CS300/BIO300). This exam is closed notes and closed book; you are not allowed to run any code on any computer to answer any questions. The exam is to be completed in Alden hall only. By the submission of your exam, you are agreeing to adhere to the honor code pledge.

Best of luck to all,
Dr. Bonham-Carter

Part 1

1

Below are four sequence fragments that can be assembled to create a single contig sequence. Show your alignment and the final contig sequence.

Fragments:

CGGACCAGA
ATCGGA
AGACTTTTTTACCAA
GACCAGACTTT

Answer Point Value: 10.0 points
Model Short Answer:

```
                AGACTTTTTTACCAA
          GACCAGACTTT
     CGGACCAGA
ATCGGA
```

ATCGGACCAGACTTTTTTACCAA

Feedback: --------

2

Name and briefly describe two key differences between annotating a eukaryotic genome and a prokaryotic genome.
1. Eukaryotic genomes annotation approach to annotation (5pts)

2. Prokaryotic genome approach to annotation (5pts)

Answer Point Value: 10.0 points
Model Short Answer:
Some ideas:
Eukaryotic genomes are larger with multiple chromosomes – simply more genome to annotate.

Eukaryotic protein coding genes often exist in pieces – exons separated by introns – not only does one have to search for open reading frames (ORFs) but you also need to identify exon/intron boundaries.

Eukaryotic promoter sequences are not as highly conserved as prokaryotic promoter sequences – not all promoters contain the same elements and there can be variation within the DNA sequence of an element across genes, making it harder to identify promoter elements in eukaryotes as compared to prokaryotes.

Prokaryotic annotation algorithms may search for Shine-Dalgarno subsequences to signify a coding region.

Feedback: --------

3

In clear and meaningful language, explain what are the PAM and BLOSUM matrices?

Answer Point Value: 5.0 points
Model Short Answer: Matrices used to score dynamic programming alignments between proteins.
Feedback: PAM 120 and BLOSUM 62 and are general-use substitution matrices and are used for the alignment of proteins.

4

In clear and meaningful language, explain what is a major difference between PAM 120 &BLOSUM 62 and, PAM 250 &BLOSUM 45, in terms of their applications? Explain your reasoning.

Answer Point Value: 5.0 points
Model Short Answer: PAM120 and BLOSUM 62 are for protein comparisons using tools such as BLAST and others from dynamic programming where-as, PAM250 and BLOSUM 45 are generally used to compare proteins of distantly related species using these same tools.
Feedback: --------

5

In clear and meaningful language, explain what key factor gives a protein its ultimate function.

Explain how malfunctions may occur from this factor?

Answer Point Value: 10.0 points
Model Short Answer: Structure from folding gives proteins its function. A improperly folded protein may induce disorders.
Feedback: --------

6

In clear and meaningful language comprising one or two sentences, explain what are open reading frames (ORFs)?

Answer Point Value: 10.0 points
Model Short Answer: An open reading frame (ORF) is the part of a reading frame that has the ability to be translated.
Feedback: --------

7

Imagine that you have just used BLAST to process an unknown protein sequence that you were given. When the results come up after processing, you note that there are several suggested protein sequences. Describe how you would use those sequences to determine the function of your unknown protein. In your discussion, describe what the E-value would tell you as a piece of meta-data for each of the suggested sequences.

Answer Point Value: 10.0 points
Model Short Answer: --------
Feedback: --------

8

Imagine that you are conducting an annotation analysis over the same sequence using an online tool and an offline tool written in Python3. In your analysis, your online tool is the ORFfinder tool from NCBI, and your offline tool is a locally installed program written in Python3. After completing your experiment, you note the results from each tool describe some common findings. However, you also note that there is actually a lot of disagreement between the results. For instance, the online tool has found several results which were overlooked by the Python3 tool.

In clear and meaningful language, describe a likely cause for the differences between the output of each tool when using the same sequence as an input. Which set of results are you more inclined to accept? Why?

Answer Point Value: 10.0 points
Model Short Answer: The online tool also checks the results against a database to reduce false positives. The offline tool used an algorithmic approach that looks for statistically possible annotations which may not be verified.
Feedback: --------

9

What is the Shine-Dalgarno Sequence? Justify.

A. A ribosomal binding site in bacterial and archaeal messenger RNA, generally located around 8 bases found before (upstream of) a start codon AUG. This sequence may serve as a genetic landmark outside of the gene for predicting the beginning of a region of gene code.

Feedback: --------

B. A ribosomal binding site in bacterial and archaeal messenger RNA, generally located around 8 bases found after (downstream of) a start codon AUG and is a landmark inside a gene sequence.

Feedback: --------

C. a ribosomal binding site in bacterial and archaeal messenger protein, which has been observed in the upstream sections of the starts of the codons in gene transcription TATA regulator expressions.

Feedback: --------

D. a ribosomal binding site found principally in cows, whales and some types of buffalo. This binding site has a length of about one amino acid and can serve as a landmark to help determine relations between these organisms.

Feedback: --------

E. A ribosomal binding site in bacterial and archaeal messenger RNA, generally located around 8 bases found after (downstream of) a stop codon. This genetic landmark serves to mark the exact end of a gene region.

Feedback: --------

Answer Point Value: 10.0 points
Answer Key: A
Correct Feedback: --------
Incorrect Feedback: --------

10

Which list represents the correct order for the steps of genome sequencing and assembly?

- A. Genomes are cloned, genomes are broken into fragments, fragments are sequenced, overlapping fragments are assembled to construct the genome sequence. Feedback: --------
- B. Genomes are broken into fragments, fragments are sequenced, genomes are cloned, overlapping fragments are assembled to construct the genome sequence. Feedback: --------
- C. Genomes are cloned, fragments are sequenced, genomes are broken into fragments, overlapping fragments are assembled to construct the genome sequence. Feedback: --------
- D. Overlapping fragments are assembled to construct the genome sequence, genomes are cloned, genomes are broken into fragments, fragments are sequenced. Feedback: --------

Answer Point Value: 10.0 points
Answer Key: A
Correct Feedback: --------
Incorrect Feedback: --------

11

What is the purpose of the tool, BLAST. Justify your answer.

- A. It is an open reading frame discovery tool. Feedback: --------
- B. It is used to find regions of local similarity between sequences. Feedback: --------
- C. It is a database whose principle task is to store the countless reads from an assembly task. Feedback: --------
- D. It is a database which stores genes and supplies meta data concerning its origins, in addition to who was responsible for curating this information. Feedback: --------
- E. It is a system to determine what products are transcribed and translated from original DNA. Feedback: --------

Answer Point Value: 10.0 points
Answer Key: B
Correct Feedback: --------
Incorrect Feedback: --------