

BIIN 200: Bioinformatics I

Dr. Craig Struble

Fall 2008, Midterm Exam

closed book, closed notes, one US Letter cheat sheet, calculators OK

100 points

13 questions, 7 pages

Name: _____

Answer each question in the space provided. Do not use more space than provided. Write neatly. You may use the back of your test pages for scratch paper.

1. [5 pts.] Suppose that you see the following location information in a GenBank entry for a gene or CDS:

```
complement(join(122306..123386,123440..124053))
```

Explain how this relates to the underlying biology of the gene/CDS.

2. [5 pts.] Are *accession numbers* unique identifiers for database entries?
3. [5 pts.] Give the recursive formula used by the dynamic programming algorithm for **local** alignments.
4. [5 pts.] Draw a dot plot for the following two sequences: CACGAC and GATCACG. Assume a window and stringency of 1.

9. We have discussed two primary means for finding sequences in large databases: *keyword searching* and *sequence-based searching*.

(a) [5 pts.] Give an example of keyword searching and of sequence-based searching.

(b) [5 pts.] Briefly describe at least two ways in which these search techniques are similar.

(c) [5 pts.] Give a situation in which sequence-based searching be more appropriate to use than keyword searching.

(d) [5 pts.] Can these two techniques be combined? Give an example or explain why they can't.

10. Suppose that the score for an A to T substitution is -3 in a PAM-like DNA scoring matrix *measured in bits*. Assuming the substitution rates for transitions and transversions are the same, answer the following questions.

(a) [5 pts.] What is the probability of the A to T substitution?

(b) [5 pts.] What is the probability of an A to A substitution?

(c) [5 pts.] What is the score of an A to A substitution?

(d) [5 pts.] Assuming a gap opening penalty of -8 and gap extension penalty of -1, what is the score of the alignment using the scores you have derived in this question.

```
ACT-CTGGAAT
ACTTCT---AT
```

11. [5 pts.] Suppose in a Python script the variable `seq` contains a string for a DNA sequence. In a simple English sentence, state what the following line of code does:

```
loc = seq.find('ATG')
```

12. [5 pts.] Assuming `mat` is a scoring matrix, `abet` is a sequence alphabet, and `s1,s2` are sequences, what does the following Python code do

```
def foo(mat, abet, s1, s2):  
    l1, l2 = len(s1), len(s2)  
    if l1 == l2:  
        s = 0  
        for i in range(l1):  
            x1 = abet.index(s1[i])  
            x2 = abet.index(s2[i])  
            s += mat[x1, x2]  
        return s  
    else:  
        return None
```

13. [10 pts.] Suppose you are given a file named `seqs.gb` containing data in GenBank format. Write a short Python script that will print the accession number of each sequence in the file. You may use either BioPython or plain old Python.