BSCI348S – Fall 2003 – Midterm 1

Multiple Choice: select the single best answer to the question or completion of the phrase. (5 points each)

1. The field of bioinformatics…

a. uses biomimetic algorithms to develop more efficient software.
b. integrates concepts and techniques from information technology and molecular biology.
c. requires complete genome sequences to be useful.
d. has only developed in the last 5 years.

2. What is the principle application of the BLAST family of algorithms?

a. Identifying sequences that are similar to a protein or nucleotide sequence in a biological sequence database.
b. Aligning two nucleotide sequences from end to end.
c. Identifying the best possible alignment of two short protein sequences.
d. Finding the minimum-energy configuration of a polypeptide sequence.

3. Which of the following is NOT true of a BLOSUM80 matrix?

a. It gives the log-odds of substitution of any given pair of amino acids.
b. It is calculated based on the Gibbs free energy (ΔG) of amino acid substitutions.
c. It is based on the BLOCKS database of sequence motifs.
d. It is best used for analysis of very distantly related proteins.

4. At the NCBI web site, the default scoring matrix for protein-protein BLAST analyses (BLASTP) is the BLOSUM62 matrix. Which of the following is a likely reason that this matrix was selected as the default?

a. It represents a compromise between information content for each residue and the amount of information that contributes to the matrix.
b. Only this matrix has been show to be appropriate for all protein comparisons.
c. It incorporates 62 position-specific scoring patterns.
d. Unlike PAM matrices, this matrix is based on explicit phylogenetic information.

5. Which of the following organisms has the largest genome size?

a. *Mycoplasma genitalium*
b. *Escherichia coli*
c. *Amoeba dubia*
d. *Homo sapiens*

Definitions: Provide a 1-2 sentence definition of each term listed below. (2 points each)

6. Homolog

7. Open Reading Frame (ORF)

8. Swiss-Prot

9. e-value (in BLAST analyses)

10. contig

Short Answer: Answer the question in the space provided. Brevity is desirable, and it should be possible to answer the question in a few sentences. (5 points each)

11. Random clone assembly requires that substantially more primary sequence be determined than does map-based assembly. Why is random clone assembly now used more commonly for the determination of complete genome sequences?

12a. What is the difference between a primary database and a secondary database?

b. Is the GenBank nr database primary or secondary? Explain.

13a. Why does it make sense to include gap characters in a pairwise sequence alignment rather than just accepting that some characters will be mismatched?

b. What defines the "best" pairwise sequence alignment?

More grandiose questions: provide a thorough but concise response to the question or problem. Excessively lengthy or disorganized responses will be penalized.

14. Use the Needleman-Wunsch algorithm to align the following sequences. Use a gap-creation penalty of 8. Show your work. An portion of the BLOSUM62 scoring matrix is given below. (15 points)

WGHE
WHE

	E	G	H	W
E	5	-2	0	-3
G	-2	6	-2	-2
H	0	-2	8	-2
W	-3	-2	-2	11

15. For each of the following cases, determine how you would approach the problem, and make recommendations for what analyses to perform. (15 points).

a. Vikram has determined 16,000 500 base-pair reads from a random clone library made from the genome of a newly identified bacterial pathogen whose genome size is thought to be about 1 Mb. What should he do next?

b. Lisa determined the sequence of a human hormone-receptor gene, but is not sure whether this gene is unique, or is a member of gene family. What should she do next?

c. André is interested in the gene triose phosphate isomerase (TPI) in insects. He has access to a newly determined genome sequence from a species of mosquito, but the genome has not yet been annotated. By performing searches with a TPI cDNA sequence from the same organism, he has found the (single-copy) gene in the unannotated genome and has noticed that it contains an intron. He would like to determine exactly where in the protein-coding sequence the intron occurs. What should he do next?

16. Compare and contrast the Smith-Waterman algorithm and the "classic" BLAST algorithm. Be sure to address advantages and disadvantages of each algorithm, and when each should be used. (20 points).