BSCI348S – Fall 2004 – Midterm 1

Multiple Choice: select the single best answer to the question or completion of the phrase. (4 points each)

1. Which of the following is an appropriate application of the Needleman-Wunsch algorithm?

a. Searching a large database for similar sequences.
b. Aligning two sequences from end to end.
c. Finding a local alignment between two very dissimilar sequences.
d. Preparing an entry for the BLOCKS database.

2. At least one *Propionibacterium acnes* gene seems to have been incorporated into the published human genome sequence. Assuming that this gene is not really part of the human genome, which of the following is most likely to have been responsible for that mistake?

a. Craig Venter had really bad acne the day he donated his DNA for sequencing.
b. The *Propionibacterium* sequence is so similar to the human sequence that the difference between these two sequences could not be recognized.
c. A technician inadvertently put a piece of the *Propionibacterium* genome sequence into the contig assembler.
d. The sequence was in an area of low coverage and was probably assembled on the basis of short overlaps of relatively repetitive sequence.

3. What pathogen did Steven Salzberg work on in greatest detail?

a. *Bacillus anthracis*
b. *Haemophilus genitalium*
c. *Mycoplasma influenzae*
d. *Propionibacterium acnes*

4. BLOSUM matrices are derived from the BLOCKS database. What kind of data reside in the BLOCKS database?

a. Short, ungapped multiple sequence alignments of amino-acid data that can be represented by a specific sequence motif.
b. 1472 specific residues selected for uniformity and congruence on the basis of phylogenetic comparisons.
c. DNA sequence data representing at least one species from each of the three domains of life (Bacteria, Archaea, Eukarya)
d. All of the above.

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

a. *Mycoplasma genitalium*
b. *Propionibacterium acnes*
c. *Amoeba dubia*
d. *Homo sapiens*

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

6. Bioinformatics

7. Analogous

8. Markov process

9. Bit (in the context of information theory)

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. What is the distinction between an exact algorithm and a heuristic algorithm?

12. The restriction enzyme EcoR1 has the recognition sequence CAATTC and is expected to cut a genome of 1 megabase ($1x10^6$ BP) approximately 244 times. What assumptions underlie this statement? (5 points)

13. In his 1964 article "Strong Inference" Platt argued that certain methods of scientific thinking were particularly productive. What general approach did he advocate, and how can his insights be applied to the practice of bioinformatics? (5 points)

14. When calculating the odds that a hit would occur at random, BLAST uses a Gumbel extreme value distribution rather than a normal distribution. Why is this appropriate? (5 points)

15. BLAST and its variants are substantially faster than exact pairwise alignment algorithms such as the Smith-Waterman algorithm. Identify three ways in which BLAST reduces the computational time needed to identify regions of similarity. (15 points).

a)

b)

c)

16a. What are the scoring rules for the Smith-Waterman algorithm? (5 points)

16b. There are ten blank cells in the matrix below! Fill in the correct values for these cells using the Smith-Waterman algorithm. A BLOSUM62 scoring matrix is provided on the next page. Assume a monotonic gap penalty of 8. (10 points)

-	-	a	k	q	v	g	r	k	p	y
-	0	0	0	0	0	0	0	0	0	0
k	0	0	5	1	0	0	2	5	0	0
q	0	0	1	10	2	0	1	3	4	0
v	0	0	0	2		6	0	0	1	3
v	0		0	0	6	11	3	0	0	0
g	0	0	0	0	0	12	9		0	0
v	0	0	0	0	4	4	9	7	0	0
p	0	0	0	0	0	2	2	8	14	6
k	0	0	5	1	0	0	4	7	7	12
q	0	0	1	10	2	0	1	5	6	6
v	0	0	0	2	14	6	0	0	3	5
v	0	0	0	0	6	11	3	0	0	2
g	0	0	0		0	12	9	1	0	0
r	0	0	2		0	4				0
p	0	0	0	1	0	0	9			10

16c. Show the trace-back. Is this the only non-trivial diagonal on this matrix? (5 points)

16d. Referring to the dynamic programming matrix above, show the pairwise sequence alignment that will be output by the algorithm. (5 points)

16e. Referring again to the dynamic programming matrix on the previous page, you will note that there is something interesting about this comparison. What do you observe? What biological explanation do you have for this? Is the pairwise alignment you show above a good representation of this phenomenon? Explain your reasoning. (5 points)

17. Consider the BLOSUM62 matrix shown below. Note that the numbers along the diagonal are not identical. What do values along the diagonal represent, and why are they not identical? (5 points)

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```