T-61.5120 Computational genomics Final exam 20.12.2007

Name:

Student ID:

Calculator and two sheets (2xA4) of hand-written notes are allowed. Put your name on the notes and return them together with this final exam. No other material is allowed. Write down your answers directly to these pages and return all pages.

---

1. True/False statements. Mark the following statements whether they are TRUE (T) or FALSE (F). (10p)

a. In proteins mutation rate in core amino acids is usually more than on surface amino acids.

b. Substitution rate in 5' UTR sequences is in general bigger than in 5' Flank sequences.

c. It is estimated that approximately 25% of the genes are alternatively spliced.

d. The UPGMA hierarchical clustering methodology follows the single (minimum) linkage.

e. An E-value for a BLAST analysis is between zero and one.

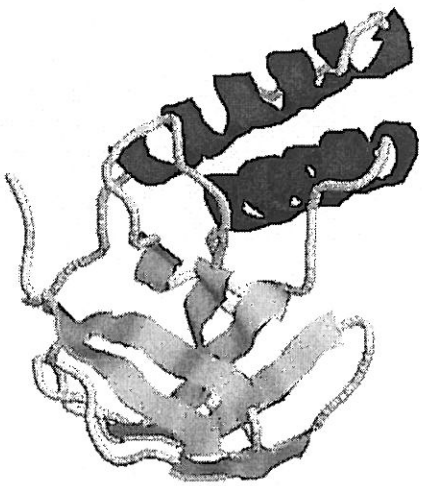
f. Oncogenes suppress tumour formation.

g. Significant structural similarity for proteins does not necessarily indicate an evolutionary relationship

h. Genome between two individuals differs approximately 0.5%.

i. Needleman-Wunsch algorithm is very good solution to generate local alignments.

j. The protein below belongs to the protein class α/β.



2. Answer the following questions. Max three sentences per question. (10p)

a. What is a motif? Mention one computational method to identify motifs and briefly describe how it works.

b. What are micro-RNAs (miRNAs) and how miRNAs regulate gene expression?

c. When is Bonferroni correction needed? Give an example.

d. Often in cancers CpG islands are methylated. How does it affect transcriptional activity of cancer-progression genes having CpG islands occurring in transcription starting sites?

e. What is pseudoknot (in the context of RNA secondary structure prediction)? Explain briefly why predicting pseudoknot structures is challenging.

3. Method optimality. For each of the following methods, circle "Yes" if the method in question is guaranteed to find a globally optimal solution to the problem it is designed to solve. Circle "No" if it is not. (10p)

a) The dynamic programming algorithm for pair-wise sequence alignment.

Yes No

b) The Gibbs Sampler algorithm.

Yes No

c) The Viterbi algorithm.

Yes No

d) The Zuker RNA folding algorithm if all non-pseudoknotted RNA structures are considered.

Yes No

e) The energy minimization approach to protein structure determination.

Yes No

4. Microarray analysis. Assume Napoleon from the lab nearby is doing a microarray data analysis. There are 20,000 genes on the array, 40 cancer samples and 40 reference (healthy) samples. Napoleon's goal is to identify differentially expressed genes (DEGs). As he knows you are an expert in computational genomics, he comes to discuss the project. (10p)

a) Napoleon would like to use t-test to identify DEGs. Is it a good idea? If yes, briefly explain how t-test can be used to identify DEGs. If not, explain why t-test is not a good choice.

b) Based on your recommendations Napoleon analyzes the data. He soon runs back to you and is excited as he found 12 genes with $p < 0.01$. What does a p-value < 0.01 mean? Would you consider these 12 genes truly statistically significant? If yes or no, explain why. If you believe you need more information, what would you ask from Napoleon?

c) Napoleon has read from an article on Gene Ontology analysis. Would it be applicable in Napoleon's project? Explain why.

5. RNA secondary structure prediction. An application of Nussinov's method to predict RNA secondary structure resulted in the matrix below. Perform traceback step and sketch the resulting RNA secondary structure. (10p)

	C	U	G	U	U	A	A	U	G	C	U	A	A
C	0	0	1	1	1	2	3	3	3	4	4	5	6
U		0	0	0	0	1	2	2	2	3	3	4	5
G			0	0	0	1	2	2	2	3	3	4	5
U				0	0	1	2	2	2	3	3	4	5
U					0	1	1	2	2	3	3	4	4
A						0	0	1	1	2	3	3	3
A							0	1	1	2	2	3	3
U								0	0	1	1	2	3
G									0	1	1	2	2
C										0	0	1	1
U											0	1	1
A												0	0
A													0

6. Parsimony method to find the best phylogenetic tree. One method to build phylogenetic trees is character-based parsimony method. Given the sequences below write down all three unrooted trees that are possible. Then, use the sequences at each position to find which unrooted tree is the right one. (10p)

	Position						
Sequence	1	2	3	4	5	6	7
1	**G**	**A**	**T**	**T**	**A**	**C**	**A**
2	**T**	**T**	**T**	**T**	**A**	**C**	**T**
3	**G**	**A**	**T**	**C**	**G**	**C**	**A**
4	**T**	**A**	**T**	**C**	**G**	**C**	**T**

BLOSUM62 substitution matrix and amino acid table:

	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

First Position	Second Position: U code	U Amio Acid	C code	C Amio Acid	A code	A Amio Acid	G code	G Amio Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G