T-61.5120 Computational genomics
Exam 23.10.2012

Kirsti Laurila

A student can have **a function calculator**, graphic calculators are forbidden. All the other material is forbidden.

All papers must be returned!

1) Explain the following concepts briefly (max 5 sentences /concept)

a) Paralogs
b) Unsupervised learning
c) ORF
d) Complete linkage method
e) Forward algorithm
f) Motif discovery
g) Type II error
h) BLAST
i) Frame-shift mutation
j) k-means clustering

2) You have a PFM for TF SOX10

A	0	8	0	0	0	0
C	19	2	1	0	3	0
G	0	2	4	0	19	1
T	3	10	17	22	0	21

a) Compute the information content (for sequence logo) for each site in binding site, draw the sequence logo . Use pseudocount $\frac{1}{4}\sqrt{N}$ for each base.

b) What kind of problems there may occur when predicting TFBSs with PWMs?

3) a) Perform quantile normalization for the following gene expression data, when the columns represent samples and rows are the genes.

45	34	12	45
76	51	34	32
32	64	54	75
31	32	66	32
67	98	14	43
56	34	77	43
25	76	24	76

b) Explain why preprocessing/normalization/quality control is needed in microarray and sequencing data analysis.

4) You have protein sequences GRCRTGSWIF and RGRDTASWI. Find the optimal global alignment when using BLOSUM62 matrix as a scoring matrix (next page) and -6 as the penalty score.

---

5) Plan a HMM that can be used in gene finding for DNA sequences in eukaryotes (i.e. you can deduce whether some region is in gene or intergenic regions from the HMM states.). Draw the model structure and explain why you chose that kind of model.

Additionally, explain what parameters are needed in the model. How one can infer the parameters, if one has only sequences as starting material? You can see a model of a gene structure in next page.

---

6. a) Are the following statements true or false? If your answer is right (true/false) you get +1 point/statement, if your answer is wrong you get -1/statement point. If you don't answer, you get 0 point/statement. However, whole the whole assignment, minimum number of points is 0 (and maximum 5).

i) p-value is the probability that the null hypothesis is true

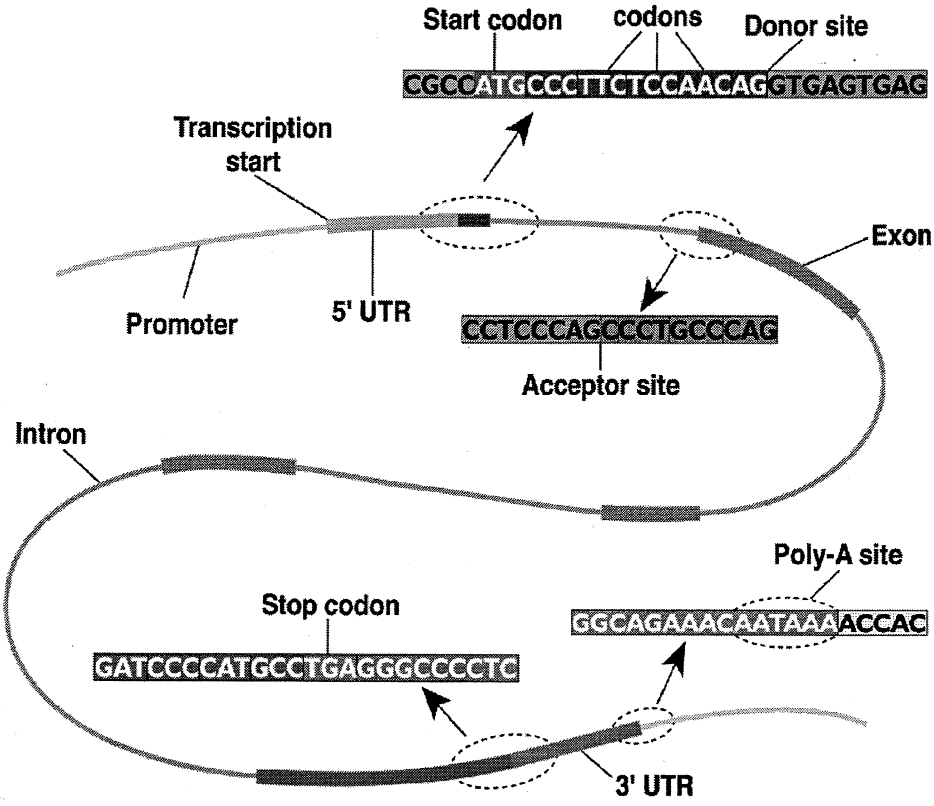
ii) When using distant species in your sequence alignment, you should use PAMxx matrix with high xx number.

iii) The longer promoter sequence you use, the better TFBS finding results you get.

iv) With Baum-Welch algorithm, you can find the optimal HMM parameters.

v) When performing BLAST, you get a bitscore of 20. This means you need a database with ~1 050 000 letters to get similar score by chance.

b) Justify your answers in the a)-part. Explain briefly why you answered true/false. This part doesn't affect the points of the a)-part and you don't get minus points if your answer is wrong.



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W