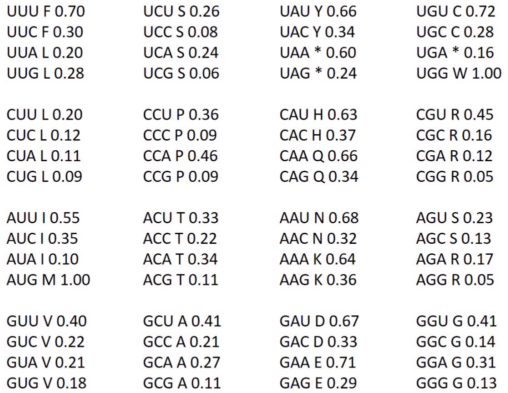
CS-E5865 Computational genomics, Exam, October 20, 2020

Lecturer: Pekka Marttinen

Time: October 20, from 9 am to 1 pm.
You have 3.5 hours for the exam and 0.5 hours for submitting it. All answers must be written on pen and paper and converted into a PDF. The submission must be done as a single PDF in MyCourses and the deadline is at 1 pm. You may use a scientific calculator with memory erased and all materials provided on the course: lecture slides, videos, assignments, and model solutions. Use of other materials and communicating with other students by any means during the exam is not allowed. For more information, see the Exam information announcement in MyCourses.

Q1) Sequence statistics

The codon usage table of *Streptococcus pneumoniae* is given below. Calculate the codon adaptation index for RNA sequence: AUGGCUCGUUGCUUAUAA. What is the biological interpretation of this? (**4p**)



UUU F 0.70	UCU S 0.26	UAU Y 0.66	UGU C 0.72
UUC F 0.30	UCC S 0.08	UAC Y 0.34	UGC C 0.28
UUA L 0.20	UCA S 0.24	UAA * 0.60	UGA * 0.16
UUG L 0.28	UCG S 0.06	UAG * 0.24	UGG W 1.00
CUU L 0.20	CCU P 0.36	CAU H 0.63	CGU R 0.45
CUC L 0.12	CCC P 0.09	CAC H 0.37	CGC R 0.16
CUA L 0.11	CCA P 0.46	CAA Q 0.66	CGA R 0.12
CUG L 0.09	CCG P 0.09	CAG Q 0.34	CGG R 0.05
AUU I 0.55	ACU T 0.33	AAU N 0.68	AGU S 0.23
AUC I 0.35	ACC T 0.22	AAC N 0.32	AGC S 0.13
AUA I 0.10	ACA T 0.34	AAA K 0.64	AGA R 0.17
AUG M 1.00	ACG T 0.11	AAG K 0.36	AGG R 0.05
GUU V 0.40	GCU A 0.41	GAU D 0.67	GGU G 0.41
GUC V 0.22	GCC A 0.21	GAC D 0.33	GGC G 0.14
GUA V 0.21	GCA A 0.27	GAA E 0.71	GGA G 0.31
GUG V 0.18	GCG A 0.11	GAG E 0.29	GGG G 0.13

Q2) Gene finding:

The figure below shows a DNA fragment that contains the beginning of a gene. Identify the coding and template strands, and give their polarities. What is the sequence of bases in the resulting mRNA strand? What are the first three amino acids in the resulting protein (it is enough to give the 'letter' instead of the full name of the amino acid)? (**6p**)

ACTTGTTAACAAGTAATCC

TGAACAATTGTTCATTAGG

Q3) Statistical significance:

a) Suppose you have detected the following ORF: ATGCCAACACCAGCGTGA. Calculate the p-value for this ORF assuming the multinomial sequence model, where you assume that the probabilities of the three stop codons are given by:

$$P(TAA) = 0.02, P(TGA) = 0.02, P(TAG) = 0.01.$$

Interpret the result. (**4p**)

b) Write pseudo-code to estimate the false discovery rate for some threshold δ for ORFs detected from a given DNA sequence $\mathbf{s}$ using permutation sampling. Explain how you interpret the output of the algorithm. (**4p**)

Q4) HMMs:

Consider a HMM with 2 states: A and B, and assume that the observed sequence consists of letters a and b. The transition T and emission E probabilities are given in the figure below, and you can assume equal probabilities for the two initial states. The observed sequence is $\mathbf{s} = (a, b, a)$.

a) Calculate the probability $p(\pi = \pi^*, \mathbf{s})$, where $\pi^* = (A, B, B)$ (**2p**).

b) Calculate the probability, $p(\pi_2 = B|\mathbf{s})$ (**4p**).

Justify all answers by showing the formulas and intermediate results.

$$\mathbf{T} = \begin{array}{c|cc} & \mathbf{A} & \mathbf{B} \\ \hline \mathbf{A} & 0.70 & 0.30 \\ \mathbf{B} & 0.20 & 0.80 \end{array} \qquad \mathbf{E} = \begin{array}{c|cc} & \mathbf{a} & \mathbf{b} \\ \hline \mathbf{A} & 0.90 & 0.10 \\ \mathbf{B} & 0.30 & 0.70 \end{array}$$

Q5) Neighbor-joining algorithm

Suppose nodes A and B have been merged into a new node Z during the NJ algorithm, as shown below in the figure. You know that

1. the distance between A and B is equal to 5.
2. the average distance between A and the nodes $\{v_1, v_2, v_3\}$ is equal to 13.
3. the average distance between B and the nodes $\{v_1, v_2, v_3\}$ is equal to 14.
4. the distance between A and v_1 is equal to 10.
5. the distance between B and v_1 is equal to 12.

Using this information, calculate the following:

a) the distance between Z and v_1. (2p)

b) the length of branch from A to Z. (2p)

c) the neighbor distance M between A and B. (2p)

Remember to justify your answers properly.