Ime in priimek (s tiskanimi črkami) / Name (please print): _____

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

2. izpitni rok / Second Examination Term

16. februar 2015 / February 16, 2015

Naloga / Exercise	1	2	3	4	5	6	Vsota / Sum
Vrednost / Max	5	6	5	5	2	6	29
Točk / Points							

[4] 1. (a) Given is a nucleotide sequence for which we would like to find all open reading frames (ORFs). Assume `ATG` for a start codon, and {`TAA`, `TAG`, and `TGA`}) for the end codons. Report only on proteins with at least four amino acids.

[1] (b) There is one error in the sequence, marked with question mark "?". Which is the most likely nucleotide at that position (given the English word encoded in the genome).

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter
TTG L Leu i	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu i	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met i	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly

(a) V danem zaporedju želimo poiskati vse možne odprte bralne okvire (ORF) in prevesti v zaporedje aminokislin na podlagi podane standardne tabele (začetek z `ATG`, konec z {`TAA`,`TAG`,`TGA`}). Poročaj le o proteinih z vsaj štirimi aminokislinami.

(b) V zaporedje se je prikradla napaka, označena z vprašajem "?". Kateri nukleotid je tam najbolj verjetno pravilen (glede na angleško besedo, ki je zakodirana v genomu).

CTCCATGATTTCTT?TATTAACTAACGTTAAGTCTCACAGTTCTCAACCTGCTCAGACATCC

Page for your solutions. / Stran za vaše rešitve.

The same sequence is printed twice. / Isto zaporedje je izpisano dvakrat.

CTCCATGATTTCTT?TATTAACTAACGTTAAGTCTCACAGTTCTCAACCTGCTCAGACATCC

CTCCATGATTTCTT?TATTAACTAACGTTAAGTCTCACAGTTCTCAACCTGCTCAGACATCC

2. Given are short sequences of DNA fragments from four hypothetical species.

[2] (a) Explain Jukes-Cantor (JC) correction in one or two sentences.

[2] (b) Compute a pairwise distance matrix (mismatch frequency) between the sequences. Correct the matrix using JC correction. Answer should include both the original and the corrected matrix.

[2] (c) Perform hierarchical clustering (UPGMA) and draw a dendrogram of the four sequences, using the JC-corrected matrix. Use the *average* linkage method (see image).

Podana so kratka zaporedja fragmetov DNA štirih hipotetičnih vrst.

(a) V enem ali dveh stavkih razložite Jukes-Cantorjev (JC) popravek.

(b) Izračunajte matriko medsebojnih razdalj (frekvenco različnih nukleotidov) med zaporedji. Popravite vrednosti z uporabo popravka JC. Odgovor naj vključuje tako prvotno kot popravljeno matriko.

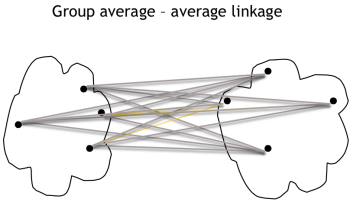
(c) Izvedite hierarhično razvrščanje v skupine (UPGMA) in narišite dendrogram štirih sekvenc na osnovi popravljene matrike. Pri združevanju skupin merite razdaljo med dvema povprečnima točkama (centoma) obeh skupin (slika).

centroma

a ATTCCATTTT
b GATTCATTTC
c TTTCCATTTA
d GTTCCATTTA

$d_{\mathrm{JC}} = -\frac{3}{4}\ln(1 - \frac{4}{3}d)$

$\frac{1}{|A|\cdot|B|}\sum_{x\in A}\sum_{y\in B} d(x,y)$



Page for your solutions. / Stran za vaše rešitve.

[5] 3. We are observing a dice thrower, choosing a fair (F) or a loaded (L) die. Suppose we know the parameters of the Hidden markov model behind the process. We observe the sequence of three dice rolls. What were the probabilities of both states at the second throw, after observing the sequence?

Opazujemo metalca kock, ki izbira med pošteno (F) in uteženo (L) igralno kocko. Predpostavimo, da poznamo parametre skritega markovskega modela, ki opisuje proces. Opazujemo sekvenco treh metov kocke. Kakšni sta verjetnosti obeh stanj ob drugem metu, potem ko smo opazovali zaporedje?

Throw number	1	**2**	3
Observed sequence	1	**6**	6

Transition probabilities

State	$\mathcal{B}$	F	L
$\mathcal{B}$	0	0.5	0.5
F	0	0.75	0.25
L	0	0.75	0.25

Emission probabilities

State/Outcome	1	2	3	4	5	6
F	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
L	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{3}{10}$	$\frac{2}{5}$

	Viterbi	Forward
Initialisation:	$v_0(0) = 1$, $v_k(0) = 0$ for $k \neq 0$	$f_0(0) = 1, f_k(0) = 0$ for $k \neq 0$
$i = 1 \dots L$:	$v_l(i) = e_l(x_i) max_k(v_k(i-1)t_{kl})$	$f_l(i) = e_l(x_i) \sum_k f_k(i-1)t_{kl}$
Termination:	$P_v = \max_k(v_k(L))$	$P_f = \sum_k f_k(L)$

	Backward	Posterior decoding
Initialisation:	$b_k(L) = 1$ for all k	
$i = L-1 \dots 1$:	$b_k(i) = \sum_l t_{kl} e_l(x_{i+1}) b_l(i+1)$	$pd_k(i) = \frac{f_k(i)b_k(i)}{P_f}$
Termination:	$P_b = \sum_l t_{0l} e_l b_l(1)$	$P_{pd}(i) = \text{argmax}_k\ pd_k(i)$

Page for your solutions. / Stran za vaše rešitve.

4. The code below reads a nucleotide sequence and then uses it to perform some operation.

[3] (a) What does the code do?

[2] (b) The code outputs 0.003. How would you interpret this number, what does it tells us?

Spodnji program prebere nukleotidno zaporedje in potem nekaj izračuna.

(a) Kaj počne program, čemu je namenjen?

(b) Program izpiše 0.003. Kaj pomeni ta številka? O čem nam ta rezultat govori?

```
from Bio import SeqIO
import random

def shuffle_string(s):
    lst = list(s)
    random.shuffle(lst)
    return "".join(lst)

original = str(SeqIO.read("data/h_influenzae.fasta", "fasta").seq)
count = original.count("CG")

n = 1000
null = []
for i in range(n):
    shuffled = shuffle_string(original)
    null.append(shuffled.count("CG"))

p = sum(1 for x in null if x > count) / float(n)
print(p)
```

Page for your solutions. / Stran za vaše rešitve.

[2] 5. Phenotype of a mutant A- (gene A was knock-out from the genome) is excessive growth, and phenotype of a mutant B- is reduced growth. Which of the following experiments could give rise to a hypothesis that gene A is epistatic to gene B (this would mean that gene A can block the influence of gene B with respect to the observed phenotype)?

(a) phenotype of mutant A+ is excessive growth

(b) phenotype of double mutant A-B- is reduced growth

(c) phenotype of double mutant A-B- is excessive growth

(d) phenotype of mutant B+ is reduced growth

(e) phenotype a double mutant A-B+ is reduced growth

Fenotip mutante A- (gen A smo zbili iz genoma) je pospešena rast, mutante B- pa zavrta rast. Kateri od spodnjih eksperimentov bi nam dal podlago za hipotezo, da je gen A epistatičen genu B (torej da gen A lahko blokira vpliv gena B)?

(a) fenotip mutante A+ je pospešena rast

(b) fenotip dvojne mutante A-B- je zavrta rast

(c) fenotip dvojne mutante A-B- je pospešena rast

(d) fenotip mutante B+ je zavrta rast

(e) fenotip dvojne mutante A-B+ je zavrta rast

Page for your solutions. / Stran za vaše rešitve.

[5] 6. (a) Given is a list of short sequence reads (k-mers, k=4) from a genome sequencing project. Your goal is to use the de Bruijn graph method to assemble the genome. Draw the de Bruijn graph, reconstruct the genome sequence and report on how you have reconstructed the genome sequence.

[1] (b) Is there more than one possible genome reconstruction? If yes, what are all the solutions?

(a) Podan imaš seznam kratkih odčitkov (nizov dolžine k=4), ki so rezultat sekvenciranja genoma. Uporabi metodo na osnovi de Bruijnovih grafov in sestavi zaporedje genoma. Nariši graf. Poročaj o genomskem zaporedju. Poročaj o tem, kako si sestavil genom.

(b) Je možnih več rešitev? Če ja, jih naštej.

```
ACCA, ACCA, AGTG, ATAC, ATAG, ATCC, CACC, CATA,
```

```
CATA, CATC, CCAT, CCAT, CCAT, TACC, TAGT, TCCA
```

Page for your solutions. / Stran za vaše rešitve.