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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

3. izpitni rok / Third Examination Term

27. avgust 2015 / August 27, 2015

Naloga / Exercise	1	2	3	4	5	6	Vsota / Sum
Vrednost / Max	6	6	5	5	6	6	34
Točk / Points							

[6] 1. V danem zaporedju želimo poiskati vse možne bralne okvire in prevesti v zaporedje aminokislin na podlagi podane standardne tabele (za pričetek vzemite le `ATG`, za konec pa {`TAA`,`TAG`,`TGA`}). Poročajte le o proteinih z vsaj štirimi aminokislinami.

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

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

CTATGTCTGAGTGTCAGAACGATTAAACGTTAAGTAGAACGAATAAACATTGATGACTCACATTCGTGATTAAGGT

Page for your solutions. / Stran za vaše rešitve.

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

CTATGTCTGAGTGTCAGAACGATTAAACGTTAAGTAGAACGAATAAACATTGATGACTCACATTCGTGATTAAGGT

CTATGTCTGAGTGTCAGAACGATTAAACGTTAAGTAGAACGAATAAACATTGATGACTCACATTCGTGATTAAGGT

[6] 2. Given are two sequences

```
CAGA
CATAGG
```

and a scoring function

$$\sigma(a,b) = \begin{cases} 1 & a = b \\ -2 & a = - \text{ or } b = - \\ 0 & \text{otherwise} \end{cases}$$

Propose all possible global alignments with a maximal score. Do this by computing the dynamic programming table, highlight all trace-backs, report on alignment score and show the aligned sequences for all alignments with a maximal score.

Dani sta zaporedji:

```
CAGA
CATAGG
```

in ocenjevalna funkcija

$$\sigma(a,b) = \begin{cases} 1 & a = b \\ -2 & a = - \text{ ali } b = - \\ 0 & \text{v ostalih primerih} \end{cases}$$

Globalno poravnaj zaporedji tako, da bo ocena poravnave maksimalna, in izpiši vse poravnave z najvišjo oceno: pripravi in izračunaj tabelo dinamičnega programiranja, označi “trace-back” za vse poravnave z najvišjo oceno, poročaj o oceni poravnave in prikažite vse poravnave zaporedij z najvišjo oceno.

$$M_{i,j} = \max\left(M_{i-1,j} + \sigma(s_i, -), M_{i,j-1} + \sigma(-, t_j), M_{i-1,j-1} + \sigma(s_i, t_j)\right)$$

Page for your solutions. / Stran za vaše rešitve.

[5] 3. Construct a hidden Markov model from hidden and observable (here DNA) sequences below.

Zgradite skriti Markov model iz skritega in vidnega zaporedja DNA, ki sta zapisani spodaj.

```
IIIIGGGGGGGGGGGGGGGGIIIIIIIIIIII
GTATATGGTAGAACGATATTGATAACAATTCTAT
```

Page for your solutions. / Stran za vaše rešitve.

[5] 4. A hypothetical organism has ten genes: A, B, C, D, E, F, G, H, I, J, and two essential metabolic pathways: B-C-D in A-G-H-I-J. All genes are expressed in the control group, while only genes A, B, C, E, H, J are expressed in infected individuals. Malfunction of which metabolic pathway is more likely associated with the disease? (Hint: consider genes that are expressed differently in disease).

Hipotetični organizem ima deset genov: A, B, C, D, E, F, G, H, I, J, ter dve metabolni poti: B-C-D in A-G-H-I-J, ki sta bistveni za delovanje organizma. V primerjavi s kontrolno skupino, pri kateri so izraženi vsi geni, opazimo, da se ob prisotnosti bolezni izrazijo samo geni A, B, C, E, H, J. Katera metabolna pot je bolj verjetno podvržena posledicam bolezni? (Namig: premisli, kateri geni se ob bolezni izražajo drugače?).

$$P(K=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad \binom{N}{n}=\frac{N!}{(N-n)!n!}$$

Page for your solutions. / Stran za vaše rešitve.

5. 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) Draw a dendrogram of the four sequences, using the JC-corrected matrix. Use the single linkage method (see image).

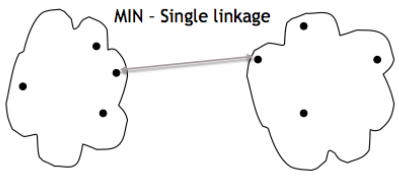
Podana so kratka zaporedja fragmetov DNA štirih hipotetičnih vrst.

(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) Narišite dendrogram štirih sekvenc na osnovi popravljene matrike. Pri združevanju merite razdaljo med dvema najbližjima točkama dveh skupin (slika).

```
ATTCCATTTA
GATTCATTTC
TTTCCATTTT
GTTCCATTTA
```



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

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.

```
AGAG, AGGC, ATCA, CAGA, CCAG, CGAT, GAGG, GATC, GCCA, GCGA, GGCC, GGCG, TCAG
```


```
AGAG, AGGC, ATCA, CAGA, CCAG, CGAT, GAGG, GATC, GCCA, GCGA, GGCC, GGCG, TCAG
```

Page for your solutions. / Stran za vaše rešitve.

Page for your solutions. / Stran za vaše rešitve.