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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

3. izpitni rok / Third Examination Period

11. september 2014 / September 11, 2014

Naloga / Exercise	1	2	3	4	5	6	Vsota / Sum
Vrednost / Max	6	6	6	6	6	6	36
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.

[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.

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

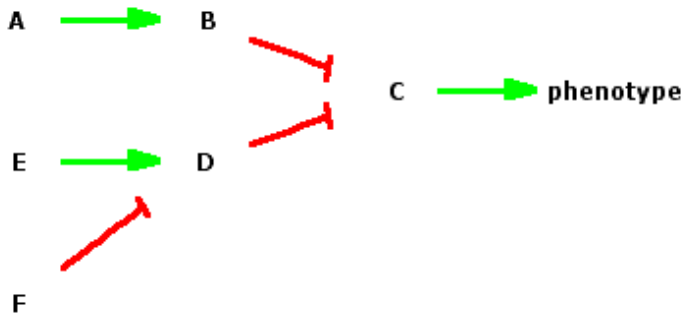
Page for your solutions. / Stran za vaše rešitve.

[6] 3. Gene regulation networks can be inferred through epistasis, where one gene can block the other one in the pathway for a specific phenotype.

Given is a set of 9 experiments, where we have observed a phenotype for the wild type organism (E1), different single mutants (E2 to E5), and different double mutants (E6 to E9). Genes were either knocked-out (e.g, B-) or over-expressed (e.g., A+). The phenotype can have tree values: n (decreased), 0, p (increased), where phenotype 0 is a wild-type phenotype.

ID	Gene 1	Gene 2	phenotype	Confidence	Comments	Ignore	Edit	Delete
E1			0	1.00		I	E	D
E2	A+		n	0.50		I	E	D
E3	B-		p	0.50		I	E	D
E4	D-		p	0.50		I	E	D
E5	E+		n	0.50		I	E	D
E6	B-	C-	n	0.20		I	E	D
E7	D-	C-	n	0.20		I	E	D
E8	E+	D-	p	0.20		I	E	D
E9	F-	D-	p	0.20		I	E	D

We are still missing (at least) three crucial experiments to confirm our hypothesis about what the network should look like:



Which are the three missing experiments (on single or double mutants) that we should perform? What should the outcome (phenotype) of those experiments be?

Pri gradnji genskih regulacijskih mrež iz fenotipskih podatkov o mutantih smo govorili o epistazi, pojavu, kjer en gen lahko blokira druge na regulacijski poti do fenotipa.

Razpredelnica podaja nabor devetih eksperimentov, kjer smo opazovali fenotip pri nemutiranem organizmu (E1), enojnih (E2 do E5) in dvojnih mutantih (E6 do E9). Gene smo pri eksperimentih ali izničili (npr. B-) ali jih čezmerno izrazili (npr. A+). Opazovani fenotip smo zajeli kvalitativno z vrednostmi n (znižan), 0, p (povečan). Fenotip divjega osebka je 0.

Kateri so trije manjkajoči eskperimenti (enojni ali dvojni mutanti in njihovi fenotipi), s katerimi bi lahko dokazali prikazano mrežo genske regulacije (slika).

Page for your solutions. / Stran za vaše rešitve.

4. 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).

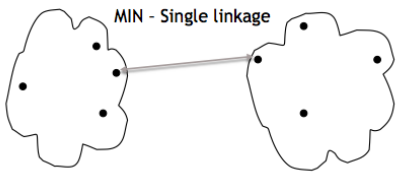
Podana so kratka zaporedja fragmetov DNA štirih hipotetičnih vrst.

(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.

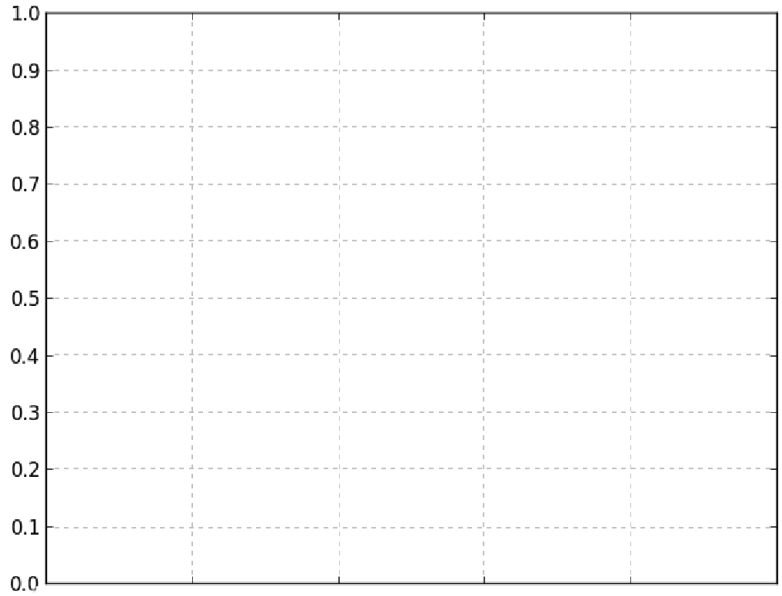
(c) Narišite dendrogram štirih sekvenc na osnovi popravljene matrike. Pri združevanju merite razdaljo med dvema najbližjima točkama dveh skupin (slika).

```
ATTCCATTTA
GATTCATTTC
TTTCCATTTT
GTTCCATTTA
```



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

Page for your solutions. / Stran za vaše rešitve.



[2] 5. (a) What are pseudocounts and why do we use them?

[4] (b) Construct a hidden Markov model from hidden and observable (here DNA) sequences given below. Use a pseudocount of 2.

(a) Kaj so "pseudocount"-i in zakaj jih uporabljamo?

(b) Zgradite skriti Markov model iz skritega in vidnega zaporedja, ki sta zapisani spodaj. Uporabite "pseudocount" 2.

```
IIIIGGGGGGGGGGGGGGGGIIIIIIIIIIII
GTATATGGTAGAACGATATTGATAACAATTCTAT
```

6. You have identified a group of highly conserved genes (given in set C). Explain how to calculate the p-value for the enrichment of a Gene Ontology term T in your set C. Genes known to be associated to the Gene Ontology term T of interest are given in set T_{genes}. Use the hypergeometric distribution (formula below).

[3] (a) How should you set parameters m, n, and N of the hypergeometric distribution? What is k?

[2] (b) Write the formula to compute the p-value (the probability of finding such or better results at random).

[1] (c) Do lower or higher p-values correspond to more significant enrichments?

Odkrili ste skupino dobro ohranjenih genov C. Razložite, kako bi izračunali p-vrednost obogatenosti nekega pripisa T iz genske ontologije (Gene Ontology term) v odkriti skupini C. Množica vseh genov v skupini je T_{genes}. Uporabite hipergeometrijsko porazdelitev (spodnja formulo).

(a) Kako nastaviti parametre hipergeometrijske porazdelitve m, n, in N? Kaj je k?

(b) Napiši formulo za izračun p-vrednosti (verjetnost, da dobite tak ali boljši rezultat po naključju).

(c) Kakšne p-vrednosti pomenijo bolj značilno obogatenost skupine: manjše ali večje?

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

Page for your solutions. / Stran za vaše rešitve.