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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

1. izpitni rok / First Examination Term

29. januar 2015 / January 29, 2015

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

[5] 1. 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 aminoacids.

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

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.

TGATGGGTGAGAACATGTAATATTAATACAACATCTCAGAAGAAGCCATTTG

Page for your solutions. / Stran za vaše rešitve.

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

TGATGGGTGAGAACATGTAATATTAATACAACATCTCAGAAGAAGCCATTTG

TGATGGGTGAGAACATGTAATATTAATACAACATCTCAGAAGAAGCCATTTG

[6] 2. Given are two sequences

```
CAGG
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:

```
CAGG
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 prikaži 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. We have observed a phenotype for the wild type organism (E1), different single mutants (E2 to E6), and different double mutants (E7 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.

We performed epistasis analysis and derived a gene network. But then, part of the experimental data and part of the network got deleted accidentally.

You are presented with the partial information, shown below. Your task is to propose the missing parts of data (experiments E6 and E9). Also, complete the network by drawing the type of influence between nodes in the network (use appropriate arrows: $->$ or $-|$ to indicate the influence type). Briefly explain the reasoning behind each suggested influence.

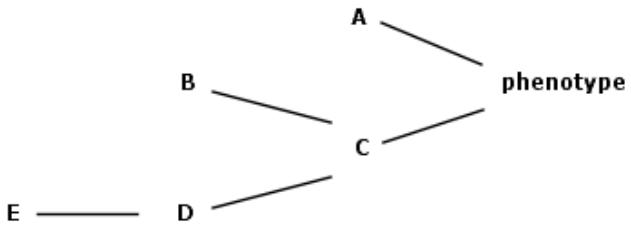
Pri gradnji genskih regulacijskih mrež iz fenotipskih podatkov o mutantih smo govorili o epistazi, pojavu, kjer en gen lahko blokira druge na regulacijski poti do fenotipa.

Razpredelnica podaja nabor devetih eksperimentov, kjer smo opazovali fenotip pri nemutiranem organizmu (E1), enojnih (E2 do E6) in dvojnih mutantih (E7 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.

Na podlagi analize epistaze smo zgradili gensko mrežo. Nato pa smo ponesreči izgubili del eksperimentalnih podatkov in del mreže.

Soočen si z delno informacijo, prikazano na spodnjih dveh slikah. Naloga je predlagati vrednosti manjkajočih podatkov (eksperimenta E6 in E9). Na mreži nariši tudi tipe medsebojnih vplivov genov v mreži (mrežo dopolni s povezavami : $->$ ali $-|$). Na kratko razloži, zakaj predlagaš posamezne tipe povezav.

ID	Gene 1	Gene 2	phenotype
E1			0
E2	A+		n
E3	B-		p
E4	C-		n
E5	D-		p
E6	E+		
E7	B-	C-	n
E8	D-	C-	n
E9			p



Page for your solutions. / Stran za vaše rešitve.

4. Genetic distances have been estimated among homolog sequences in four species, given in the table below (d_{ij}).

[4] (a) Produce the output of the initial step of the neighbor joining algorithm - the matrix D and the (partial) tree. Indicate the branch distances. Which two nodes need to be joined first?

[2] (b) Describe the optimization goal of the neighbor-joining algorithm.

Iz krajših homolognih genskih zaporedji štirih različnih vrst smo ocenili genske razdalje (d_{ij}), ki so podane v tabeli.

(a) Izračunaj prvi korak algoritma združevanja najbližjih sosedov tako, da izračunaš matriko D in narišeš (vmesno) drevo. Na drevesu označi dolžine povezav. Kateri dve vrsti je potrebno najprej združiti?

(b) Algoritem združevanja najbližjih sosedov gradi filogenetska drevesa tako, da pri tem zasleduje specifičen cilj. Kakšen cilj je to oziroma kaj optimizira ta algoritem?

	B	C	D
A	14	14	12
B		16	14
C			6

$$U_i = \sum_{j=1}^{N} d_{ij}$$

$$D_{ij} = d_{ij} - \frac{U_i + U_j}{N-2}$$

$$d_{ik} = \frac{1}{2}\left(d_{ij} + \frac{U_i - U_j}{N-2}\right)$$

$$d_{jk} = d_{ij} - d_{ik}$$

$$d_{km} = \frac{1}{2}\left(d_{im} + d_{jm} - d_{ij}\right)$$

Page for your solutions. / Stran za vaše rešitve.

[5] 5. 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.

[6] 6. Given is a list of short sequence reads (k-mers, k=3) 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.

Podan imaš seznam kratkih odčitkov (nizov dolžine k=3), 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.

AAC, ACA, AGA, ATG, CAT, GAA, GAG, GGT, GTG, TGA

Page for your solutions. / Stran za vaše rešitve.

Page for your solutions. / Stran za vaše rešitve.