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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

2. izpitni rok / Second Examination Period

17. februar 2016 / February 17, 2016

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	4	7	4	6	6	27
Točk / Points						

[4] 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`, `TGA`} for the end codons. Report only on proteins with at least four amino acids.

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.

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	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	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

GCGGGATGCCTCGTGCTGTTATTGGTTAAATTTTTAACCCTGATCAACCTCAGAAGGCATTC

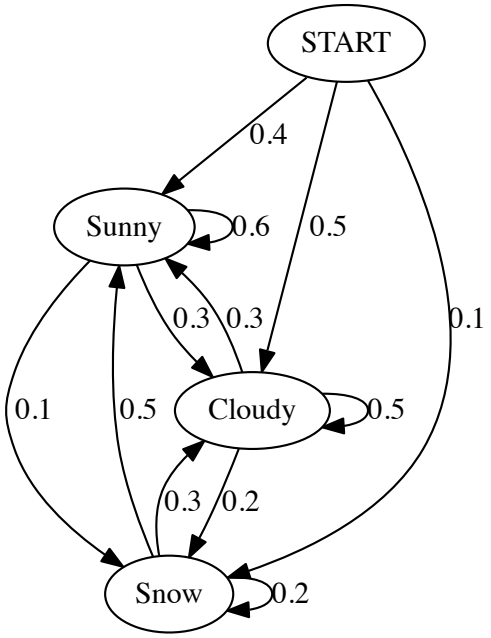
Page for your solutions. / Stran za vaše rešitve.

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

GCGGGATGCCTCGTGCTGTTATTGGTTAAATTTTTAACCCTGATCAACCTCAGAAGGCATTC

GCGGGATGCCTCGTGCTGTTATTGGTTAAATTTTTAACCCTGATCAACCTCAGAAGGCATTC

2. In a winter vacation time the weather in Ljubljana changes like given by a Markov model from the figure.



Depending on a weather, we are considering various types of day-trips (Krvavec is a skiing place, Piran is on the coast, "Home" is just staying at home). The probabilities for each of the trips are given in the table below.

Trip	Sunny	Cloudy	Snow
Krvavec	0.5	0.1	0.1
Piran	0.4	0.5	0.1
Home	0.1	0.4	0.8

In calculating the probabilities below assume that the history (previous day) is not known, and hence we start from the "START" state.

[1] (a) What is the probability of five sunny days in a row?

[1] (b) What is the probability of three snowy days followed by two cloudy days?

[2] (c) What is the probability of two sunny days followed by cloudy day where each day we went to Krvavec (skiing)? Does this probability change if we would go instead to Piran (all three days)? Compute and report on both probabilities.

[3] (d) What was the most probable sequence of weather conditions for the following sequence of our activities: Home, Home, Krvavec? Present your solution in the form of the Viterbi table and clearly mark most probable state transitions and show your derivations of probabilities.

Zgoraj (slika, tabela) je dan skriti markovski model za vreme v Ljubljani in pripadajoče verjetnosti enodnevnih izletov (Piran, Krvavec). Pri izračunih verjetnosti upoštevaj, da podatkov o vremenu pred prvim opazovanim dnem nimamo in moramo zato upoštevati apriorne verjetnosti (prehod iz stanja "START").

(a) Kakšna je verjetnost pet zaporednih sončnih dni?

(b) Kakšna je verjetnost, da tri dni zapored sneži, potem pa je dva dni oblačno?

(c) Trikrat zapored smo šli na Krvavec. Kakšna je verjetnost, da je prva dva dni bilo sončno, potem pa oblačno? Bi bil ta niz vremenskih slik bolj verjen, če bi izvedeli, da smo namesto na Krvavec šli v Piran? Izračunajte in poročajte o verjetnostih.

(d) Kakšno je najbolj verjetno zaporedje vremenske slike za naslednje zaporedje dnevnih aktivnosti: Home, Home, Krvavec? Svojo rešitev predstavi z Viterbijevo tabelo. Jasno označi najbolj verjetne prehode. Iz rešitve naj bodo razvidni tudi izračuni posameznih verjetnosti.

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

$$v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl})$$

[4] 3. A hypothetical organism has 100 genes in total. You have performed a stress test, which identified 5 genes to be differentially regulated under the stress.

There is one gene function of special interest to you, let's call it *functionX*. There are 33 genes that are known to perform *functionX*.

From the 5 genes you have identified, at least how many need to be associated with *functionX* in order to claim the association (enrichment) statistically significant ($p < 0.15$)? Clearly write and explain the calculation that supports your answer.

Hipotetični organizem ima 100 genov. Opravil si stresni test, s katerim si določil 5 genov, ki se odzovejo na stres.

Predvsem te zanima določena funkcija genov, imenujmo jo *funkcijaX*. Vseh genov, za katere vemo, da so povezani s *funkcijoX*, je 33.

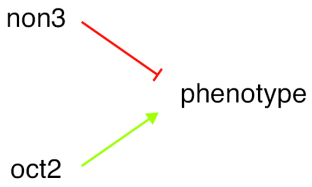
Med 5 geni, ki si jih odkril, vsaj kolikim mora biti pripisana *funkcijaX*, da lahko trdimo, da je funkcijska obogatitev statistično značilna ($p < 0.15$)? Podrobno opiši izračun, ki podpira odgovor.

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

Page for your solutions. / Stran za vaše rešitve.

4. Gene regulation networks can be inferred through epistasis, where one gene can block the other one in the pathway for a specific phenotype. Experiments can consist of wildtype (unchanged conditions), single or double mutants (gene knockdown or activation). The following phenotypes can be measured: **- -** highly decreased expression, **-** decreased expression, **0** wildtype expression, + increased expression, ++ highly increased expression.

[2] (a) Given is a small, hypothetical network of two genes *non3* and *oct2* and their influence on the phenotype. List a sufficient set of experiments to reconstruct the network shown on figure below. Fill in the missing experiments in Table 1. Briefly explain the reasoning behind each suggested experiment.



exp. ID	gene 1	gene 2	phenotype	comment
E1			0	wildtype

Tabela 1: Experiments for part (a).

[4] (b) The experiments in Table 2 describe programmed cell death in *C. elegans*. Reconstruct the genetic network using the experimental evidence. Briefly argue your choice of edges in the network.

exp. ID	gene 1	gene 2	phenotype	comment
E1			0	wildtype
E2	ced9-		+	
E3	ced4-		-	
E4	egl1-		-	
E5	ced3-		-	
E6	ced3+		+	
E7	ced3-	ced9-	-	
E8	ced4-	ced3-	- -	
E9	ced9-	ced4-	-	
E10	egl1-	ced9-	+	

Tabela 2: Experiments for part (b).

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. Eksperimenti lahko obsegajo nespremenjen organizem (*divji tip*) enojne ali dvojne mutante (odstranitev ali aktivacija gena). Merimo lahko naslednja stanja fenotipa: **- -** zelo zmanjšano izražanje, **-** zmanjšano izražanje, **0** nivo izražanja divjega tipa, + povečanjo izražanje, ++ zelo povečano izražanje.

(a) Podano je manjše hipotetično gensko omrežje, sestavljeno iz dveh genov *non3* in *oct2* ter pripadajočega fenotipa. Naštej zadostno skupino eksperimentov, ki omogoča sestavljanje omrežja na sliki. Dopolni Tabelo 1. Za vsakega od predlaganih eksperimentov podaj kratko utemeljitev.

(b) V Tabeli 2 je podan seznam eksperimentov, ki opisuje pojav proženja celične smrti (apoptoze) pri organizmu *C. elegans*. Sestavi gensko omrežje na podlagi eksperimentalnih podatkov. Utemelji izbiro posameznih povezav.

[5] 5. (a) Given is a list of short sequence reads (k-mers, k=4) from a genome sequencing project. Your goal is to build a graph on which you can use the Eulerian Path approach to reconstruct the genome sequence. Draw the 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? If no, explain why not.

(a) Podan imaš seznam kratkih odčitkov (nizov dolžine k=4), ki so rezultat sekvenciranja genoma. Uporabi metodo na osnovi iskanja Eulerjeve poti in tako sestavi zaporedje genoma. Nariši ustrezen graf. Poročaj o sestavljenem genomskem zaporedju. Podrobno poročaj o postopku, kako si sestavil genom.

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

```
AGAT    ATAG    ATGA    ATGC    CGAT    GATA    GATG    GATG    TAGA    TGAT
```

$k\text{-}mer_1 \rightarrow k\text{-}mer_2$, if suffix($k\text{-}mer_1$) = prefix($k\text{-}mer_2$) (e.g., $TAA \rightarrow AAG$)

prefix(k-mer) $\rightarrow$ suffix(k-mer), for each k-mer (e.g., for $TAA : TA \rightarrow AA$)

(prefix(k-mer) returns first $k-1$ letters of k-mer; suffix(k-mer) returns last $k-1$ letters of k-mer.)

Page for your solutions. / Stran za vaše rešitve.

Page for your solutions. / Stran za vaše rešitve.