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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

1. izpitni rok / First Examination Period

26. januar 2017 / January 26, 2017

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	6	7	6	6	6	31
Točk / Points						

1. Given is a part of a dynamic programming table for global alignment of two sequences:

	-	T	C	T	A
-	0	-2	-4	-6	-8
A	-2	-1	-3	-5	-4
T	-4	0			
G	-6	-2			
A	-8	-4			

[2] (a) What is a scoring function that we have used?

[2] (b) Complete the dynamic programming table. That is, propose the values for its nine missing elements.

[2] (c) What is the global alignment of the two sequences with a maximal score according to our scoring function. Show the aligned sequences and highlight the trace-back in the dynamic programming table.

Dana je tabela dinamičnega programiranja za globalno poravnavo dveh zaporedij.

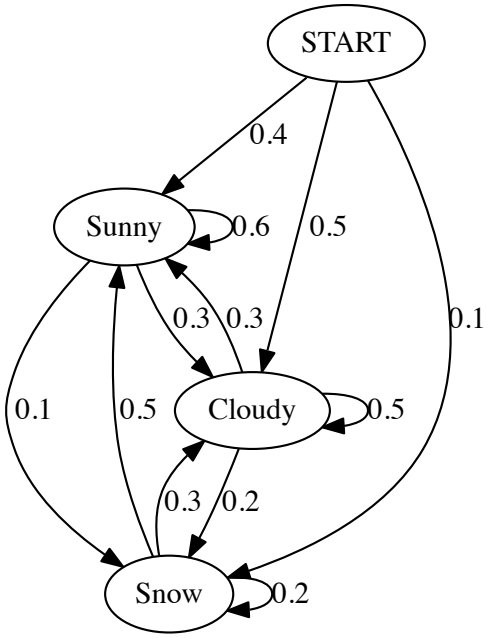
(a) Kakšno cenovno funkcijo smo uporabili?

(b) Dopolni manjkajoči del tabele.

(c) Kakšna je globalna poravnava dveh zaporedij iz tabele, ki ima najvišjo oceno? Zapiši poravnani zaporedji in prikaži sled poravnave (angl. trace-back) v tabeli dinamičnega programiranja.

Page for your solutions. / Stran za vaše rešitve.

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 verjeten, 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.

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

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

3. A hypothetical organism has 97 genes in total. You have performed a stress test, which identified 6 genes to be differentially regulated under the stress.

   Out of total 97 genes, there are 43 genes connected to metabolism.

[2] (a) When is an observation termed *statistically significant*? What is the p-value?

[4] (b) Suppose that 5 out 6 genes that react to stress are also connected to metabolism. Is this observation statistically significant? Justify your answer (by means of calculation).

Hipotetični organizem ima 97 genov. Opravil si stresni test, s katerim si določil 6 genov, ki se odzovejo na stres.

Predvsem te zanimajo geni, ki so povezani z metabolizmom. Vseh genov, za katere vemo, da so povezani z metabolizmom, je 43.

(a) Kdaj pravimo, da je meritev *statistično značilna*? Kaj je p-vrednost?

(b) Ali je meritev statistično značilna, če je med 6 geni, ki se odzovejo na stres, 5 takih, ki so povezani z metabolizmom? Utemelji odgovor (s pomočjo izračuna).

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

Page for your solutions. / Stran za vaše rešitve.

[6] 4. 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 11 experiments, where we have observed a phenotype for the wild type organism (E1), different single mutants (E2 to E7), and different double mutants (E10, E11, E13, E14). Genes were knocked-out (e.g. A-). The phenotype can have tree values: − (decreased), + (normal), ++ (increased), where phenotype + is a wild-type phenotype.

Use the experimental data to reconstruct the network.

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 enajstih eksperimentov, kjer smo opazovali fenotip pri nemutiranem organizmu (E1), enojnih (E2 do E7) in dvojnih mutantih (E10, E11, E13, E14). Gene smo pri eksperimentih izničili (npr. A-). Opazovani fenotip smo zajeli kvalitativno z vrednostmi − (znižan), + (normalen), ++ (povečan). Fenotip divjega osebka je +.

Na osnovi eksperimentalnih podatkov zgradi gensko mrežo.

ID	Gene 1	Gene 2	phenotypeX
E1			+
E2	A-		++
E3	B-		-
E4	C-		++
E5	D-		-
E6	E-		++
E7	F-		-
E10	D-	F-	-
E11	A-	B-	++
E13	A-	D-	-
E14	E-	F-	++

Page for your solutions. / Stran za vaše rešitve.

[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.

`ATCA` `ATCG` `ATCT` `CATC` `CATC` `CTAT` `GCAT` `TATC` `TCAT` `TCTA`

$k\text{-}mer_1 \rightarrow k\text{-}mer_2$, 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.