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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

1. izpitni rok / First Examination Period

2. februar 2016 / February 2, 2016

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	6	8	5	8	6	33
Točk / Points						

1. Consider the program provided below.

[2] (a) What is printed on the output when this program is run?

[2] (b) What is the meaning of this output for the particular choice of sequences x and y. What does a function lcs, in general, do?

[2] (c) Propose two sequences x and y each of length 7 and $x \neq y$ for which the output of the program would be "4".

Spodaj je podan kratek program.

(a) Kaj ta program izpiše?

(b) Kaj nam ta izpis pove glede na dani izbor zaporedij x in y. Kaj v splošnem počne funkcija lcs?

(c) Predlagaj zaporedji x and y, kjer sta obe zaporedji dolgi 7 znakov in velja $x \neq y$ tako, da bi za ti zaporedji program izpisal "4".

```
from collections import defaultdict

def lcs(s, t):
    table = defaultdict(int)
    for i, si in enumerate(s):
        for j, tj in enumerate(t):
            table[i, j] = max(
                table[i-1, j],
                table[i, j-1],
                table[i-1, j-1] + (si == tj)
            )
    return table

x = "AACCTTGG"
y = "ACACTGTGA"
table = lcs(x, y)
print(table[len(x)-1, len(y)-1])
```

Solution: The output of the program is 6. Function lcs computes the length of the longest common subsequence (note: this is different from longest common substring). One such example could be

```
x="1234000"; y="1234111"
```

Page for your solutions. / Stran za vaše rešitve.

2. Consider a hidden Markov model from the figure below.

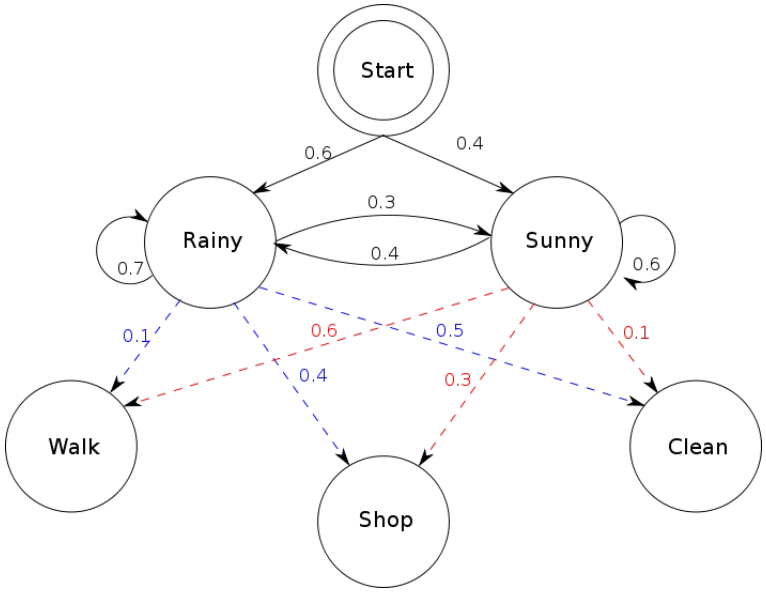
[1] (a) What is the probability of five rainy days in a row?

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

[2] (c) What is the probability of two sunny days followed by two rainy days, within which one will walk, then shop, clean and then shop again (in this sequence)?

[3] (d) What is the most probable sequence of hidden states (Rainy and Sunny) for observed sequence of Walk, Walk, Clean?

[1] (e) Compare the probability of the sequence of observed and hidden states from the question above with the probability of the observed sequence under three rainy days or under three sunny days (compute each of the probabilities).



Na sliki zgoraj je podan skriti markovski model.

(a) Kakšna je verjetnost, da pet dni zapored dežuje?

(b) Kakšna je verjetnost, da dežuje tri dni zapored, potem pa je dva dni sončno?

(c) Kakšna je verjetnost dveh sončnih dnevov ki jima sledita dva deževna dneva, kjer gremo prvi dan na sprehod, drugi dan v trgovini, tretji dan čistimo in četrti dan spet obiščemo trgovino?

(d) Kakšno je najbolj verjetno zaporedje skritih stanj (Rainy in Sunny) za zaporedje Walk, Walk, Clean?

(e) Primerjaj verjetnost zaporedja skritih in opaženih stanj iz prejšnjega vprašanja z verjetnostjo, da smo zaporedje opaženih stanj zabeležili pri treh deževnih ali zabeležili pri treh sončnih dnevih (izračunaj pripradajoče verjetnosti).

Solution:

(a) $0.6 \times 0.7^4 = 0.14$

(b) $0.6 \times 0.7 \times 0.7 \times 0.3 \times 0.6 = 0.053$

(c) $0.4 \times (0.6 \times 0.6) \times (0.3 \times 0.4) \times (0.5 \times 0.7) \times (0.4 \times 1.0) = 0.00242$

(d) sunny, sunny, rainy with $p = 0.0173$,
Viterbi table {Rainy: 0.06, Sunny: 0.24}, {Rainy: 0.0096, Sunny: 0.0864}, {Rainy: 0.0173, Sunny: 0.00518}

(e) three rainy days: $0.6 \times (0.1 \times 0.7)^2 \times 0.5 = 0.00147$, three sunny days: $0.4 \times (0.6 \times 0.6)^2 \times 0.1 = 0.00518$

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

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

Page for your solutions. / Stran za vaše rešitve.

3. Using a sophisticated experimental technique, you have identified a group of genes of interest (listed in set C).

   Now, you would like to understand the functional role of genes in C. There are two Gene Ontology (GO) terms that you suspect are specially relevant to the experiment and should be enriched in C: GO terms T_1 and T_2.

   All genes known to belong to GO terms T_1 and T_2 are listed in sets G_{T_1} and G_{T_2}, respectively.

[2] (a) Explain, how to use the formula for the hypergeometric distribution (given below) to calculate the probability of finding, by chance, the observed overlap between the genes in a term (T_1, T_2) and the identified group of genes C. How should you set parameters m, n, and N? What is k? Explain for one of the terms, e.g., explain for T_1.

[2] (b) Write the formula to compute the p-value (the probability of finding such or better overlaps at random). For each of the terms (T_1 and T_2), write a formula to calculate the corresponding p-value.

[1] (c) Assume that p_1 is the p-value calculated for term T_1, and p_2 is the p-value for term T_2. When can we claim that group C is more significantly associated with T_1 than with T_2?

Z napredno eksperimentalno tehniko si določil skupino genov C.

Zdaj bi rad razumel funkcijo genov v skupini. Za dve funkcijski kategoriji genske ontolgije (GO) še posebej domnevaš, da sta povezani s funkcijo eksperimentalno določene skupine C: to sta funkcijski skupini T_1 in T_2.

Vsi geni, za katere je znano, da so pripisani funkcijski skupini T_1, so navedeni v množici G_{T_1}. Vsi povezani s T_2 so navedeni v množici G_{T_2}.

(a) Opiši, kako uporabiti formulo hipergeometrijske porazdelitve (podana spodaj) za izračun verjetnosti, da bi po naključju opazili takšno prekrivanje med geni v funkcijski skupini (T_1, T_2) in eksperimentalno določeni skupini C. Razloži pomen parametrov m, n in N? Kaj je k? Razloži na primeru izračuna ene funkcijske skupine, recimo za T_1.

(b) Napiši formulo za izračun p-vrednosti (verjetnost, da bi po naključju dobili takšno ali večje prekrivanje). Zapiši formulo za izračun p-vrednosti, za vsako funkcijsko skupino (T_1 in T_2) posebej.

(c) Recimo, da je p_1 predstavlja p-vrednost izračunano za skupino T_1, p_2 pa vrednost za skupino T_2. Kdaj lahko trdimo, da je skupina C bolj značilno obogatena s pripisom T_1 kot pa s pripisom T_2?

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

Solution:

m - velikost G_{T_1} (ali G_{T_2})

n - velikost C

N - število vseh genov, kakorkoli se ga dobi

k - število genov v preseku C in G_{T_1} (ali G_{T_2})

(b) - seštevek od velikost preseka (vključno) do največjega možnega ali 1 - od 0 do vključno (velikosti preseka - 1)

(c) manjša p-vrednost pomeni večjo obogatenost: $p_1 < p_2$.

Page for your solutions. / Stran za vaše rešitve.

4. Below is the output of a partially completed Neighbor joining algorithm, which was run on 6 species: A, B, C, D, E, F. The pairwise distances between nodes are given in the table (only the lower triangular part is shown for brevity).

[6] (a) Run the algorithm to completion and complete the tree below. Draw the missing nodes (if any), edges, and write the missing edge lengths. Make sure to perform and to document all the missing steps of the algorithm.

[2] (b) Describe the Neighbor joining algorithm. What does the algorithm optimize?

Podan je izpis delno izvedenega algoritma združevanja sosedov (ang. *neighbor joining*), na šestih bioloških vrstah: A, B, C, D, E, F. Trenutne medsebojne razdalje so podane v tabeli (prikazana je samo polovica tabele).

(a) Dokončaj izvajanje algoritma tako, da dopolneš drevo z manjkajočimi vozlišči (skupnimi predniki, če obstajajo), povezavami in oznakami dolžin povezav. Izvedi in jasno zapiši vse korake do konca izvajanja algoritma.

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

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

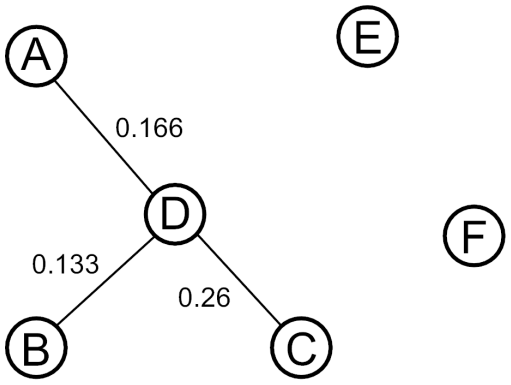
$$D_{ij} = d_{ij} - \frac{U_i + U_j}{N-2}$$

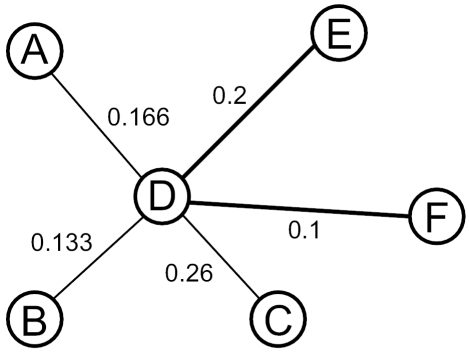
$$d_{ik} = \frac{1}{2}\Big(d_{ij} + \frac{U_i - U_j}{N-2}\Big)$$

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

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

	A	B	C	D	E	F
A						
B	0.299					
C	0.426	0.393				
D	0.166	0.133	0.26			
E	0.366	0.333	0.46	0.2		
F	0.266	0.233	0.36	0.1	0.3	





Solution: (a)

Corrections U_i:

```
U(D) = 0.3
U(E) = 0.5
U(F) = 0.4
```

Corrected distances D_{ij}:

```
DE = 0.2 - (0.3+0.5) = -0.6
DF = 0.2 - (0.3+0.5) = -0.6
EF = 0.2 - (0.3+0.5) = -0.6
```

Choosing any pair is equivalent. Choosing DE and defining new node G yields.

```
d(DG) = 0.5 * (0.2 + 0.3 - 0.5) = 0
        implying D = G.
d(EG) = 0.5 * (0.2 + 0.5 - 0.3) = 0.2
d(FG) = 0.5 * (0.3 + 0.1 - 0.2) = 0.1
```

(b)
The Neighbour joining algorithm is a greedy algorithm, optimizing according to balanced minimum evolution. The algorithm minimizes the length of the tree, such that the distance of each path along the edges of the tree (from one original taxon to another) will correspond to the true distance.

[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 Hamiltonian 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 Hamiltonove 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.

`ATAC` `ATTG` `CGAT` `GATA` `GATT` `TGAT` `TTGA`

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

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

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

Solution:

```
genome (10): CGATTGATAC
GENOME RECONSTRUCTION USING SINGLE k-mers
4-mer composition (7): ['ATAC', 'ATTG', 'CGAT', 'GATA', 'GATT', 'TGAT', 'TTGA']

Hamilton path approach:
ATTG -> TTGA
CGAT -> GATA, GATT
GATA -> ATAC
GATT -> ATTG
TGAT -> GATA, GATT
TTGA -> TGAT

only one solution (reconstruction), Hamilton path:
CGAT -> GATT -> ATTG -> TTGA -> TGAT -> GATA -> ATAC
```

```
In case they make a de Bruijn graph (which was not requested), it should be:
de Bruijn graph: {'ATA': ['TAC'], 'CGA': ['GAT'], 'GAT': ['ATA', 'ATT'], 'ATT': ['TTG']
ATA ['TAC']
ATT ['TTG']
CGA ['GAT']
GAT ['ATA', 'ATT']
TGA ['GAT']
TTG ['TGA']
cycle: ['CGA', 'GAT', 'ATT', 'TTG', 'TGA', 'GAT', 'ATA', 'TAC']
genome        : CGATTGATAC

no other solution (no other Eulerian path)

 +0.5 if some graph
 -3 if graph not correct
 -3 if genome sequence wrong
 -0.5 if last nucleotide of reconstructed sequence missing
```

Page for your solutions. / Stran za vaše rešitve.

Page for your solutions. / Stran za vaše rešitve.