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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

3. izpitni rok / Third Examination Period

25. avgust 2017 / August 25, 2017

3

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

[5] 2. Given is a program in Python. What does it output? Give an exact solution, that is, write the output of the program, rather than giving a conceptual answer, in a sense "It outputs a dynamic programming table.".

Dan je spodnji program. Kaj izpiše? Podaj torej konkreten izpis programa, in ne konceptualen odgovor tipa "Izpiše tabelo dinamičnega programiranja.".

```
def sigma(a, b):
    return -2 if ("-" in [a, b]) else (1 if a == b else -1)

def dpt(s, t):
    table = {(0, 0): 0}
    table.update({(i+1, 0): sigma("-", c) * (i+1) for i, c in enumerate(s)})
    table.update({(0, i+1): sigma("-", c) * (i+1) for i, c in enumerate(t)})

    for i in range(1, len(s)+1):
        for j in range(1, len(t)+1):
            table[i, j] = \
                max(table[i-1, j] + sigma(s[i-1], "-"),
                    table[i, j-1] + sigma("-", t[j-1]),
                    table[i-1, j-1] + sigma(s[i-1], t[j-1]))
    return table

def pp(s, t, table):
    print("  " + " ".join("%2s" % c for c in "-" + t))
    for i, c in zip(range(len(s)+1), "-" + s):
        print(c + " " + " ".join("%2d" % table[i, j] for j in range(len(t)+1)))

s, t = "ATGA", "ATCTA"
pp(s, t, dpt(s, t))
```

Page for your solutions. / Stran za vaše rešitve.

[5] 3. Construct a hidden Markov model from hidden state path (in first row: I=intron, G=gene) and observable DNA sequence (second row: four letters of the DNA alphabet) sequences.

Zgradi skriti Markov model iz zaporedja skritih stanj (prva vrstica: I=intron, G=gen) in vidnega zaporedja (druga vrstica: štiri črke abecede DNA), ki sta zapisani spodaj.

```
IIIIGGGGGGGGGGGGGGGGIIIIIIIIIIII
GTATATGGTAGAACGATATTGATAACAATTCTAT
```

Page for your solutions. / Stran za vaše rešitve.

4. We have conducted a study of effects of genetic modification of tomato. From $N = 10$ tomatoes in the study, $m = 6$ were genetically modified, while $N - m = 4$ where borrowed from our neighbor's garden. Our neighbor is a hair dresser and she is (otherwise) not involved in genetic experiments.

   Our aim was to study the effect of genetic modification to the expression of the genes *obi1* and *obi2*. We have noticed that *obi1* is overexpressed with five tomatoes ($n = 5$), od which three ($k = 3$) are genetically modified. Gene *obi2* was overexpressed with four tomatoes ($n = 4$), of which three ($k = 3$) were genetically modified.

[2] (a) Estimate (or, if really needed, compute) on which of the gene expressions (e.g., *obi1* vs *obi2*) the genetic modification has a larger effect.

[2] (b) Is this effect arbitrary? How likely it would be obtained if the expression of the two genes would not be related to our experiment?

[2] (c) How would you compute probability of obtaining the same or better result (in terms of number of tomatoes where the two genes are overexpressed). Propose the formula, you do not need to compute the probability.

V študiji gensko spremenjenega paradižnika smo vključili $N = 10$ primerkov. Šest paradižnikov ($m = 6$) je bilo takih, ki so gensko spremenjeni, štiri ($N - m = 4$) pa smo si sposodili iz vrta sosede. Soseda je po poklicu frizerka in se v prostem času ne ukvarja z genetiko.

Zanimal nas je vpliv genskih sprememb na izražanje genov *obi1* in *obi2*. Izražanje teh genov smo zmerili in opazili, da je *obi1* nadpovprečno izražen pri petih paradižniki ($n = 5$), od katerih so trije ($k = 3$) gensko spremenjeni. Gen *obi2* se je visoko izrazil pri štirih paradižnikih ($n = 4$), od katerih so trije ($k = 3$) gensko spremenjeni.

(a) Oceni (lahko pa tudi izračunaš), na katerega od genov *obi1* in *obi2* je genska sprememba bolj vplivala?

(b) Je ta vpliv naključen? Kakšna je verjetnost, da bi tak rezultat dobil z naključnim žrebom.

(c) Opiši, kako bi izračunal, kakšna je verjetnosti, da bi dobil tak ali boljši rezultat, kot smo ga dobili v eksperimentu. Rezultata ti ni potrebno izračunati. Najbolje, če tudi lahko podaš kar enačbo, pri kateri si morda lahko pomagaš z izrazom na dnu strani.

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

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.

```
ATAT    ATCG    ATCT    CATA    CTTA    GCAT    TATC    TATC    TCTT    TTAT
```

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