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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

2. izpitni rok / Second Examination Period

10. februar 2017 / February 10, 2017

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	4	5	5	6	6	26
Točk / Points						

[4] 1. 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 (2 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] 2. We sequenced the DNA of a simple organism and obtained the sequence on the following page. The machine could not determine the nucleotides at four positions. These were marked as "?".

Luckily, we also obtained an independent and more reliable amino acid sequence of the three proteins encoded by the genome:

```
MVERY
MSHQRT
MGENQME
```

Identify the ORFs for the three known proteins. Then use the information on protein amino acid sequence to determine the most likely values of the missing nucleotides. If a nucleotide can not be determined unambiguously, explain why is it so.

Use the standard codon tabel (below, ORFs start with `ATG` and end with {`TAA`,`TAG`,`TGA`}).

Sekvencirali smo zaporedje DNA enostavnega organizma in dobili zaporedje na naslednji strani. Med sekvenciranjem je prišlo do napake branja štirih nukleotidov. Stroj je ta mesta označil z "?".

K sreči so bila eksperimentalno neodvisno in tehnično bolj zanesljivo določena tudi zaporedja aminokislin vseh treh proteinov, ki jih organizem lahko tvori:

```
MVERY
MSHQRT
MGENQME
```

Najprej določi položaje genov (ORFov) za tri znane proteine. Nato uporabi podatke o aminokislinskem zaporedju proteinov in karseda natančno določi vrednost čimveč manjkajočih nukleotidov. V primeru, da nukleotida ni možno določiti nedvoumno, navedi razloge, zakaj ni možno.

Pomagaj si s standardno kodno tabelo (spodaj, za pričetek ORF-a vzemi le `ATG`, za konec pa {`TAA`,`TAG`,`TGA`}).

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

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

GTATATGGTAGAACGATATTGATAA?AATTCTATT?CATCTGGTTTTCCCCCATTACATGTCACA?CAAC?AACCTAAATGGG

GTATATGGTAGAACGATATTGATAA?AATTCTATT?CATCTGGTTTTCCCCCATTACATGTCACA?CAAC?AACCTAAATGGG

[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. Marie and Pierre are leading researchers in the field of a rare genetic disease. They are currently evaluating computer models of the disease. They do not share data between each other.

   Marie's group has 100 human genome sequences at their disposal, out of which 12 are known to be positive cases of the disease. Their computer model predicted 20 potential cases from all 100, and 11 predictions were correct.

   Pierre's group have 80 genomes, out of which 14 are known positive cases of the disease. Pierre's model has identified 10 potential cases.

[3] (a) Is Marie's result statistically significant? Declare a p-value threshold on your own.

[3] (b) How many of Pierre's predictions have to be correct, for his result to be more significant than Marie's?

   Support you conclusions with necessary calculations. Tip: some intermediary results can perhaps be reused.

Marija in Peter sta vodilna raziskovalca na področju redke genske bolezni. Trenutno vrednotita računalniške modele te bolezni. Med seboj ne delita podatkov.

Marijin laboratorij ima na voljo 100 zaporedij človeških genomov, od katerih je 12 povezanih z omenjeno boleznijo. Njihov model je izmed vseh 100 kandidatov označil 20 potencialno obolelih, pri čemer je bilo 11 napovedi pravilnih.

Petrov laboratorij ima na voljo 80 genomov, od tega 14 povezanih z boleznijo. Model, ki so ga razvili, je od 80 kandidatov označil 10 potencialno obolelih.

(a) Ali je delež pravilnih napovedi Marijinega algoritma statistično značilen? Stopnjo značilnosti določi sam/a.

(b) Najmanj koliko napovedi Petrovega modela mora biti pravilnih, da bo rezultat bolj statistično značilen od Marijinega?

Zaključke podpri s potrebnimi izračuni. Namig: nekateri vmesni rezultati so morda ponovno uporabni.

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

Page for your solutions. / Stran za vaše rešitve.

5. Below is the output of a partially completed Neighbor joining algorithm, which was run on 5 species: A, B, C, D, E. The pairwise distances between nodes are given in the table (only the lower triangular part is shown for brevity).

[5] (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.

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

Podan je izpis delno izvedenega algoritma združevanja sosedov (ang. *neighbor joining*), na petih bioloških vrstah: A, B, C, D, E. 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?

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

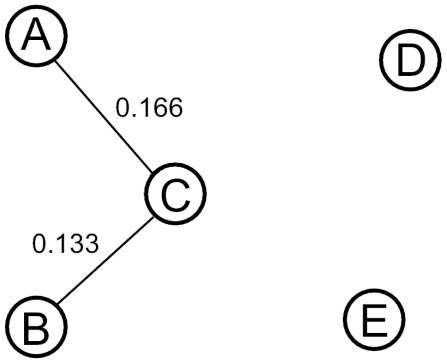
$$D_{ij} = d_{ij} - \frac{U_i + U_j}{N-2}$$

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

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

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

	A	B	C	D	E	
A						
B	0.299					
C	0.166	0.133				
D	0.366	0.333	0.2			
E	0.266	0.233	0.3	0.4		



Page for your solutions. / Stran za vaše rešitve.