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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

3. izpitni rok / Third Examination Period

22. avgust 2016 / August 22, 2016

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

GGTGTATGCCTCATGATGGTCGTGGTCGTTAACCCATTAAGTCATAGTCATATCATGAGGCATTG

Page for your solutions. / Stran za vaše rešitve.

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

GGTGTATGCCTCATGATGGTCGTGGTCGTTAACCCATTAAGTCATAGTCATATCATGAGGCATTG

GGTGTATGCCTCATGATGGTCGTGGTCGTTAACCCATTAAGTCATAGTCATATCATGAGGCATTG

Solution:

```
start codons: ['ATG']
stop codons: ['TAA', 'TAG', 'TGA']
Forward table
TTT F    TCT S    TAT Y    TGT C
TTC F    TCC S    TAC Y    TGC C
TTA L    TCA S    TAA *    TGA *
TTG L    TCG S    TAG *    TGG W

CTT L    CCT P    CAT H    CGT R
CTC L    CCC P    CAC H    CGC R
CTA L    CCA P    CAA Q    CGA R
CTG L    CCG P    CAG Q    CGG R

ATT I    ACT T    AAT N    AGT S
ATC I    ACC T    AAC N    AGC S
ATA I    ACA T    AAA K    AGA R
ATG M    ACG T    AAG K    AGG R

GTT V    GCT A    GAT D    GGT G
GTC V    GCC A    GAC D    GGC G
GTA V    GCA A    GAA E    GGA G
GTG V    GCG A    GAG E    GGG G


Back-translate amino acid sequence
MPHDGRGR*
ATGCCTCATGATGGTCGTGGTCGTTAA

MPHDMTMT*
ATGCCTCATGATATGACTATGACTTAA
RC: TTAAGTCATAGTCATATCATGAGGCAT

+:
GGTGTATGCCTCATGATGGTCGTGGTCGTTAACCCATTAAGTCATAGTCATATCATGAGGCATTG 65
G  V  C  L  M  M  V  V  V  V  N  P  L  S  H  S  H  I  M  R  H
 V  Y  A  S  *  W  S  W  S  L  T  H  *  V  I  V  I  S  *  G  I
  C  M  P  H  D  G  R  G  R  *  P  I  K  S  *  S  Y  H  E  A  L

-:
CAATGCCTCATGATATGACTATGACTTAATGGGTTAACGACCACGACCATCATGAGGCATACACC
Q  C  L  M  I  *  L  *  L  N  G  L  T  T  T  T  I  M  R  H  T
 N  A  S  *  Y  D  Y  D  L  M  G  *  R  P  R  P  S  *  G  I  H
  M  P  H  D  M  T  M  T  *  W  V  N  D  H  D  H  H  E  A  Y  T


kriteriji:
```

[6] 2. Given are two sequences

```
AGTCTG
GGTATCG
```

and a scoring function

$$\sigma(a,b) = \begin{cases} 1 & a = b \\ -2 & a = - \text{ ali } b = - \\ -1 & \text{otherwise} \end{cases}$$

Find the global alignment with a maximal score. Do this by computing the dynamic programming table, highlight the trace-back, report on alignment score and show the aligned sequences.

Dani sta zaporedji:

```
AGTCTG
GGTATCG
```

in ocenjevalna funkcija

$$\sigma(a,b) = \begin{cases} 1 & a = b \\ -2 & a = - \text{ ali } b = - \\ -1 & \text{v ostalih primerih} \end{cases}$$

Globalno poravnaj zaporedji tako, da bo ocena poravnave maksimalna (pripravi in izračunaj tabelo dinamičnega programiranja, poročaj o oceni poravnave in prikažite poravnani zaporedji).

Solution:

```
Tabela:
s: GGTATCG
t: AGTCTG

  0  -2  -4  -6  -8 -10 -12
 -2  -1  -1  -3  -5  -7  -9
 -4  -3   0  -2  -4  -6  -6
 -6  -5  -2   1  -1  -3  -5
 -8  -5  -4  -1   0  -2  -4
-10  -7  -6  -3  -2   1  -1
-12  -9  -8  -5  -2  -1   0
-14 -11  -8  -7  -4  -3   0
ocena globalne poravnave: 0
GGTATCG
 || | |
AGTCT-G


\begin{verbatim}
Kriteriji:
+3 tabela, ki vsebuje pravilne vrednosti (za napake odbijaj)
+2 poravnava je pravilna (ena od 6 moznih)
+0.5 ocena poravnave je navedena oz. oznacena
+0.5 ocena poravnave je pravilna
```

$$M_{i,j} = \max\Big(M_{i-1,j} + \sigma(s_i, -), M_{i,j-1} + \sigma(-, t_j), M_{i-1,j-1} + \sigma(s_i, t_j)\Big)$$

Page for your solutions. / Stran za vaše rešitve.

3. Given are short sequences of DNA fragments from four hypothetical species.

[1] (a) Explain Jukes-Cantor (JC) correction in one or two sentences.

[2] (b) Compute a pairwise distance matrix (mismatch frequency) between the sequences. Correct the matrix using JC correction. Answer should include both the original and the corrected matrix.

[2] (c) Draw a dendrogram of the four sequences, using the JC-corrected matrix. Use the *average* linkage method (see image).

[2] (d) What is the computational complexity (storage and time) of hierarchical clustering in terms of the number of sequences (n), using the *average* linkage method?

Podana so kratka zaporedja fragmetov DNA štirih hipotetičnih vrst.

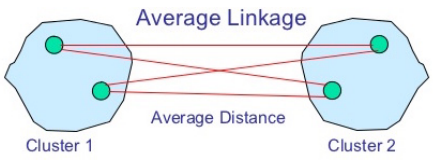
(a) V enem ali dveh stavkih razložite Jukes-Cantorjev (JC) popravek.

(b) Izračunajte matriko medsebojnih razdalj (frekvenco različnih nukleotidov) med zaporedji. Popravite vrednosti z uporabo popravka JC. Odgovor naj vključuje tako prvotno kot popravljeno matriko.

(c) Narišite dendrogram štirih sekvenc na osnovi popravljene matrike. Pri združevanju skupin merite razdaljo med dvema povprečnima točkama (centroma) obeh skupin (slika).

(d) Kakšna je prostorska in časovna zahtevnost algoritma za hierarhično gručenje v odvisnosti od števila sekcenc (n), če je za združevanje skupin uporabljena omenjena razdalja med centroma skupin?

```
ATTCCATTTT
GATTCATTTC
TTTCCATTTA
GTTCCATTTA
```



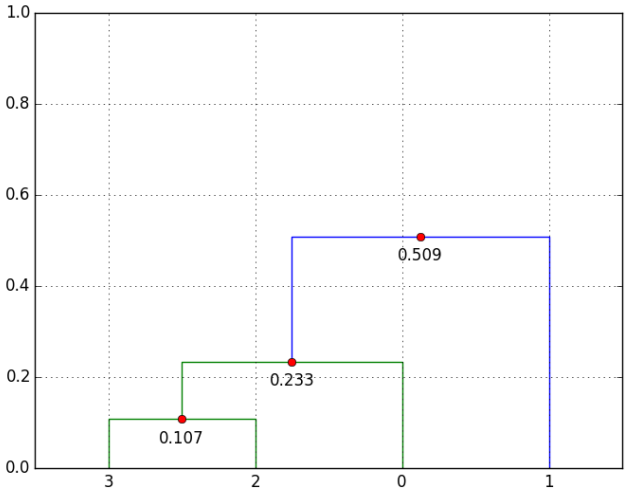
$d_{\mathrm{JC}} = -\frac{3}{4}\ln(1 - \frac{4}{3}d)$

Page for your solutions. / Stran za vaše rešitve.

Solution:

```
  0.   0.4  0.2  0.2
  0.4  0.   0.4  0.3
  0.2  0.4  0.   0.1
  0.2  0.3  0.1  0.

Distances (corrected)
 0.          0.57160504  0.2326162   0.2326162
 0.57160504  0.          0.57160504  0.38311922
 0.2326162   0.57160504  0.          0.10732563
 0.2326162   0.38311922  0.10732563  0.
```



Both time and storege complexities are $O(n^2)$

```
-0.5 if proportions ok, but y-axis scale is wrong
-1 if wrong neighbours are joined
-1 if using absolute frequency instead of relative
```

Page for your solutions. / Stran za vaše rešitve.

[4] 4. A hypothetical organism has 100 genes in total. You have performed a stress test, which identified 6 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 20 genes that are known to perform *functionX*.

From the 6 genes you have identified, at least how many need to be associated with *functionX* in order to claim the association (enrichment) statistically significant ($p < 0.05$)? Clearly write and explain the calculation that supports your answer.

Hipotetični organizem ima 100 genov. Opravil si stresni test, s katerim si določil 6 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 20.

Med 6 geni, ki si jih odkril, vsaj kolikim mora biti pripisana *funkcijaX*, da lahko trdimo, da je funkcijska obogatitev statistično značilna ($p < 0.05$)? Podrobno opiši izračun, ki podpira odgovor.

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

Solution:

```
N=100
m=20
n=6
k: P(k)     sum(P)   sum(P) < 0.05
6: 0.0000   0.0000   True
5: 0.0010   0.0011   True
4: 0.0128   0.0139   True
3: 0.0786   0.0925   False
2: 0.2521   0.3446   False
1: 0.4033   0.7479   False
0: 0.2521   1.0000   False
```

[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.

```
AGAC  AGCA   ATAG   ATGC  CAGA   CATA  CATG   GCAG   GCAT  TAGC   TGCA
```

for each *k-mer*: prefix(*k-mer*) $\rightarrow$ suffix(*k-mer*), e.g., for $k = 3$, $TAA : TA \rightarrow AA$

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

Solution:

```
genome (14): CATAGCATGCAGAC
GENOME RECONSTRUCTION USING SINGLE k-mers
4-mer composition (11): ['AGAC', 'AGCA', 'ATAG', 'ATGC', 'CAGA', 'CATA', 'CATG', 'GCAG', 'GCAT',
de Bruijn graph: {'GCA': ['CAG', 'CAT'], 'ATG': ['TGC'], 'ATA': ['TAG'], 'AGC': ['GCA'], 'AGA':

AGA ['GAC']
AGC ['GCA']
ATA ['TAG']
ATG ['TGC']
CAG ['AGA']
CAT ['ATA', 'ATG']
GCA ['CAG', 'CAT']
TAG ['AGC']
TGC ['GCA']

WARNING, more than one solution
cycle: ['CAT', 'ATG', 'TGC', 'GCA', 'CAT', 'ATA', 'TAG', 'AGC', 'GCA', 'CAG', 'AGA', 'GAC']
genome        : CATAGCATGCAGAC
reconstructed : CATGCATAGCAGAC
False


 +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.