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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

3. izpitni rok / Third Examination Period

5. september 2012 / September 5, 2012

Naloga / Exercise	1	2	3	4	5	6	Vsota / Sum
Vrednost / Max	24	24	10	0	8	16	82
Točk / Points							

[4] 1. (a) Napišite funkcijo `izberi`, ki glede na verjetnosti nukleotida v multinomskem modelu za neko zaporedje, ki ga tvorimo, izbere nov nukleotid. Multinomski model je podan v pythonskem slovarju. Pomagajte si s funkcijo `random`, ki vam z enakomerno porazdelitvijo vrne število med 0 in 1.

```
model = {'A': 0.3, 'C': 0.2, 'T': 0.4, 'G': 0.1}
def izberi(model):
   """vrne nukleotid z verjetnostjo, podano z model"""
   # vaša koda tu
```

[4] (b) Zapišite funkcijo `multinomski(model, dolzina)`, ki generira zaporedje podane dolžine za podan multinomski model.

[4] (c) Zapišite funkcijo `markov(mmodel, dolzina)`, ki generira zaporedje podane dolžine za podan markovski model. Markovski model je podan kot pythonska terka z verjetnostmi $P(x_0)$ začetnega nukleotida x_0 (kot prvi element terke) in z verjetnostmi naslednjega nukleotida $P(x_i)$ glede na prejšnjega x_{i-1} (kot drugi element terke; ključ predstavlja x_{i-1}, vrednost pa $P(x_i)$ za nukleotid x_i):

```
mmodel = ({'A': 0.23, 'C': 0.27, 'T': 0.37, 'G': 0.13},
          {'A': {'A': 0.2, 'C': 0.12, 'T': 0.3, 'G': 0.38},
           'C': {'A': 0.25, 'T': 0.25, 'C': 0.25, 'G': 0.25},
           'T': {'A': 0.1, 'C': 0.3, 'T': 0.4, 'G': 0.2},
           'G': {'A': 0.1, 'C': 0.3, 'T': 0.3, 'G': 0.3}}
 )
```

---

[4] (d) Create a Python function `choice` that returns a nucleotide according to the multinomial model. The model is given as a Python dictionary. You can use function `random` which returns a number between 0 and 1 from a uniform distribution.

```
model = {'A': 0.3, 'C': 0.2, 'T': 0.4, 'G': 0.1}
def choice(model):
   """return a nucleotide according to the model"""
   # your code here
```

[4] (e) Design another function `multinomial(model, len)`, which returns a nucleotide sequence of a given length `len` for a given multinomial model.

[4] (f) Design a function `markov(mmodel, len)`, which generates a sequence of a given length `len` for a given Markov model. The model is a Python tuple with probabilities $P(x_0)$ of a star tining nucleotide x_0 (first element of a nucleotide sequence) and transition probabilities $P(x_i)$ with respect to the previous nucleotide x_{i-1} in the sequence (the key of the dictionary is x_{i-1}, the value $P(x_i)$ for nucleotide x_i):

```
mmodel = ({'A': 0.23, 'C': 0.27, 'T': 0.37, 'G': 0.13},
          {'A': {'A': 0.2, 'C': 0.12, 'T': 0.3, 'G': 0.38},
           'C': {'A': 0.25, 'T': 0.25, 'C': 0.25, 'G': 0.25},
           'T': {'A': 0.1, 'C': 0.3, 'T': 0.4, 'G': 0.2},
           'G': {'A': 0.1, 'C': 0.3, 'T': 0.3, 'G': 0.3}}
 )
```

2. Za dve regiji DNA imamo podani zaporedji, za prvo pa še strukturo genov na niti + z vrednostmi _ (medgenska regija), O (ORF), I (intron) in E (ekson, ki ni ORF). Radi bi napovedali strukturo genov na + niti še za regijo, kjer strukture ne poznamo.

   Obe regiji sta dolgi 100.000 baznih parov in izgledata takole:

```
Prva:  TAGGATAATGTCTT ... TTCAAGTATG
       _______OOOOOOO ... OOOOOIIIII

Druga: GCATGTAACCTTTT ... CTAACTCAAT
                - ne poznamo -
```

[4] (a) Opišite, kako napovedati strukturo genov na drugi regiji DNA. Bodite natančni!

[4] (b) Strukturo genov na drugi regiji DNA lahko lepo napovemo na dva načina oziroma z dvema različnima metodama, ki smo ju spoznali na predavanjih in vajah. V čem se razlikujeta rezultata, ki jih vrneta ti dve metodi?

[4] (c) Kako lahko ugotovimo, kateri način je boljši za podane podatke. Predlagaj način preverjanja in metriko (oceno kvalitete), ki bi jo pri tem opazoval!

Given are sequences of two DNA regions. The first sequence is sense-annotated with labels _ (intergenic region), O (ORFs), I (intron) in E (exon which is not ORF). Our goal is to predict the structure of the second sequence in the sense direction.

Both regions contain 100.000 base pairs and look something like:

```
First:  TAGGATAATGTCTT ... TTCAAGTATG
       _______OOOOOOO ... OOOOOIIIII

Second: GCATGTAACCTTTT ... CTAACTCAAT
           - annotation unknown -
```

[4] (a) Describe the procedure to predict sequence structure of the second region. Be exact and precise!

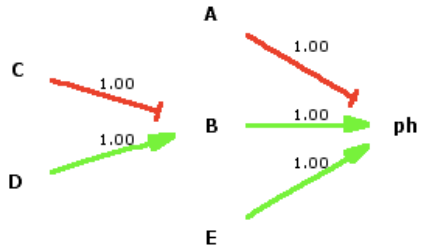
[4] (b) We have discussed on two possible approaches to predict the structure of the second region. What is the difference in the type of results these two methods return?

[4] (c) How would we determine which of the two different methods is better? Propose a scoring technique and a method with which you would assess the score and in this way estimate the performance.

3. 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.

   Na podlagi eksperimentalnih podatkov in pravil sklepanja o epistazi in vplivu posameznih genov na fenotip je možno zgraditi spodnjo gensko mrežo. Možni so trije fenotipi: n, 0, p. Fenotip 0 je fenotip divjega osebka (ang. wild-type).

ID	Gene 1	Gene 2	ph	Confidence	Comments	Ignore	Edit	Delete
E1			0	1.00		I	E	D
E2	A-		p	1.00		I	E	D
E3	B-		n	1.00		I	E	D
E4	C-		p	1.00		I	E	D
E5	D-		0	1.00		I	E	D
E6	E-		n	1.00		I	E	D
E7	C-	B-	n	1.00		I	E	D
E8	B-	D-	n	1.00		I	E	D



   Katere eksperimente bi bilo potrebno še izvesti in kakšni bi morali biti njihovi izidi (fenotip *ph*), da bi dokazali naslednje relacije (točke a do e). Če relacije ni možno dokazati, navedite razlog oz. kateri obstoječi eksperiment(i) nasprotujejo želeni relaciji.

[2] (a) $E \dashv A$

[2] (b) $E \rightarrow A$

[2] (c) $E \dashv B$

[2] (d) $C \dashv E$

[2] (e) $D \dashv A$

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 mutant-based experiments and the inferred gene regulation network (see table and figure on the previous page). The experiments report on three different phenotypes, n, 0, p, where phenotype 0 is a wild-type phenotype.

   Which experiments we would have to conduct and which would be their expected outcomes that would be, in addition to the set of existing experiments, needed to hypothesize a given set of relations (items a to e on the previous page). Also, where needed, specify if it is not possible to come up with such an experiment, or there is a conflicting existing experiments.

[8] 5. V danem zaporedju želimo poiskati vse možne bralne okvire in prevesti v zaporedje aminokislin na podlagi podane standardne tabele (za pričetek vzemite le `ATG`, za konec pa {`TAA`,`TAG`,`TGA`}). Poročajte 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 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

`TAAATGCGTGAGTCTATTACTGAGGTTTAAGAATGAATATTTAAGCTTAAGCAACAGCACGAGGCATAAAT`

Given is a nucleotide sequence for which we would like to find all open reading frames. Assume `ATG` for a start codon, and {`TAA`, `TAG`, and `TGA`}) for the end codons. Report only on proteins with at least four aminoacids.

6. V študijo prehranjevalnih navad smo vključili 10 oseb. Štiri osebe so se prehranjevale kot običajno, šest pa jih je jedlo le en (ogromen) obrok dnevno. Študija je trajala teden dni.

   Zanimal nas je vpliv prehrane na izražanje genov *SIRT1* in *PPAR*. Izražanje teh genov smo izmerili in opazili, da je *SIRT1* nadpovprečno izražen pri petih osebah, od katerih so trije jedli le en obrok dnevno. Gen *PPAR* se je visoko izrazil pri treh osebah, od katerih sta dva jedla le en obrok dnevno.

[3] (a) Oceni (in oceno utemelji - lahko pa tudi izračunaš) na katerega od genov *SIRT1* in *PPAR* je genska sprememba bolj vplivala?

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

[2] (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.

In the study we have examined eating habits of 10 different students. Four students eat normally, while six eat only one (extra-large) meal per day. The study lasted for one week.

We were interested about the influence of the eating habit to expression of the genes *SIRT1* and *PPAR*. We have measured their expression and noticed that gene *SIRT1* is overexpressed with five different students, from whcih three eat only one meal daily. Gene *PPAR* is overexpressed with three different students, of which two ate only one meal daily.

[3] (a) On which of the two genes does the eating habit have a larger influence (estimate and explain; no need to derive precise numbers).

[3] (b) What is the probability that such a result would be obtained by chance?

[2] (c) Give an equation with which you can estimate the probability that you would obtain the same or better result as in our experiment. You do not need to compute the result.

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