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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

2. izpitni rok / First Examination Period

13. februar 2013 / February 13, 2013

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	5	4	6	8	6	29
Točk / Points						

[5] 1. Construct a hidden Markov model from hidden and observable (here DNA) sequences below.

Zgradite skriti Markov model iz skritega in vidnega zaporedja DNA, ki sta zapisani spodaj.

```
IIIIGGGGGGGGGGGGGGGGIIIIIIIIIIII
GTATATGGTAGAACGATATTGATAACAATTCTAT
```

[4] 2. We wanted to evaluate the predictive performance of our model based on data in the table below. The data include observed mutant phenotypes and phenotypes predicted with our model.

After we calculated recall to be 0.75 (3/4), we accidentally erased the information on the observed phenotype for mutants m2, m6 and m9. These are now marked as "?".

Try to determine the missing data on observed phenotypes (m2, m6 and m9). After you fill in the missing data, calculate the other measure of predictive performance: precision.

Na podlagi podatkov o opazovanem (stolpec "observed phenotype") in napovedanem fenotipu (stolpec "predicted phenotype") v tabeli smo želeli izračunati priklic in točnost našega napovednega modela.

Najprej smo izračunali priklic (0.75 oz. 3/4), nakar pa smo po nesreči zbrisali podatke o opazovanem fenotipu mutantov m2, m6 in m9. Le-ti so zdaj označeni z "?".

Določite vrednosti manjkajočih fenotipov za m2, m6 in m9. Nato izračunajte še točnost.

mutant	observed phenotype	predicted phenotype
m1	+	+
m2	?	-
m3	-	-
m4	-	+
m5	+	+
m6	?	+
m7	-	+
m8	-	-
m9	?	+
m10	+	+

precision = $\frac{TP}{(TP+FP)}$, recall = $\frac{TP}{(TP+FN)}$

[6] 3. Given are two sequences

```
TCAGAC
CCATAGGC
```

and a scoring function

$$\sigma(a,b) = \begin{cases} 1 & a = b \\ -1 & a = - \text{ ali } b = - \\ 0 & \text{otherwise} \end{cases}$$

Propose 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:

```
TCAGAC
CCATAGGC
```

in ocenjevalna funkcija

$$\sigma(a,b) = \begin{cases} 1 & a = b \\ -1 & a = - \text{ ali } b = - \\ 0 & \text{v ostalih primerih} \end{cases}$$

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

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

Page for your solutions. / Stran za vaše rešitve.

4. Jukes-Cantor is expressed as:

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

[2] (a) What are d_{JC} and d in this equation?

[2] (b) Is d_{JC} expected to be bigger or smaller than d? Why?

[2] (c) A hypervariable region of mitochondrial genome is particularly well suited to measure the evolutionary distance between species. Why? What is a function of this region?

[2] (d) How does a variability of a human genome in Africans compare to inhabitants of Chile? Why?

Korekcijska enačba po Jukes-Cantor je:

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

(a) Kaj predstavljata spremenljivki d_{JC} in d?

(b) Ali pričakujemo, da je d_{JC} manjša ali večja od d? Zakaj?

(c) V namene opazovanja razdalj med vrstami je še posebej primerna hipervariabilna regija mitohondrijske DNA. Zakaj? Kakšno funkcijo ima ta regija?

(d) Primerjajte variabilnost humanega genoma prebivalcev Afrike in prebivalcev Čila. Odgovor utemeljite!

5. We wanted to determine if the nucleotide sequence TAA appears in the human mitochondria more frequently than expected by chance (given a multinomial model). We obtained reference (null) distribution of probability of TAA, which is shown in the histogram below. We then compared the histogram to the measured frequency of TAA on true data (0.025, vertical line).

[2] (a) How did we obtain the null distribution?

[1] (b) Estimate how many samples of the null hypothesis we have generated to show the histogram. In other words, how many times did we measure the probability of TAA?

[2] (c) How can we obtain the p-value (0.090) from the distribution shown on figure?

[1] (d) Can we say that sequence TAA does not appear that frequently? Why?

Želeli smo ugotoviti ali se v človeškem mitohondriju zaporedje nukleotid TAA pojavi pogosteje, kot bi pričakovali po naključju (ob predpostavki multinomskega modela). Pridobili smo referenčno (ničelno) porazdelitev verjetnosti TAA, kar kaže spodnji histrogram, in jo primerjali z izmerjeno verjetnostjo na pravih podatkih (0.025, navpična črta).

(a) Kako smo pridobili ničelno porazdelitev?

(b) Oceni, kolikokrat smo morali generirati vrednosti ob ničelni hipotezi, da smo dobili podatke za prikazani histogram. Z drugimi besedami, kolikokrat smo morali v ta namen izmeriti verjetnost TAA?

(c) Kako lahko iz distribucije na sliki razberemo p-vrednost (0.090)?

(d) Ali lahko rečemo, da se zaporedje TAA ne pojavlja posebej pogosto? Zakaj?