

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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

3. izpitni rok / First Examination Period

18. september 2013 / September 18, 2013

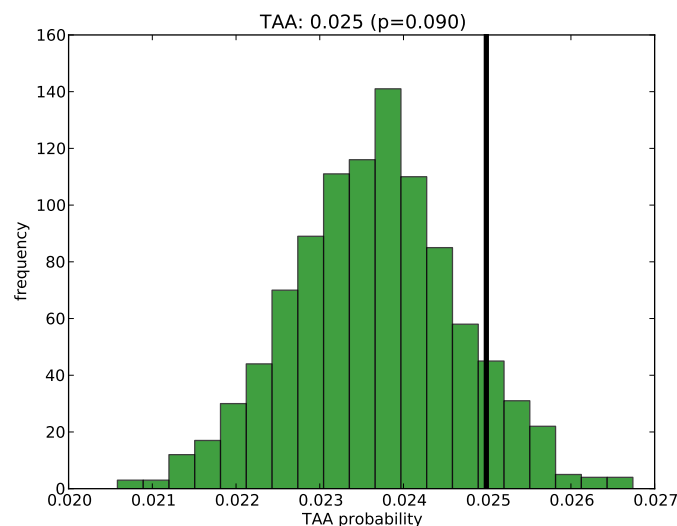
Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	6	5	6	6	6	29
Točk / Points						

1. 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 histogram, in jo primerjali z izmerjeno verjetnostjo na pravih podatkih (0.025, navpična črta).

- (a) Kako smo pridobili ničelno porazdelitev?
- (b) Oцени, 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?



Page for your solutions. / Stran za vaše rešitve.

- [5] 2. You are given a Hidden Markov Model for coin flipping, for both a fair and a biased coin.

For the fair coin (hidden state f), the probabilities of Tail (T) and Head (H) are the same. For the biased coin (hidden state b), the probability of Tail (T) is much higher.

$$t_{0f} = 0.5; t_{0b} = 0.5$$

$$t_{ff} = 0.95; t_{fb} = 0.05$$

$$t_{bf} = 0.1; t_{bb} = 0.9$$

$$e_f(T) = 0.5; e_f(H) = 0.5$$

$$e_b(T) = 0.9; e_b(H) = 0.1$$

What is the probability to obtain this sequence of four flips $x = \text{TTHH}$?

Podan imate skriti Markov model za metanje dveh kovancev, poštenega in goljufivega. Pri poštemem (skrito stanje p) sta verjetnost cifre (C) in glave (G) enaki, pri goljufivem (skrito stanje g) je cifra (C) dosti bolj verjetna.

$$t_{0p} = 0.5; t_{0g} = 0.5$$

$$t_{pp} = 0.95; t_{pg} = 0.05$$

$$t_{gp} = 0.1; t_{gg} = 0.9$$

$$e_p(C) = 0.5; e_p(G) = 0.5$$

$$e_g(C) = 0.9; e_g(G) = 0.1$$

Kakšna je verjetnost, da pri štirih metih kovanca vidimo zaporedje $x = \text{CGGG}$?

	Viterbi	Forward
Initialisation:	$v_0(0) = 1, v_k(0) = 0$ for $k \neq 0$	$f_0(0) = 1, f_k(0) = 0$ for $k \neq 0$
$i = 1 \dots L$:	$v_l(i) = e_l(x_i) \max_k (v_k(i-1) t_{kl})$	$f_i(i) = e_l(x_i) \sum_k f_k(i-1) t_{kl}$
Termination:	$P_v = \max_k (v_k(L))$	$P_f = \sum_k f_k(L)$

Page for your solutions. / Stran za vaše rešitve.

- [6] 3. A hypothetical organism has ten genes: A, B, C, D, E, F, G, H, I, J. We know that genes A, B, C, D, E are involved in reproduction, while genes E, F, G are involved in metabolism. A study discovered that genes B, C, E, F, I, J are expressed differently if the temperature is changed.

Compute the probability of finding as many or more genes from a biological process among differently expressed genes by chance (for both reproduction and metabolism separately). Which process is therefore more likely to be influenced by the temperature change?

Hipotetični organizem ima deset genov: A, B, C, D, E, F, G, H, I, J. Vemo, da geni A, B, C, D, E sodelujejo v procesu razmnoževanja, medtem ko so geni E, F, G potrebni za presnovo. Odkrili so, da se geni B, C, E, F, I, J ob spremembi temperature izražajo drugače.

Izračunaj verjetnost, da bi naključno našli vsaj toliko ali več genov biološkega procesa med drugače izraženimi geni (za razmnoževanje in presnovo, vsako posebej). Na kateri proces verjetneje vpliva sprememba temperature?

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

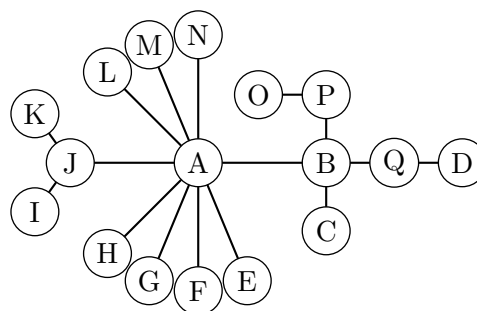
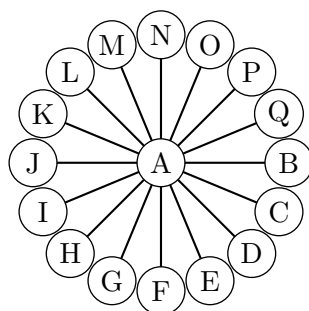
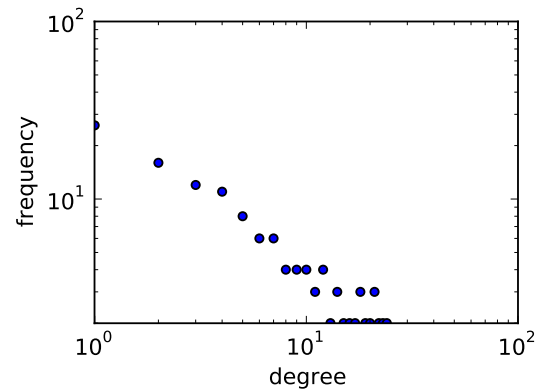
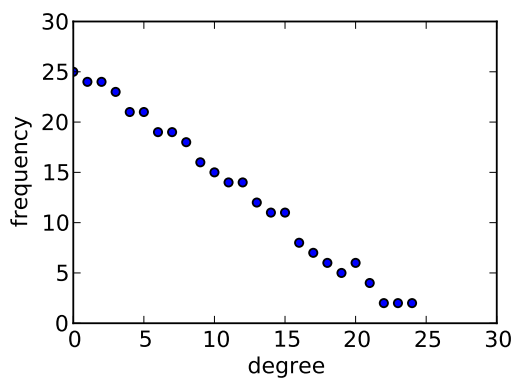
$$\binom{N}{n} = \frac{N!}{(N-n)!n!}$$

- [2] 4. (a) What are genetic algorithms? How do they work?
- [2] (b) What purpose do they serve in computer science?
- [2] (c) Why do we say they were inspired by evolution?

-
- (a) Kaj so genetski algoritmi? Kako delujejo?
- (b) Pri čem jih uporabljamo v računalništvu?
- (c) Zakaj pravimo, da jih je navdihnila evolucija?

- [2] 5. (a) Which of the degree distribution graphs shows a scale free network? Why?
- [2] (b) You have a random networks with n nodes and e edges. Which network would have a smaller diameter: a scale free network or an Erdős-Renyi network (edges are added independently of current edges)? Why?
- [2] (c) Describe an algorithm that can be used to find communities (clusters) in a network. Mark the clusters on networks below.

-
- (a) Kateri graf porazdelitve števila povezav prikazuje mrežo "scale free"? Zakaj?
- (b) Imate naključno generirane mreže z n vozlišči in e povezavami. Bi imela manjši polmer mreža "scale free" ali mreža Erdős-Renyi (povezave so dodane neodvisno od obstoječih povezav)? Zakaj?
- (c) Opiši algoritem za iskanje skupin v mrežah. Označi skupine na spodnjih mrežah.



Page for your solutions. / Stran za vaše rešitve.

Page for your solutions. / Stran za vaše rešitve.