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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

1. izpitni rok / First Examination Period

30. januar 2013 / January 30, 2013

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	5	8	6	6	5	30
Točk / Points						

[5] 1. 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`

Vaša naloga je najprej določiti položaje genov (ORFov) za tri znane proteine. Nato uporabite podatke o aminokislinskem zaporedju proteinov in karseda natančno določite vrednost čimveč manjkajočih nukleotidov. V primeru, da nukleotida ni možno določiti nedvoumno, navedite razloge zakaj ni možno.

Pomagajte si s standardno kodno tabelo (spodaj, za pričetek ORF-a vzemite 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

2. Given is a set of genes ($g_1 \ldots g_5$) and their (normalized and scaled) expression at four different conditions.

	cold	hot	acid	bacteria
g_1	1	0	0	2
g_2	2	0	3	4
g_3	0	0	1	2
g_4	4	1	2	1
g_5	0	1	4	3

We would like to represent genes as points in Euclidean space so that genes with similar profile would be placed closer together than genes with different profile.

[2] (a) Propose a criteria function that we would like to optimize in the proposed projection. Express this function in an equation and explain the terms used.

[4] (b) Draw an Euclidean plane and propose a projection for the five genes. Projection should be based on some gene distance scoring. Propose an appropriate scoring and roughly estimate the distances (present them in a distance matrix).

[1] (c) What is the name of the algorithm that can solve such a problem?

[1] (d) What are the gains of such visual representations when compared to a dendrogram obtained through hierarchical clustering?

Dan je nabor genov ($g_1 \ldots g_5$) z njihovimi izraznimi profili oziroma izrazi izmerjenimi pri štirih različnih pogojih (tabela).

Gene bi želeli predstavili kot točke v Evklidski ravnini tako, da so si geni s podobnimi profili blizu in ti z različnimi profili daleč vsaksebi.

(a) Predlagajte kriterijsko funkcijo, ki jo optimiziramo pri taki predstavitvi (projekciji).

(b) Predlagajte primerno projekcijo in postavitev točk v Evklidski ravnini. Projekcija mora temeljiti na oceni razdalj med geni. Predlagajte ustrezno metriko razdalj, izračunajte matriko razdalj in na osnovi te predlagajte ustrezno projekcijo.

(c) Kako se imenuje algoritem, s katerim lahko nalogo te vrste rešite programsko?

(d) Kakšne so prednosti take predstavitve pred predstavitvijo dendrograma hierarhičnega razvrščanja?

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. Odkrilo 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?

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

4. Two networks are given below.

[2] (a) Which of the two networks below is more likely scale free? Why?

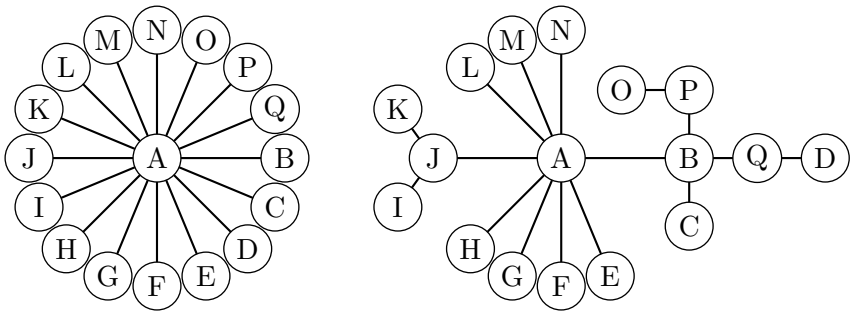
[1] (b) What are the diameters of both networks?

[3] (c) Describe an algorithm that can be used to find communities (clusters) in a network. Mark the clusters on networks below.

(a) Katera od spodnjih mrež je verjetneje "scale free". Zakaj?

(b) Kakšna sta premera obeh mrež?

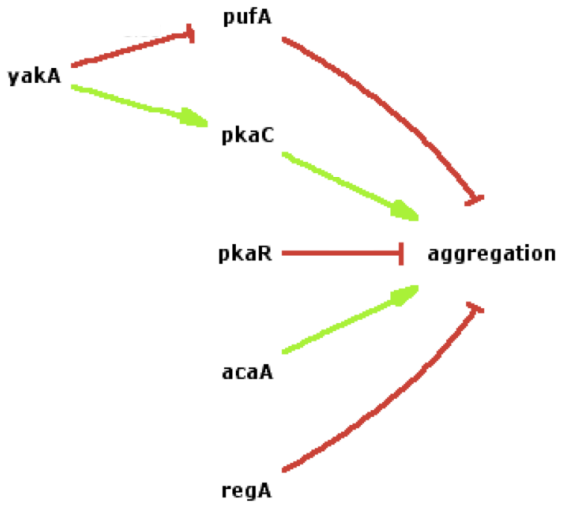
(c) Opiši algoritem za iskanje skupin v mrežah. Označi skupine na spodnjih mrežah.



5. Given is a set of 15 experiments, where we have observed a phenotype (aggregation) for the wild type organism (E1), different single mutants (E2 to E9), and different double mutants (E10 to E15). Genes were either knocked-out (e.g, yakA-) or overexpressed (e.g., acaA+). Phenotypes are qualitative, where "-" means no aggregation, "±" suppressed aggregation, "+" normal aggregation and "++" rapid aggregation.

ID	Gene 1	Gene 2	aggregation
E1			+
E2	yakA-		-
E3	pufA-		++
E4	pkaR-		++
E5	pkaC-		-
E6	acaA-		-
E7	regA-		++
E8	acaA+		++
E9	pkaC+		++
E10	pkaC-	regA-	-
E11	yakA-	pufA-	++
E12	yakA-	pkaR-	±
E13	yakA-	pkaC-	-
E14	pkaC-	yakA+	-
E15	yakA-	pkaC+	++

We used the experiments and inferred the following gene regulation network. But we overlooked (ignored) one of the experiments!



[2] (a) Which experiment was ignored in the construction of the network?

[3] (b) Correct the network to include the ignored experiment.

Razpredelnica podaja nabor 15ih eksperimentov, kjer smo opazovali fenotip "aggregation" pri nemutiranem organizmu (E1), enojnih (E2 do E9) in dvojnih mutantih (E10 do E15). Gene smo

pri eksperimentih ali izničili (npr. yakA-) ali jih čezmerno izrazili (npr. acaA+). Opazovani fenotip smo zajeli kvalitativno z vrednostmi "-" (no aggregation), "±" (suppressed aggregation), "+" (normal aggregation) in "++" (rapid aggregation).

Na osnovi eksperimentov smo zgradili mrežo genskih regulacij (slika), a pri tem pozabili na en eksperiment.

(a) Na kateri eksperiment smo pri gradnji mreže pozabili?

(b) Popravi mrežo tako, da pravilno upoštevaš tudi pozabljeni eksperiment!

Page for your solutions. / Stran za vaše rešitve.