Osnove bioinformatike / Introduction to Bioinformatics

1. izpitni rok / First Examination Period

1. februar 2012 / February 1st, 2012

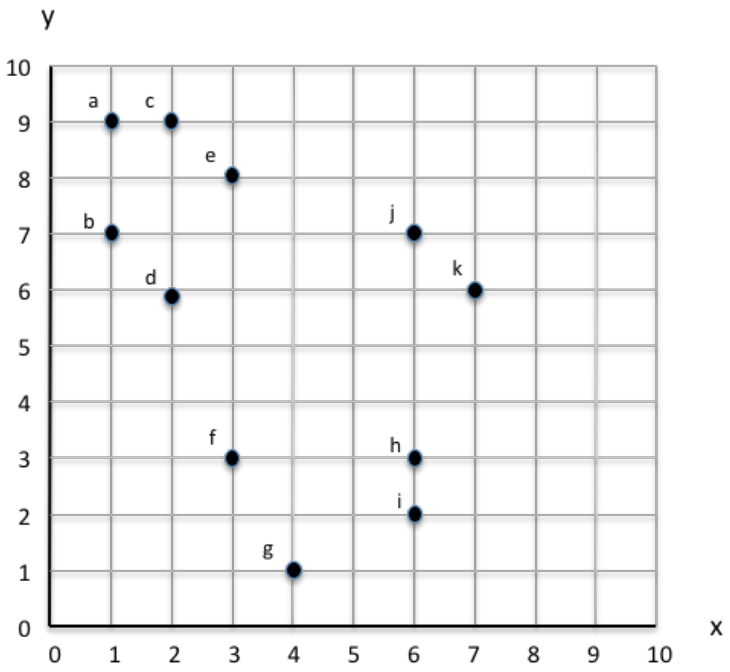
Ime in priimek / Name: _____

Vpisna številka / Student ID: _____

Ocene nalog / Grading Table

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	20	10	16	10	7	63
Točk / Points						

1. Na različnih vrstah nekega modelnega organizma (a do k) smo izmerili izražanje dveh genov, ter te po normalizaciji in skaliranju zapisali s spremenljivkami x in y. Rezultate naših eksperimentov lahko predstavimo kot točke v Evklidski ravnini:



[8] (a) Izriši dendrogram, ki ga dobiš z hierarhičnim razvrščanjem točk (vrst modelnega organizma) v skupine. Kot mero za podobnost uporabi Manhattansko razdaljo, kjer je razdalja med točkama i in j določena kot $d_{ij} = |x_i - x_j| + |y_i - y_j|$. Podobnost med dvema skupinama meri s tehniko maksimalne razdalje med pari točk iz različnih skupin (t. im. *complete linkage*).

[2] (b) Uporabi izrisani dendrogram in na podlagi njega predlagaj razdelitev primerov v tri skupine (na dendrogramu izriši vertikalo, ki točke razdeli v tri skupine). Izpiši, kateri primeri pripadajo posamezni skupini.

We have observed the expression of two genes at different types of the model organism (a to k), and after the data preprocessing (normalization and scaling) recorded these with variables x and y. The results can be presented in the Euclidean space (see figure).

[8] (a) Draw the dendrogram that visualizes the results of hierarchical clustering of types of model organism. Use Manhattan distance ($d_{ij} = |x_i - x_j| + |y_i - y_j|$). Use complete linkage to estimate the similarity between two clusters.

[2] (b) Use the dendrogram to propose the split of the set of model organisms into three groups. Mark this split with the vertical line in the histogram. Report on cluster membership.

[10] 2. Dani sta zaporedji:

```
AATAAGTTA
GACTGTA
```

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).

Given are two sequences

```
AATAAGTTA
GACTGTA
```

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 and report on alignment score.

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

3. V študiji gensko spremenjenega paradižnika smo vključili $N = 10$ primerkov. Šest paradižnikov ($m = 6$) je bilo takih, ki so gensko spremenjeni, štiri ($N - m = 4$) pa smo si sposodili iz vrta sosede. Soseda je po poklicu frizerka in se v prostem času ne ukvarja z genetiko.

   Zanimal nas je vpliv genskih sprememb na izražanje genov *obi1* in *obi2*. Izražanje teh genov smo zmerili in opazili, da je *obi1* nadpovprečno izražen pri petih paradižniki ($n = 5$), od katerih so trije ($k = 3$) gensko spremenjeni. Gen *obi2* se je visoko izrazil pri treh paradižnikih ($n = 3$), od katerih sta dva ($k = 2$) gensko spremenjena.

[2] (a) Oceni (lahko pa tudi izračunaš), na katerega od genov *obi1* in *obi2* je genska sprememba bolj vplivala?

[4] (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.

We have conducted a study of effects of genetic modification of tomato. From $N = 10$ tomatoes in the study, $m = 6$ were genetically modified, while $N - m = 4$ where borrowed from our neighbor's garden. Our neighbor is a hair dresser and she is (otherwise) not involved in genetic experiments.

Our aim was to study the effect of genetic modification to the expression of the genes *obi1* and *obi2*. We have noticed tat *obi1* is overexpressed with five tomatoes ($n = 5$), od which three ($k = 3$) are genetically modified. Gene *obi2* was overexpressed with three tomatoes ($n = 3$), of which two ($k = 2$) were genetically modified.

[2] (a) Estimate (or, if really needed, compute) on which of the gene expressions (e.g., *obi1* vs *obi2*) the genetic modification has a larger effect.

[4] (b) Is this effect arbitrary? How likely it would be obtained if the expression of the two genes would not be related to our experiment?

[2] (c) How would you compute probability of obtaining the same or better result (in terms of number of tomatoes where the two genes are overexpressed). Propose the formula, you do not need to compute the probability.

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

4. 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.

[2] (a) Kakšna je omejitev sklepanja o epistazi glede na število opazovanih genov?

A. Epistaza je relacija med parom genov, postopek je torej neodvisen od števila opazovanih genov.

B. Epistazo lahko opazujemo le, če imamo v podatkih natančno dva gena.

C. Epistazo lahko opazujemo le na manjših naborih podatkov.

D. Epistazo lahko opazujemo le, če smo v eksperimentih opazovali do 10 genov.

E. Ta omejitev je odvisna od opazovanega organizma.

[3] (b) Fenotip mutante a^-b^- in mutante b^- je zavrta rast. Rast nemutiranega organizma je normalna. Rast mutante a^- je pospešena. Kakšna je relacija med genoma a, b in fenotipom ($\rightarrow$ označuje eksitacijo oziroma pozitivno regulacijsko zvezo, $-|$ pa inhibicijo oziroma negativno regulacijsko zvezo)?

A. gena a in b sta na paralelnih poteh za rast

B. $b -| a \rightarrow$ rast

C. $b \rightarrow a -|$ rast

D. $a -| b \rightarrow$ rast

E. $a \rightarrow b \ -|$ rast

Gene regulation networks can be inferred through epistasis, where one gene can block the other one in the pathway for a specific phenotype.

[2] (a) Is epistasis analysis limited to the number of observed genes?

A. Epistasis is a relation between pair of genes, and the inference procedure does not depend on the total number of observed genes.

B. Epistasis can be observed only if the data includes exactly two genes.

C. Epistasis can be studied only a smaller data sets.

D. Epistasis can be observed only on the data sets of about 10 genes.

E. This limitation is organism-specific.

[3] (b) Phenotype of a a^-b^- mutant and of mutant b^- is inhibited growth. The growth of a wild-type organism is normal. Growth of a mutant a^- is accelerated. What is the relation between genes a, b and the phenotype? Below, $\rightarrow$ marks excitation (positive regulation) and $-|$ inhibition (negative regulation)?

A. genes a and b act in parallel

B. $b -| a \rightarrow$ rast

C. $b \rightarrow a -|$ rast

D. $a -| b \rightarrow$ rast

E. $a \rightarrow b \ -|$ rast

5. Iz krajših homolognih genskih zaporedjih petih različnih vrst ($N = 5$) smo ocenili genske razdalje (d_{ij}), ki jih podaja spodnja tabela:

	B	C	D	E
A	5	4	9	8
B		5	10	9
C			7	6
D				7

[5] (a) Na podlagi algoritma združevanja najbližjih sosedov (angl. *neighbor-joining method*) oceni, katere dve vrsti sta si evolucijsko najbližje. Določi tudi oddaljenost teh dveh vrst do njunega skupnega prednika (to je, do novega vozlišča, ki ga uvedeš ob združitvi najbližjih vrst v filogenetsko drevo. Vprašanje sprašuje po oddaljenost ene in druge vrste do tega prednika, kjer ti razdalji tipično nista enaki).

[2] (b) Algoritem združevanja najbližjih sosedov gradi filogenetska drevesa tako, da pri tem zasleduje specifičen cilj. Kakšen cilj je to oziroma kaj optimizira ta algoritem?

We used shorter homolog gene sequences to estimate the genetic distances d_{ij} for five ($N = 5$) different species (see the table above).

[5] (a) Use neighbor-joining method to estimate which of the two species are evolutionary the closest. Estimate their distance to their joint predecessor (that is, to the node that is introduced after joining of two closest species in the phylogenetic tree. Notice that the two distances are typically not equal to each other).

[2] (b) What does neighbor-joining algorithm try to optimize?

$$U_i = \sum_{j=1}^{N} d_{ij}$$

$$D_{ij} = d_{ij} - \frac{U_i + U_j}{N-2}$$

$$d_{ik} = \frac{1}{2}\Big(d_{ij} + \frac{U_i - U_j}{N-2}\Big)$$

$$d_{jk} = d_{ij} - d_{ik}$$