Name, SURNAME and ID ⇒ SOLUTIONS

Middle East Technical University
Department of Computer Engineering

CENG 465

Introduction to Bioinformatics

Spring '2005-2006

Midterm Exam II

- **Duration: 50** minutes.
- **Exam:**
  - This is a **closed book, closed notes** exam. The use of any reference material is strictly forbidden.
  - No attempts of cheating will be tolerated. In case such attempts are observed, the students who took part in the act will be prosecuted. The legal code states that students who are found guilty of cheating shall be expelled from the university for **a minimum of one semester**!
- **About the exam questions:**
  - The points assigned for each question are shown in parenthesis next to the question.
- **This exam consists of *6* pages including this page. Check that you have them all!**
- **GOOD LUCK !**

Question 1	40
Question 2	40
Question 3	20
Total ⇒	100

1 **(40 pts)**

40

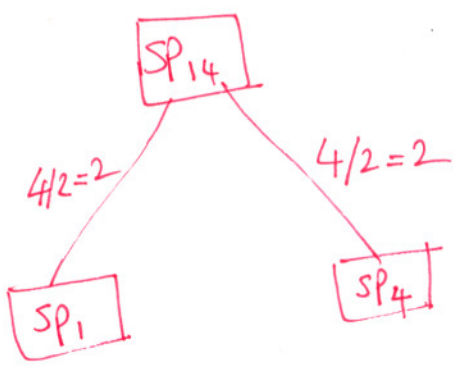
(a)(20 pts) Given the following distance matrix **D** for the four species sp_1, sp_2, sp_3, and sp_4, construct a phylogenetic tree using the UPGMA method. Show intermediate matrices and subtrees in your construction.

D:

	sp_1	sp_2	sp_3	sp_4
sp_1	0	12	10	4
sp_2	12	0	6	14
sp_3	10	6	0	12
sp_4	4	14	12	0

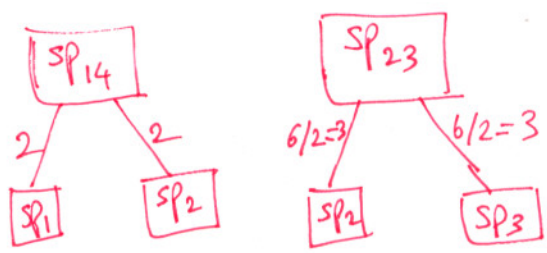
step (1) sp_1 and sp_4 are closest, merge them.

	sp_{14}	sp_2	sp_3
sp_{14}	0	(12+14)/2 = 13	(10+12)/2 = 11
sp_2	13	0	6
sp_3	11	6	0

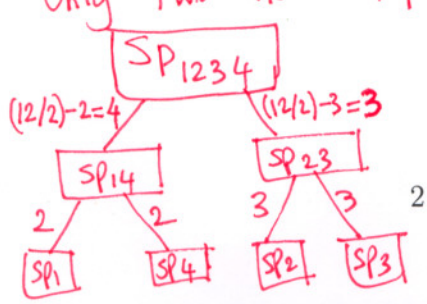


step (2) sp_2 and sp_3 are closest, merge them

	sp_{14}	sp_{23}
sp_{14}	0	(13+11)/2 = 12
sp_{23}	12	0



step (3) Only two nodes left to merge, merge them



(b)(10 pts) Based on the tree you constructed in part **a** fill-in the following *path distance matrix* **d** using the path distances between the four species.

d:

	sp_1	sp_2	sp_3	sp_4
sp_1	0	12	12	4
sp_2	12	0	6	12
sp_3	12	6	0	12
sp_4	4	12	12	0

(c)(10 pts) How close is the tree you generated in part **a** to the actual (i.e., true) phylogenetic tree which preserves the original distances between the four species? In other words, what is the error between matrix **D** and matrix **d** using LSQ (least squares) error measure? (You may use upper or lower triangle of the matrix when computing the LSQ error.)

(upper triangle)

$$\text{LSQ error} = \underbrace{(12-12)^2}_{0} + \underbrace{(12-10)^2}_{4} + \underbrace{(4-4)^2}_{0} + (6-6)^2 + (14-12)^2 + \underbrace{(12-12)^2}_{0} = 4+4 = \boxed{8}$$

$$LSQ = \sum\sum (d_{ij} - D_{ij})^2$$

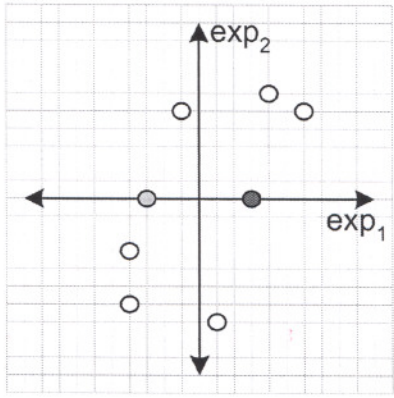
2 **(40 pts)**

40

Consider the following gene expression values from 2 microarray experiments for 8 genes.

	exp_1	exp_2
$gene_1$	-4	-3
$gene_2$	6	5
$gene_3$	1	-7
$gene_4$	-4	-6
$gene_5$	4	6
$gene_6$	-1	5
$gene_7$	-3	0
$gene_8$	3	0

Negative expression values in the table mean that the gene is downregulated (i.e., expressed less) in the experimented cell, and positive values mean that the gene is upregulated (i.e., overexpressed). A biologist is trying to find out whether these 8 genes can be separated into two groups based on their behavior in the experimented conditions. In order to visualize the relationships, she sketched a 2-dimensional plot of the genes shown below.



Use **k-means clustering** to cluster these **8** genes into **2** clusters. Use $gene_7$ (light grey data point on the 2-d plot) as the initial cluster center for $cluster_1$ and use $gene_8$ (dark grey data point) as the initial cluster center for $cluster_2$. Indicate which data point belongs to which cluster and also give the coordinates of the centroids at each iteration. Iterate until convergence, i.e., until no change in clusters. When computing distances to centroids, you may use the squared distance (no need to take the square-root).

iteration 1: Calculate (squared) distances to cluster centroids

cluster 1: $gene_1 = (-4-(-3))^2+(-3-0)^2 = 10$

$gene_2 = (6-(-3))^2+(5-0)^2 = 106$

$gene_3 = (1-(-3))^2+(-7-0)^2 = 65$

$gene_4 = (-4-(-3))^2+(-6-0)^2 = 37$

$gene_5 = (4-(-3))^2+(6-0)^2 = 85$

$gene_6 = (-1-(-3))^2+(5-0)^2 = 29$

$gene_7 = 0$

$gene_8 = (3-(-3))^2+(0-0) = 36$

cluster 2: $gene_1 = (-4-3)^2+(-3-0)^2 = 58$

$gene_2 = (6-3)^2+(5-0)^2 = 34$

$gene_3 = (1-3)^2+(-7-0)^2 = 53$

$gene_4 = (-4-3)^2+(-6-0)^2 = 85$

$gene_5 = (4-3)^2+(6-0)^2 = 37$

$gene_6 = (-1-3)^2+(5-0)^2 = 41$

$gene_7 = (-3-3)^2+(0-0)^2 = 36$

$gene_8 = 0$

cluster 1 and cluster 2 distances are compared and cluster assignment are shown in boxes.

Extra page

Compute new cluster centroids:

cluster1 : $gene_1, gene_4, gene_6, gene_7$

$$(x,y) = \left(\frac{-4-4-1-3}{4}, \frac{-3-6+5+0}{4}\right) = (-3,-1)$$

cluster2 : $gene_2, gene_3, gene_5, gene_8$

$$(x,y) = \left(\frac{6+1+4+3}{4}, \frac{5-7+6+0}{4}\right) = (3.5,1)$$

iteration 2 : calculate new squared distances to cluster centroids:

cluster1 :
- gene1 $= (-4+3)^2 + (-3+1)^2 = 5$
- gene2 $= (6+3)^2 + (5+1)^2 = 117$
- gene3 $= (1+3)^2 + (-7+1)^2 = 52$
- gene4 $= (-4+3)^2 + (-6+1)^2 = 26$
- gene5 $= (4+3)^2 + (6+1)^2 = 98$
- gene6 $= (-1+3)^2 + (5+1)^2 = 40$
- gene7 $= (-3+3)^2 + (0+1)^2 = 1$
- gene8 $= (3+3)^2 + (0+1)^2 = 10$

cluster2 :
- gene1 $= (-4-3.5)^2 + (-3-1)^2 > 5$
- gene2 $= (6-3.5)^2 + (5-1)^2 < 117$
- gene3 $= (1-3.5)^2 + (-7-1)^2 > 52$
- gene4 $= (-4-3.5)^2 + (-6-1)^2 > 26$
- gene5 $= (4-3.5)^2 + (6-1)^2 < 98$
- gene6 $= (-1-3.5)^2 + (5-1)^2 = 36.25$
- gene7 $= (-3-3.5)^2 + (0-1)^2 > 1$
- gene8 $= (3-3.5)^2 + (0-1)^2 < 10$

cluster 1 and cluster 2 distances are compared and new cluster assignments are shown in boxes.

Compute new cluster centroids:

cluster1 : gene1, gene3, gene4, gene7

$$(x,y) = \left(\frac{-4+1-4-3}{4}, \frac{-3-7-6+0}{4}\right) = (-2.5,-4)$$

cluster2 : gene2, gene5, gene6, gene8

$$(x,y) = \left(\frac{6+4-1+3}{4}, \frac{5+6+5+0}{4}\right) = (3,4)$$

iteration 3: calculate new squared distances to cluster centroids:

cluster1 :
- gene1 $= (-4+2.5)^2 + (-3+4)^2 < 98$
- gene2 $= (6+2.5)^2 + (5+4)^2 > 10$
- gene3 $= (1+2.5)^2 + (-7+4)^2 < 13.5$
- gene4 $= (-4+2.5)^2 + (-6+4)^2 < 149$
- gene5 $= (4+2.5)^2 + (6+4)^2 > 5$
- gene6 $= (-1+2.5)^2 + (5+4)^2 > 17$
- gene7 $= (-3+2.5)^2 + (0+4)^2 < 52$
- gene8 $= (3+2.5)^2 + (0+4)^2 > 16$

cluster2 :
- gene1 $= (-4-3)^2 + (-3-4)^2 = 98$
- gene2 $= (6-3)^2 + (5-4)^2 = 10$
- gene3 $= (1-3)^2 + (-7-4)^2 = 125$
- gene4 $= (-4-3)^2 + (-6-4)^2 = 149$
- gene5 $= (4-3)^2 + (6-4)^2 = 5$
- gene6 $= (-1-3)^2 + (5-4)^2 = 17$
- gene7 $= (-3-3)^2 + (0-4)^2 = 52$
- gene8 $= (3-3)^2 + (0-4)^2 = 16$

$\Rightarrow$ clusters don't change - STOP.

3 **(20 pts)**

20

Short answer questions:

(a)(10 pts) When examining the quality of a structural alignment, is it enough to inspect RMSD alone? Why/Why not?

It is NOT enough to inspect RMSD alone. Because we also need the length of the alignment. It is possible to get very good RMSD values for very short alignments. However, this does not imply a high quality alignment. Therefore, length of alignment is a critical information that should be inspected along with RMSD.

(b)(10 pts) What is the difference between single-linkage and complete-linkage clustering techniques?

In single-linkage clustering, the distance between two clusters is defined as "the distance between the closest members" of respective clusters. However, in complete-linkage clustering, the distance is defined as "the distance between the farthest members" of the respective clusters.