Introduction to bioinformatics, Autumn 2006, Exercise 4

13.10.2006

1. (Chapter 12, Exercise 11) This exercise illustrates an important principle in calculating likelihoods on trees. Probability of observing particular bases at the leaves of a tree with $n = 3$ species is

$$p(i_1, i_2, i_3) = \sum_a \sum_b \pi_a q_{ai_3}(t_2) q_{ab}(t_2 - t_1) q_{bi_2}(t_1) q_{bi_1}(t_1).$$

   This can also be written in the form

$$p(i_1, i_2, i_3) = \sum_a \pi_a q_{ai_3}(t_2) \sum_b q_{ab}(t_2 - t_1) q_{bi_2}(t_1) q_{bi_1}(t_1).$$

   Evaluate carefully how many addition and multiplication operations are performed in these two formulae, and deduce that the first form is less efficient than the second.

2. (Chapter 11, Exercise 3) Perform hierarchical clustering of the yeast data in Section C.4. Use the Euclidean distance metric with standardized data. How does cluster membership obtained with hierarchical clustering compare with the result from K-means (Computational Example 11.2)?

   You can find the data in the course folder in the course material room C127.

3. (Chapter 11, Exercise 4) Repeat the calculation in the previous exercise, except use the correlation coefficients to calculate distances. [Hint: Use the R function as.dist().] In previous calculations using these data, we standardized expression levels for each gene (row) prior to clustering. Is this standardization needed for hierarchical clustering when correlation coefficients are used for distances? Why or why not?

4. (Chapter 11, Exercise 7) Intensity measurements from a single slide (Fig. 11.5) for four replicated features corresponding with the *twist* gene are presented below. Nucleic acid corresponding with dsd^D flies was labeled with Cy5 (R), and nucleic acid corresponding with wild-type flies was labeled with Cy3 (G).

Feature number	Intensity[a] at 635 nm	Intensity[a] at 532 nm
1175	1125	1683
2329	819	1621
3407	273	532
5717	1420	1888

[a] Intensity values after substraction of background intensity

(a) Correct the data using the global normalization factor k = 1.2229 obtained in Computational Example 11.1.

(b) What is the probability that the expression levels of *twist* in dsd^D flies differs from expression levels of *twist* in normal flies?

(c) What is the probability that $R/G > 2$?

5. Answer the course questionnaire at `http://ilmo.cs.helsinki.fi/kurssit/servlet/Valinta?kieli=en`.