ECS 124A: Theory and Practice of Bioinformatics

Lab Assignment 4

Instructor: Ilias Tagkopoulos (iliast@ucdavis.edu)
TA: Linh Huynh (huynh@ucdavis.edu)

Key dates:
Assigned: Wednesday 11/20/2014
Due by: Wednesday 12/04/2014

Scope:
The goal of this lab is for you to (a) familiarize with the basic techniques of unsupervised learning and apply these techniques in real datasets, (b) learn about analysis techniques for biological networks, (c) use nucleotide evolution models and phylogenetic trees.

Deliverables:
Hand in answers to all the exercises/questions in the lab instructions.
Grading:
Exercise 1: 30 points
Exercise 2: 50 points
Exercise 3: 20 points (bonus)
TOTAL : 80 points (+20 bonus points)

Exercise 1: Network Inference and GO analysis

For this exercise, you will (a) build a gene regulatory network from gene expression data and (b) perform a functional analysis for that network. On Smartsite, under resources > Labs > Lab4 you will find the csv file GDS2768. Please note that the first two rows in the file are headers. This is a dataset from a microarray experiment of *Escherichia coli* K-12 strain to investigate biofilm formation, for the first 24 hours of culturing:

Domka J, Lee J, Bansal T, Wood TK. Temporal gene-expression in Escherichia coli K-12 biofilms. *Environ Microbiol*. 2007 Feb;9(2):332-46.

http://www.ncbi.nlm.nih.gov/pubmed/17222132

Question 1.1: Use the tool GENIE3, which can be downloaded from

http://homepages.inf.ed.ac.uk/vhuynht/software.html

to infer the gene regulatory network of *Escherichia coli* K-12 from the dataset GDS2768 above. Report the top 100 interactions (i.e. source, target, type of edge) their score and the statistical significance of the score.

Question 1.2: Use the DAVID tool from

http://david.abcc.ncifcrf.gov/

to find all functional categories of genes that relate to 100 interactions above. Make a chart to represent the number of genes for each category.

Exercise 2: Phylogenetic analysis

Question 2.1. Calculate the pair-wise Jukes-Cantor distance between these three sequences that were taken from three different species:

S1: AAAATCGATCAAATCAT
S2: AATCTCGATCAATTCAT
S3: ATATTCGATAAATTAAT

What can tell from these calculations regarding the evolutionary similarity of the corresponding species?

Question 2.2. What are the assumptions that the Jukes-Cantor Model makes? What would be the Jukes-Cantor distance between two sequences, where the first has a frame shift mutation (indel) but otherwise is identical? For example assume the following genomic sequences:

S1: ATTCGAAA …. **TCAAAATGCA** …. ATTGDSAAAA
S2: ATTCGAAA …. **CAAAATGCAT** …. ATTGDSAAAA

Only the sequences in bold are given (second sequence identical to the first but shifted one position to the left). Calculate the JC distance.

Question 2.3. Derive the Jukes-Cantor probability P that a nucleotide will remain unchanged over a period of T time units, which is given by:

$$P = \frac{1}{4}\left(1+ 3e^{-4aT}\right)$$

Assume rate of change α for a nucleotide mutating to any of the 3 other nucleotides (i.e. the mutation probability for any given nucleotide during an infinitesimal period of time is $3\alpha\Delta t$). You can follow the same reasoning that we used in class and then set the derivative of the substitution matrix to zero to get the retention/substitution probabilities.

Question 2.4. Below is a distance matrix between different species.

	HUMAN	MOUSE	RAT	DOG	CAT
HUMAN	0	3	5	10	11
MOUSE		0	2	7	8
RAT			0	5	8
DOG				0	2
CAT					0

Is this an ultra-metric distance matrix? Why? Use UPGMA to create the corresponding tree.

Exercise 3: Unsupervised learning (bonus)

The purpose of this exercise it for you to get a hands-on experience with unsupervised learning. This is an optional exercise (bonus points) that aims to extend your expertise in bioinformatics and hence it will need you to spend some time by yourselves to familiarize with R and bioconductor. You do **not** need to complete this exercise and the TA may not be able to help you troubleshoot.

There are several tool boxes to perform clustering, including Cluster 3.0, which uses a Perl script for clustering (Algorithm::Cluster for Perl) that can be found here:

http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm

or WEKA that can be found here:

http://www.cs.waikato.ac.nz/ml/weka/

For this exercise we *will use the R statistical programming language and Bioconductor*. The package “cluster” that is part of Bioconductor includes several methods for unsupervised clustering. For this exercise, use ***one*** of the methods *agnes, diana and mona* to perform hierarchical clustering and the methods *pam*, *clara and fanny*, to perform partitioning (non-hierarchical clustering). The reference manual for the package is under cluster.pdf and can be accessed from:

http://cran.r-project.org/web/packages/cluster/index.html

In the Smartsite, under resources->Labs>Lab4 you will find the DeLuc_Grapes_ Dataset1 (two versions, one xls and another txt, use whichever you find more easy to handle; the files have only the first 2000 genes of the real dataset). This is a dataset from a microarray experiment over 112 days and it looks at the gene expression profile of grapes while ripening. The first row in the file is the header (geneID/days in time series). The paper that describes the methods and results is also under the same folder (Deluc2007).

Question 3.1. Use any of the partitioning methods to cluster the Deluc_Grapes_Dataset1. Do that for number of clusters k=5,10,20. For each cluster only, in each of these cases (i.e. for k=5,10,20) report (a) the top 3 genes, (b) the cluster center, (c) the number of genes in the cluster. Visualize the results (any heatmap will do).

Question 3.2. Use any of the hierarchical methods and visualize the results (any heatmap you select will do).

Question 3.3. Use the methods we learned in class (BIC or AIC) to calculate the number of clusters that optimize the objective functions (which as we discussed penalize large number of clusters).

END OF LAB 4