Problem Set 4

Question 0: Project Teams & Times (5 points)

List the name(s) of your partner(s) for the final project: _____
_____

Oral presentations for final projects will take place at the following times and locations. Please indicate below if your team can or cannot present during the following timeslots. If your team **cannot** present at a given time, please also list the reason(s).

Lec#	Time	Can Attend	Cannot attend (+ reason)
Lec 12	12-2pm		
Lec 12	5:30-7:30pm		
Lec 13	12-2pm		
Lec 13	5:30-7:30pm		
Lec 14	12-2pm		
Lec 14	5:30-7:30pm		

Problem 1: Clustering (33 points)

Microarray and DNA chip technologies have made it possible to study expression patterns of thousand of genes simultaneously. The amount of data coming out of these efforts is overwhelming. A powerful strategy for analysis of microarray data is the clustering of expression profiles. Expression profiles can be clustered by gene or by condition. Golub *et al.* (*Science*, **286**, 531-7. pdf, supplemental website) clustered different types of leukemia expression data using non-hierarchical Self-organizing Maps (SOMs). Now you will write a Perl program to cluster the same data using an alternative hierarchical clustering algorithm.

I) I) Briefly describe the two major goals of this paper. (2 pts)

II) II) Describe the major steps of the SOMs training algorithm without using code. (4 pts)

III) III) The authors used Affymetrix GeneChip, which is very different from ratio-based cDNA microarray in the way of measuring expression level of RNA. Data from several different GeneChip microarrays should be normalized before being compared to each other. Describe why normalization is needed, and how the authors normalized their data. (4 pts)

IV) IV) A brief summary of the hierarchical clustering algorithm that you are asked to implement can be found here. Your assignment is to cluster the normalized expression data of 50 predictor genes from Golub *et al.* using the single-linkage and complete-linkage Euclidean distance metrics. (11 pts)

a. a. Partial credits are given for the following tasks:
   i. i. Reading input data (2 pts)
   ii. ii. Constructing distance matrix (2 pts)
   iii. iii. Updating distance matrix (3 pts)
   iv. iv. Output clustering result (4 pts)
b. b. Use the sample dataset of 5 samples and its clustering result to verify your code. Print the group members and distance matrix at each iteration.
c. c. Please attach your well-annotated Perl code to the end of your problem set. You may use this template (ps4-1-template.pl) for your program.

Sample dataset of 5 samples:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-sample.txt

Clustering result of sample dataset using complete-linkage Euclidean distance:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-result.txt

Normalized training dataset:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-train.txt

Clustering result of normalized training dataset using complete/single-linkage Euclidean distance:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-result2.txt

V) V) Decide the number of clusters you want to use in your program, and explain how you decided that number according to the original paper. (1 pt)

VI) VI) Run your program on the normalized training dataset using the **complete-linkage** (farthest neighbor) Euclidean distance metric. Provide your clustering result (without distance matrix) and compare it with the original paper (table_ALL_AML_predic.txt). If your program does not work, use the provided clustering result above. (4 pt)

VII) VII) Run your program on the normalized training dataset using the **single-linkage** (nearest neighbor) Euclidean distance metric. Provide your clustering result (without distance matrix) and compare it with the original paper (table_ALL_AML_predic.txt). If your program does not work, use the provided clustering result above. (4 pt)

VIII) VIII) Describe and explain any differences or similarities between your results from VI and VII. If your program does not work, use the provided clustering result above.(3 pt)

Problem 2: Motif searching and functional enrichment (34 pts total)

You will need to read the following paper by Tavazoie et al. to answer the next part: *Nature Genetics **22:**281-5*

If two genes change expression level in the same way in response to a change in conditions, they are often assumed to be related (e.g. co-regulated, or play common roles in cellular processes).

I) I) With reference to table 1 and the "Determination of statistical significance for functional category enrichment" in "Methods" section in Tavazoie *et al.* paper, answer the following questions. (Total 9 points)

a. a. With reference to table 1 and the methods section in Tavazoie *et al.* paper, explain how the statistical significance for functional category enrichment is determined. (3pts)

b. b. Examine table 1 and figure 1. What are the clusters that show cell cycle periodicity and at which stage of cell cycle genes in each of these clusters may be required? How would you use such information to learn more about a gene with previously unknown function in one of these clusters? (3 pts)

c. c. The total number of genes within a genome is not available for many species, for example, the human genome. Looking at the calculation of the hypergeometric distribution in the methods section, make an argument in favor of using a different number for the variable represented as *g* than the total number of genes within the genome. What number would you use and why? (Hint: think about the clustering step). (3 pts)

*You might find Hughes et al., J. Mol. Biol. **296**: 1205-1214, helpful when answering the following questions*.

II) II) AlignACE uses a Gibbs sampling algorithm to identify over-represented motifs in a set of DNA sequences. The program can be accessed at the following site: http://atlas.med.harvard.edu/cgi-bin/fullanalysis.pl. Here you will use it to analyze the upstream regions of genes present in the Tavazoie *et al.*'s cluster #30 http://arep.med.harvard.edu/network_discovery/clusters_members_distances_annotations.txt (total 12 pts)

Here is the list of gene names in cluster #30 (all you need to do is copy/paste into the "Enter a list of genes below, one gene name per line (Y names only):" filed:

YAL053W
YAL067C
YAR015W
YAR052c
YBL015w
YBR085w
YBR112c
YBR155w
YBR156c
YBR213w
YBR289w
YDL059C
YDL071c
YDR213W
YDR227w
YDR252w
YDR253C
YEL007w
YEL043w
YER042w
YER132c
YFR030W
YGL013C
YGL184C
YGR058W
YGR138C
YGR239C
YHR210C
YIR017C
YJL106W
YJR010W
YJR047C
YJR127C
YJR137C
YKL001C
YKR069W
YLL061w
YLL062c
YLR048w
YLR092w
YLR228C
YLR327C
YLR364W
YMR006C
YMR190C
YMR306C-A
YNL033W
YNL191W
YNL241C
YNL277W
YOL163W
YOR267C
YOR368W
YPL002C
YPL054W

YPL116W
YPL140C
YPL188W
YPR046W
YPR104C

a. a. Report the five best motifs obtained by AlignACE in terms of MAP score (note: AlignACE lists its outcome in the order of decreasing MAP score). (6 pts)

b. b. AlignACE lists its outcome in the order of decreasing MAP score. From the result you obtained by running AlignACE, does the highest MAP score always correlate to most meaningful functional motif? Why or why not? What information do group specificity scores add when trying to infer the "real" cis-regulatory elements? (3pts)

c. c. The article by Hughes *et al.* (see above) suggests significance thresholds for the MAP and group specificity scores: >=10 and <=10^{-10}, respectively. Based on these thresholds and the statistics assigned to your motifs (part c), what can you infer about the regulation of genes in the cluster? (3 pts)

III) III) The relevance of motif results (8 pts total)

a. a. Why are motif analyses performed using clusters from gene expression data (such as microarray) instead of whole genome? (2pts)

b. b. Do all the genes in same cluster share same motif? Why or why not? (3pts)

c. c. Do all the genes sharing same motif clustered together? Why or why not? (3pts)

IV) IV) Sequence Logos as visual representations of motifs (5 pts total)
You might find the following URL helpful in answering these questions:
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

a. a. Motifs can be represented visually as sequence logos with nucleotide letters of various sizes at each site. How is each letter's size calculated? What can you conclude if there is only a tiny "A" at a given position in a motif? (3 pts)

b. b. Can you reconstruct the actual regulatory sequences that went into building the motif? What information is lost in representing motifs by sequence logos? (1 pts)

c. c. How are sequence logos an improvement over consensus sequences? (1 pt)

Problem 3: Markov Chains and Hidden Markov Models (33 pts total)
Mount pages 185-191 and Durbin chapters 3-6 will be helpful.

I) I) Describe an example of a simple Markov chain. It does not have to be a biological example. (2 pts)

II) II) What is a Hidden Markov Model? (2 pts)

III) III) Suppose you were generating a HMM to predict whether a given sequence most likely came from a CpG island, a non-CpG island, or partially from both.(12 pts total)

What two states will your model consider? (2 pts)What probabilities should your model include? Use variables to represent the probabilities, not actual numbers (i.e. P(A+ ->C-) would represent the probability of generating a C in a non-island following an A in a CpG island). (10 pts)

IV) IV) Consider the sequence, "CCGTGC." Based on your HMM described above, how would you compute the probability that the first 3 nucleotides of this sequence came from a CpG island and the remaining 3 nucleotides came from a non-island? Since we have not assigned numbers to the probabilities above, write your answer using variables (i.e. probability = P(A) x P(C) x P(G)). (3 pts)

Besides Durbin chapters 5-6, PFAM related web sites such as http://pfam.wustl.edu/index.html will also help for the next few questions.

One of the most common uses of hidden Markov models for molecular biology is in protein family classification. Suppose we want to find out what the function of a protein might be. Before heading towards the bench, we would like to get as much information as possible from existing information about known proteins. We could do a BLAST search against a protein database, which will give us pairwise alignments of our unknown sequences with every similar protein in the database. However this is not always satisfactory because sometimes our unknown protein is not very similar to any individual protein in the database. An alternative approach would be to gather information from all (or most) sequences in a protein family and compare our unknown protein with such information to examine the likelihood of our unknown sequence being related to that protein family.

V) V) Without getting into details of the probabilistic model, briefly explain how hidden Markov models can be used to classify a new protein into a known family

and/or to search a database for new proteins that may belong to a known family. (5pts)

VI) VI) Now let's turn to a practical example. Go to http://www.ncbi.nlm.nih.gov/ (or any of your favorite protein sequence database web sites) and retrieve the protein sequence with accession BAA76778. (2 pts total)

a. a. What functional information (if any) or definition do you get from the annotation in the database entry? (1 pt)

b. b. Now do a BLAST search against the non-redundant protein database. What would you conclude from the BLAST search results? You do not need to show the blast output. (1pt)

VII) VII) The profile hidden Markov models (aka Pfam) for most (if not all) protein families are readily available and can be searched with protein sequence queries. You can conveniently perform such search on web sites such as http://pfam.wustl.edu/index.html.

a. a. Do a protein search using the sequence you retrieved in problem VI (BAA76778) as the query. You will need to use the fasta format sequence for the search. Do you believe the Pfam HMM search results? Why? (Hint: Examine the scores and E-values.) (2 pts)

b. b. What would you conclude from the Pfam HMM search results? (Hint: follow the link to the Pfam entry.) (3 pts)

VIII) VIII) Can you think of an advantage and a disadvantage of protein classification using Pfam compared to BLAST search? (2 pts)