HST.480/6.092: BIOINFORMATICS AND PROTEOMICS

Lab/Assignment #1

Labs/homeworks are due on the following Thurs (after being handed out). at midnight. Please submit homework in zipped format (along with source code and relevant path/other information- so it can be executed if needed) to gilusa at this email provider: gmail.com This is also a good email for various non-urgent emails as well as large files. *Please do NOT send large files to any mit email addresses.*

If the background is unfamiliar, please read the hints in the lab solutions provided under the Solutions column for Lab 1. (All students should do all problems- except optional extension problems. But, if you are in Course 6, please check hints for HST-labeled questions and vice versa for HST students).

For all Matlab problems, turn in any relevant code and figures. Program output, including figures, can automatically be saved via "'File' -> 'Publish To' menu item from the M-file editor (type 'edit' in Matlab command prompt) or via the publish (type 'doc publish' in matlab command prompt for details).

Warm-up

1. Klenk H.-P. and colleagues published the complete genome sequence of the organism Archaeoglobus fulgidus in *Nature*. [Nature 390:364-370(1997)].

   a. Find the protein sequence of the hypothetical protein AF1226 precursor. What is the sequence of amino acids (in single letter representation) from positions 141-147?

   b. Write a Matlab function that calculates how many nucleotide sequences can give rise to an arbitrary amino acid sequence. The input to the function should be a string (amino acid sequence) and the output should be the number of potential nucleotide sequences.

   c. Using the Matlab function from part b., calculate the number of sequences that could give rise to the 7 amino acid sequence found in part a.

2. Ovarian cancer is the fifth leading cause of cancer deaths in American women. According to the American Cancer Society, there are over 25,000 new cases of ovarian cancer diagnosed each year. 16,090 women died from Ovarian cancer in 2004. One of the problems is that this type of cancer often spreads (metastasizes)

before it is detected. Promising new approaches include examining serum for protein biomarkers (e.g. CA125) and gene testing (e.g. BRCA1). Here, we look at a microarray experiment to see which genes may be involved in epithelial ovarian cancer (the most common form). The dataset is under the Supporting Files column for Lab 1.

a. Load the dataset into Matlab. For more information loading Affymetrix data into Matlab, check out this demo: http://www.mathworks.com/products/demos/bioinfo/affydemo/affydemo.html

   Then, cluster the genes using average distance metric in hierarchical clustering. How well does the data cluster?- explain. Draw a dendrogram and heatmap to show which genes are differentially expressed. Extension: Do fuzzy c-means and compare with k-means clustering. How are the two methods different- both in operation and in practical results (see part d for quantifying the practical results)?

b. Devise a metric to determine what threshold for differential expression should be used to select genes that are significantly (probabilistically speaking) up or down in terms of expression. Use this to find the differentially expressed genes. Conduct a literature search to see what the role of these genes are in ovarian cancer.

c. Does significance in terms of up/down expression necessarily imply regulatory control and/or involvement in ovarian cancer? Is it sufficient? When are these true? If not always true, give an example where it does not (it doesn't have to be bioinformatics related).

d. Divide the data into training/testing set (First 80% training/last 20% testing) and test performance of classification of ovarian cancer versus normal using: k-means and k nearest neighbor. Is performance better when you use **all** of the genes, or if you select only those that were differentially expressed based on part b? Extension: How can you select the best k parameters for these methods? Implement this in Matlab and report the best parameters.

e. Write a Matlab script to compute the confusion matrix for the test set. A confusion matrix has the following information:

	Normal (Model Prediction)	Cancer (Model Prediction)
Normal (Reality)	a	b
Cancer (Reality)	c	d

In the above table, a, b, c, d are the number of test cases within each category. Using this, calculate the accuracy, sensitivity, and specificity for your predictors.

f. Assuming such a test was used for screening, would you be interested in higher accuracy, sensitivity, or specificity? What about for confirming a suspected cancer diagnosis?
g. Extension: Performance results are better understood using an ROC (Receiver Operating Characteristic) curve. What is the best operating point on the curve assuming false alarms and detection are equally important?

HST-type question

1. A 9 year old patient comes in with:
   Café au lait macules (spots) of 7mm in diameter at maximum

Image removed due to copyright considerations. To see image, please visit:

http://www.medinfans.com/DERM-caf%E9-au%20lait-cache.html

She has crowe sign (freckles) in the axilla (armpit)

Image removed due to copyright considerations. To see image, please visit:

http://www.vh.org/pediatric/provider/radiology/PedRadSecTF/080495/Images/CROWE1.html

Upon closer examination, two Lisch nodules (iris hamartomas) are seen via slitlamp:

Image removed due to copyright considerations. To see image, please visit:

http://www.mrcophth.com/finalmrcophthmcqss/22neuroophthalmology/22neuroophthalmology.html

a. What is the diagnosis?

b. What physical finding not listed above do expect to see (as the patient gets older)?

c. What gene do you suspect is involved?

d. The Affymetrix microarray probe (from human genome array: HG-U133+ 2.0) for the gene in part c. is as follows:

   AAGTGCCATGTTCCTCAGATTTATC

   If one starts with a sequence of length ***n***, derive a formula for the probability that it will match is random 25 base sequence exactly *assuming* 1) base types are uniformly distributed within each position and 2) each base position is independent of the other.

e. Using the same assumptions as above, what is the expected number of sequences that will match within the entire human genome ($3x10^9$ base pairs) going in the 5' to 3' direction (i.e. only looking in one direction).

f. Using the same assumptions as above, what is the theoretical ratio of the number of serine to tryptophan amino acids in the human genome?

g. Are the numbered assumptions made in question parts d., e., and f. always correct? Explain why or why not.

h. Diagnosis of the disease in this question is typically done via clinical evaluation rather than by microarray analysis. Describe a use for having this gene as a probe in the HG-U133+ 2.0 Genome Array. How might it be useful in other settings, in conjunction with other technologies, or in medicine in the future?