HST.480/6.092: BIOINFORMATICS AND PROTEOMICS

AN ENGINEERING-BASED PROBLEM SOLVING APPROACH

Lab/Assignment #2/3

Labs/homework is due Jan 28, 2005, 5pm. 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 2. (All students should do all 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. You have a sample of human DNA. However, in order to analyze it, you need to 'amplify the signal' by creating multiple copies of the sequence. This can be done via a process called PCR (Polymerase Chain Reaction). In one step of the process, a DNA polymerase is used. Such polymerases typically come from organisms like: *Thermus aquaticus, Pyrococcus furiosus,* and *Thermococcus litoralis.* What would happen if you used human DNA polymerase? Why? (What do the aforementioned organisms have in common?)

2. The goal in this exercise is to look at proteins from a different perspective to gain insights on mass spec analysis and related issues. In this problem, we will look at proteins in terms of masses and see if any patterns emerge upon analysis. We will also look at cleavage of proteins and how this works to produce mature and signaling peptides.

A file (protein.gbk) contains all non-redundant protein sequences for Homo sapiens in Genbank flatfile format (see resources for this problem set). The file is from the NCBI RefSeq database: http://www.ncbi.nlm.nih.gov/RefSeq/. The file is a rather large- and you might want to break it up in order to read it in (it is just a text file). Matlab genbankread can be used to read in Genbank flatfile format.

a. After reading in the data, you can use the following scripts (see resources for this problem set) to calculate the features of the proteins (i.e. where it is cleaved and where the mature peptide is)

   parseFeatures.m
   getregweights2.m
   Please see the comments at the beginning of each file for usage/example.

   As you can will see, each protein entry can yield multiple proteins/sequences depending on various factors: cleavage, post-translational modification, etc. Calculate the masses for all resulting proteins from the file. Then, plot the number of proteins per mass unit (protein 'density') versus mass.

b. Does the data appear to be randomly distributed? If not, where do you see particular deviations from this. Focus on the 2000-12000 Da region (where SELDI mass specs typically operate). Do you notice any such deviations here?

c. Plot the empirical probability density function, or PDF, for the data (in terms of probability vs. protein counts/mass unit). Given that the distribution is describing counts, what distribution do you think the PDF might follow?

d. Model the PDF with the following distributions (all available in Matlab Statistics Toolbox): gamma (a superset of exponential distribution), negative binomial, and Poisson. Which model is the best fit (e.g. in terms of log likelihood or other metric)? Why do you think this model turned out to be best (compare with your predicted model in part c)?

3.

a. Use Boolean algebra and DeMorgan's Law (as necessary) to find a simple expression (using only non-nested primitives AND, NOT, OR) that is consistent with the following protein expression circuit perturbation-based profile. In each profile (except the one at =0), one of the protein is either removed (↓) or added (↑).

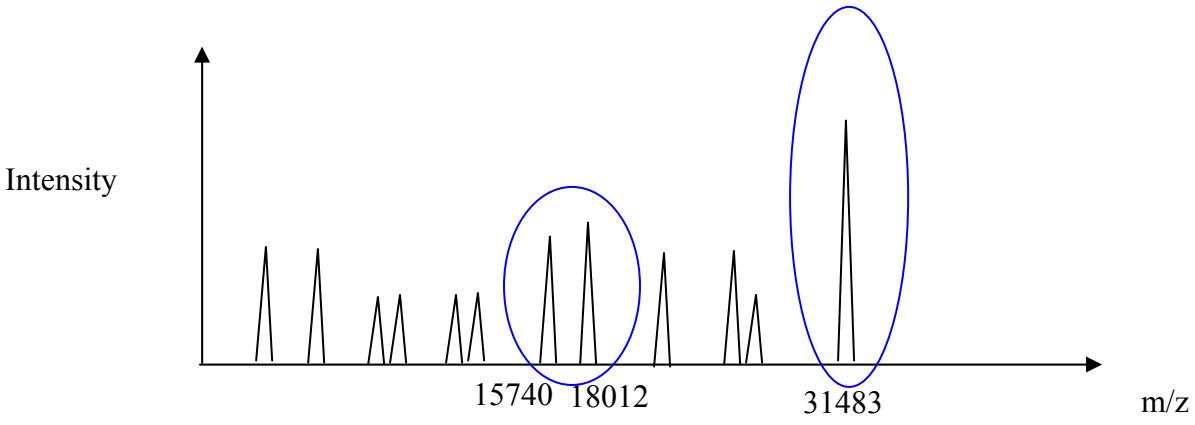
Time	**Protein 1**	**Protein 2**	**Protein 3**	**Protein 4**
t=0	**0**	**1**	**1**	**1**
t=1	**1**	**0**	**1**	**↓**
t=2	**0**	**0**	**↓**	**1**
t=3	**1**	**↓**	**1**	**1**
t=4	**↑**	**1**	**1**	**1**

b. Your collaborators' experiments suggest that all proteins except protein #3 are involved in the p53 pathway (involved in many cancers). Based on Biocarta (www.biocarta.com) pathways, what are possibilities for proteins 1, 2, and 4 given the protein circuit you derived in part a.

4. Cancer is often caused by a breakdown of a series of steps in regulation of cell growth and proliferation. One example is colon cancer, which often involves a certain temporal sequence of mutations in the following genes: APC, MCC, ras, p53, and DCC. After doing a literature search on colorectal carcinomas and citing appropriate references, construct a FSM (Finite State Machine) outlining the series of steps involved from normal colonic epithelium to adenoma to carcinoma. Extension: Label (where possible) arcs with approximate progression time between stages.

5. The Human Massome (http://chocolate.chip.org/~protcoop/protind.php) is a research project that seeks to view proteins and their interactions from a mass spec perspective. It contains the largest collection of human protein interactions (>100K protein interactions). It can be searched by inputting the masses of suspected protein interactions in the form. The first line is for the minimum and maximum expected mass of the first protein (or protein product) respectively. The second line is for the minimum and maximum expected mass of the second protein (or protein product) respectively.

a. You do a SELDI-based mass spec experiment to analyze certain binding proteins in cancer. After some analysis, you see that the following peaks (circled) that are of interest. The mass spec laser/instrument has been optimized for the range of 20000-35000 in this experiment. Assume the instrument accuracy is 400 ppm (parts per million). What are the possible protein masses for the three circled peaks? List a few potential identities for these three.



b. Your collaborator verifies that the proteins represented by the peaks, do, in fact, bind. However, there are only two proteins involved (though the collaborator is unable to identify them in time for your publication deadline). Using the "Human Massome," what might the identify of these proteins be? Why could the

collaborator only find complexes with two different types of proteins- is he/she missing something?

c. The collaborator feels bad after having missed the deadline for the publication. So, he offers to help identify some other interactions. Via gel electrophoresis techniques, he finds that a protein with a mass of around 81,372 Da interacts with the 18,012 Da protein above. Why is SELDI mass spec not really suitable to find the new protein (why is it not suitable- in this experiment as well as in general?)?
d. The collaborator is on a roll- and has found out that another protein with a mass of around 55,344 interacts with the new protein found in the last part. What could this protein be? Looking at the Biocarta (www.biocarta.com) pathway for CD95, what new pathway (not listed) does this new finding suggest? Are interactions from previous parts of this question in the Biocarta pathway diagram?
e. How might a database like the "Human Massome" be useful for researchers? Name as many uses as you can (in addition to how it is used in the above problem).