Part A: Protein structure (35 pts)

The protein Bcl-2 is the product of a human proto-oncogene located on chromosome 18; it is known to causes B-cell leukemia. When reciprocal translocation occurs with this gene locus and a region of chromosome 14 that has an upstream enhancer, Bcl-2 is made in great quantity and begins to inhibit apoptosis (cell death). Let's use this protein as an exercise in protein structure classification and prediction.

I. Protein structure classification and databases (10pts)

Four major protein structure classification databases are listed in chapter nine of the Mount *Bioinformatics* text. You will find pages 402-403, and 421-423 especially helpful with these questions.

a) Briefly give one reason why MMDB would yield different structural neighbors from the other classification methods, i.e. CATH, SCOP, and FSSP. (2pts)

The first step in studying the structure of a protein is to search for it in a structural classification database. Start at (http://www.ncbi.nlm.nih.gov/Structure/), NCBI's MMDB database. Search for structures of Bcl-2 in the MMDB (enter Bcl-2 in the search box).

b) List two structural neighbors of Bcl-2 in the MMDB classification system. (2 pts)

Select the structure entry for "Human Bcl-2, Isoform 2." Follow the PBD ID link (1GJH) to the PDB entry for Bcl-2, isoform 2. Follow the link on the right side indicating "structural neighbors" and look at the CATH classification of Bcl-2.

c) What are the classifications of this protein provided by the CATH database? (3 pts)

The FSSP database is based on a structural alignment of all pair-wise combinations of the proteins in the Brookhaven structural database by the structural alignment program DALI (**d**istance **ali**gnment tool). DALI compares existing structures using the distance matrix method and provides one convenient way to compare a new structure to existing structures in the Brookhaven structural database.

d) Give a brief explanation (2-4 sentences) of the methodology of the DALI algorithm (3 pts)

II. Prediction of protein secondary structure (8pts)

All secondary structure prediction algorithms are based on the assumption that a given stretch of sequence is more likely to form one kind of secondary structure than another. The Ramachandran plot predicts whether a given amino acid will form an alpha helix or a beta sheet.

a) What characteristics does the Ramachandran plot use to predict whether an amino acid will tend to form beta sheets or alpha helices? (1 pt)

Most widely used methods of protein secondary structure prediction use one of three major approaches (or a combination): Chou-Fasman/GOR, neural network models, and nearest-neighbor methods. Jpred (http://jpred.ebi.ac.uk) is a secondary structure prediction tool that runs a number of secondary prediction methods (NNSSP, DSC, predator and other nearest neighbor methods) and generates a consensus of the different outputs. Run Jpred on the Bcl-2, isoform 2 protein with the following settings:

- run all prediction methods in Jpred (deselect box 1)
- bypass the scan for structures in PDB (select box 3).
- Use the following sequence (this is Bcl-2, isoform 2):

mahagrtgyd nreivmkyih yklsqrgyew dagddveenr teapegtese vvhltlrqag ddfsrryrrd faemssqlhl tpftargrfa tvveelfrdg vnwgrivaff efggvmcves vnremsplvd nialwmteyl nrhlhtwiqd nggwdafvel ygpsmr

b) Describe the predicted topology and secondary structure of the protein (2-3 sentences). (2 pts)

c) How do the results from the different methods used by Jpred differ from each other and the Jpred consensus prediction? (2 pts)

d) Does the prediction agree with the classification found in the CATH database? (3pts)

III. Prediction of protein tertiary structure (17pts)

Now we will try to determine the 3D structure of Bcl-2 by homology modeling (HM). Homology modeling (or comparative structure modeling) allows one to predict protein structure using the following three steps:

1) *Identifying template proteins of known 3D structure related by primary sequence to the target protein*
2) *Constructing the model of the target sequence given its alignment with the template structure*
3) *Energy minimization and evaluation of the model structure*

In this problem, you will use a program called Swiss Pdb-Viewer which can be used both to view protein structures and for protein structural comparisons and homology modeling. The initial search for homologous proteins and energy minimization (steps 1 and 3 above) require larger and faster computers than you are likely to have on your desk, so Swiss Pdb-Viewer submits these tasks to a remote computer to carry out the heavy computing tasks. After searching for homologous proteins in the remote database (step 1), the pdb structure files of the homologous proteins come directly back to your web browser which then hands them to the Pdb-Viewer. With Pdb veiwer, construction of an initial structural model (step2) is achieved by "threading" the target amino acid sequence through the candidate fold. In threading, the quality of the sequence-structure fit is evaluated using inter-residue potentials of mean force. Although much of the threading can be automated, manual input is typically required for the modeling of regions of weakest homology. Last, the structural model generated is sent to the remote computer for final optimization which is then emailed to you.

This problem will walk you through the process of using homology modeling to predict the 3D structure of Bcl-2. First, you'll download Pdb-Viewer if your computer doesn't already have it installed (this is quite painless). Then, using Pdb-Viewer, you'll find proteins homologous to Bcl-2 with the corresponding structures, thread Bcl-2 onto them, improve the fit of several loops and fix clashes, save a copy, and send the model off for optimization. When you receive the minimized model, you will compare it with the saved copy to see the changes made during minimization.

Go to (www.expasy.ch/spdbv/text/download.htm) and download Swiss Pdb-Viewer v3.7 beta 2. Save the Bcl-2 sequence as a text file if you haven't already. Launch Pdb-Viewer and load the Bcl-2 sequence (SwissModel>Load Raw Sequence to Model). If you don't see any detail, such as side chains, you may want to zoom in on the structure, using the menu tool indicated by two overlapping boxes.

a) Briefly describe the protein secondary structure displayed (1-2 sentences). Does this differ from the secondary structural prediction you ran earlier? (3pts)

Obtain the template structures:

After entering your contact information (Preferences>Swiss-Model), obtain structures of homologous proteins (SwissModel>Find Appropriate ExPDB templates). If you have problems getting this to work, try quitting your browser before going to SwissModel>Find Appropriate ExPDB templates or go to www.expasy.ch/swissmod/SM_Blast.html and submit the Bcl-2 amino acid sequence (with fasta header) manually. Download two (or more, if you'd like) top-scoring template structures and open them in the Viewer (File>Open Pdb). You may want to color each of the structures a different color (Color>Layer) to make distinguishing them easier. Note that the control panel displays information individually about each of the structures or sequences loaded; you can switch between information on the different structures by clicking on the structure name at the top of the control panel or by using the layers window.

b) Which template(s) did you use and why?(1pt)

Fit the **template** structures (not Bcl-2!) to each other using the iterative method by selecting one of the two templates under the control panel or layers menu and then Fit>Iterative Magic Fit. You will have the option to fit by CA, backbone, sidechain, or all atoms. Look at the superposition of templates once you're done. Try zooming in and out as well as rotating the structure.

c) Describe briefly the resulting superimposed fit (1-2 sentences). Find the root mean square value (rms) associated with it (fit > calculate RMS if it's not already displayed on the top bar). What does the RMS value of this superposition indicate? (3pts)

Open the sequence alignment window (window > alignment). You should see that the two template sequences are aligned.

Structural alignment and fitting of fasl to the templates:

Now go to Fit>Generate structural alignment. Then look at the alignment window (Window>Alignment if not already open). Now all three sequences, fasl and the two templates, will be aligned. Thread the mFASL sequence onto the template structural alignment; to do this, you click on the FASL name in the layers window to make this layer active and then use Tools>Magic Fit again. Scroll through the alignment window; you can also look at the whole alignment as a text file by clicking on the page icon on the side of the alignment window.

d) Describe the sequence and structure alignment (2-3 sentences). (2pts)

Manual optimization: Make sure the current layer is FASL and then click on the little arrow located at the right of the question mark in the Align Window. The window expands, displaying the mean force potential energy as a function of residue number. This curve essentially depicts how energetically favorable the alignment is; lower energy values, and energy values below zero in general, are more favorable. The mean force potential energy is computed according to Sippl JM (1990) *J. Mol. Biol.* **213**:859.

Click on "smooth text" and set a smoothing factor of 1. It means that the energy assigned to each residue will be the average of the residue's energy plus the energy of 1 flanking residue on each side (i.e., there is a three residue window for energy state determination.) You can color the alignment with colors corresponding to energy state by selecting Swiss-Model>Auto Color by Threading Energy and then clicking on the "E= -11.9" text.

To manually carry out a partial energy minimization, select a block of residues bounded on one or both sides by a gap(s) and slide this block left or right with the arrow keys. The mean energy of this block will change as the corresponding alignment changes. If E decreases to a point where it no longer changes, you've minimized the local potential energy.

e) What is the mean force potential energy before and after your changes? Go back and look at the overall structural alignment and describe any notable changes (2-3 sentences). (3pts)

Final optimization: You're almost done! Submit your alignment to Swiss-Model (Swiss-model>Submit Modeling request). You will be prompted to save an html file that will then be opened. By default, this will also make a Swiss-Pdb Viewer project file (.pdb), with your model aligned onto the templates you used, ready for comparison. The project will be sent to you at the e-mail address you provided above.

<u>Evaluate the homology model of Bcl-2</u>:
First, look at the energy-minimized pdb file returned in Pdb-Viewer.

f) Briefly (1-2 sentences) describe its structural features. (3pts)

Look at the "quality-control" report that was mailed back to you with the energy-minimized structure.

g) How confident are you in the goodness-of-fit of your homology model and why? For example, how well does your homology model conform to the stereochemical measures reported? (2pts)

Part B: Mass Spectrometry and Proteomics (30pts)

The following reference will be useful in answering the questions in part B:
Gygi et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* 17: 994 (1999).

a) What are two cases when the proteome more informative than the transcriptome? Name an instance when the opposite is true. (3pts)

b) In tandem mass spectrometry, how does the second chamber differ from the first? What additional capabilities does this give, when combined with the first peptide-selection chamber? (2 pts)

c) Describe the ICAT strategy. What advantage does this have over traditional tandem mass spectrometry? (5 pts)

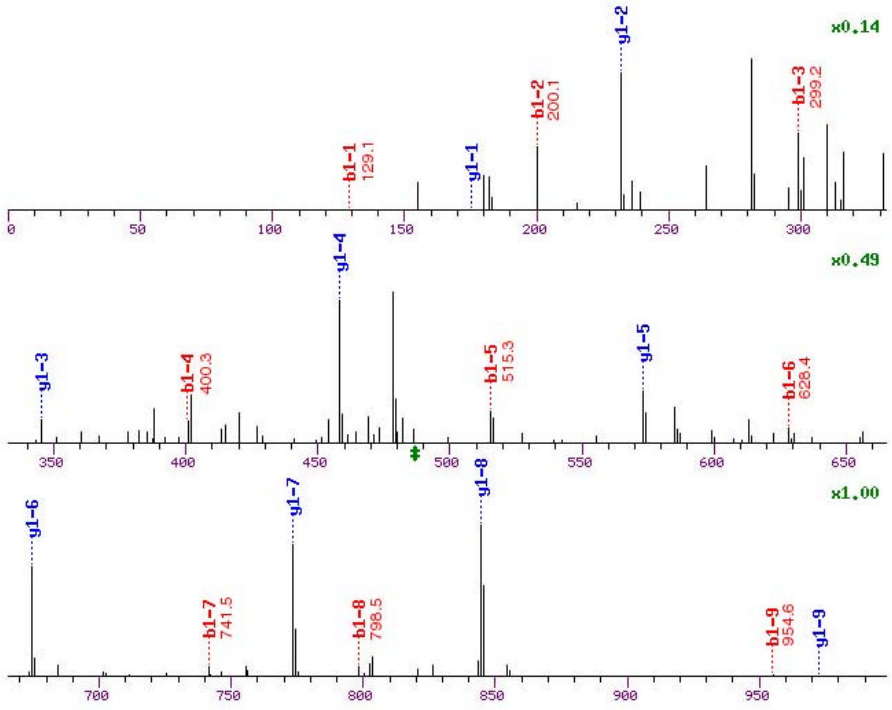
d) How can you use mass spectrometry to identify post-translational modifications? (2 pts)

e) What are 'b' and 'y' ions? How do they differ? (3 pts)

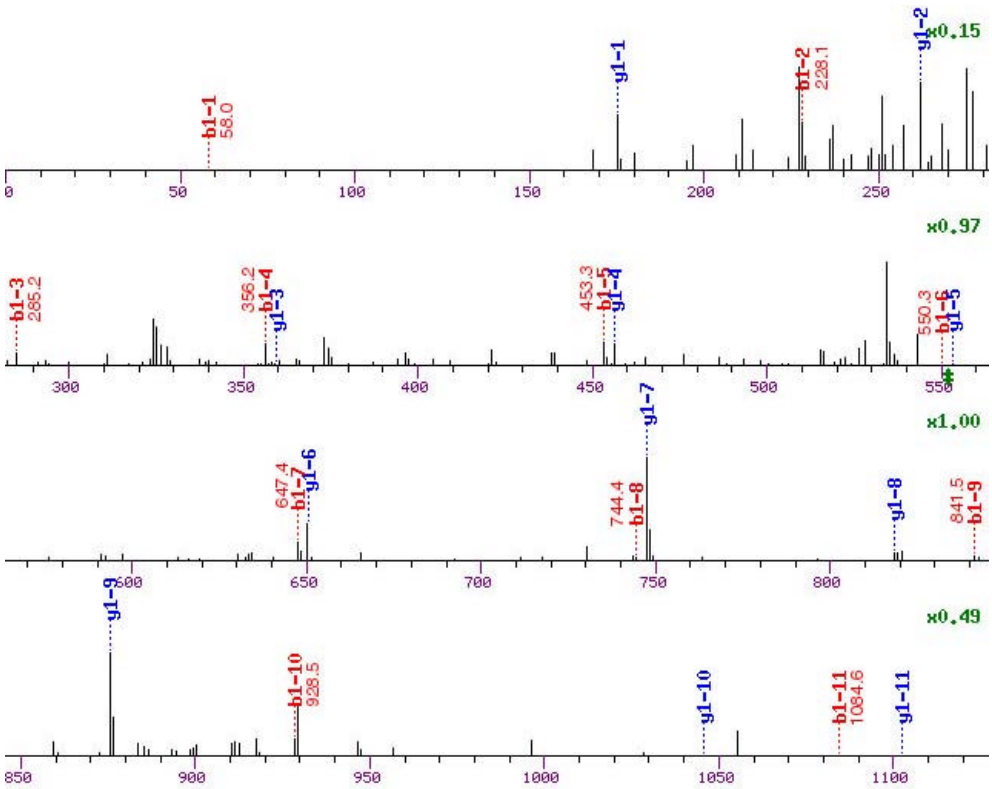
You have analyzed two unknown peptides by MS/MS. Their spectra are given below.

f) Use the 'b' ions to determine the amino-acid sequence of each peptide (hint: one of the peptides has a methylated amino acid). (6 pts per spectrum) Are you able to find distinct amino acids for each peak? Why or why not? (3 pts)

Spectrum #1:



Spectrum #2:



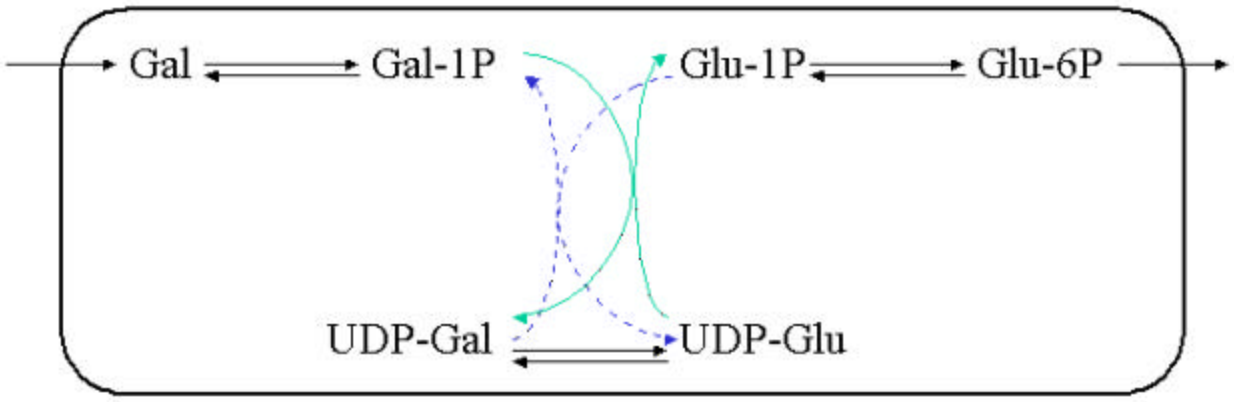
We thank G. Adelmant for sharing his MS/MS data.

Part C : Flux Balance analysis (35 points)

The following papers will be useful for answering questions in part C:

Schilling, CH and Palsson, BO, The underlying pathway structure of biochemical reaction networks. PNAS 95: 4193-4198 (1998).
(http://www.courses.fas.harvard.edu/~bphys101/problemsets/Underlying.pdf)
Schilling, CH et al. Towards metabolic phenomics: Analysis of genomic data using flux balances. Biotechnol Prog 15: 288-295 (1999).
(http://www.courses.fas.harvard.edu/~bphys101/problemsets/Toward.pdf)
Schilling, CH et al. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. Biotechnol Prog 15: 296-303.
(http://www.courses.fas.harvard.edu/~bphys101/problemsets/Path.Analy.pdf)

Note: Mathematica provides functions that can perform all the necessary linear algebra calculations that are required for solving the problems in this section. Question 3 uses a function only available in Mathematica 4.2. If you do not have Matheatica 4.2, you will either need to install an upgraded version (recall that this takes several days) or use an alternative website; see question 3 for details of where to obtain version 4.2 and/or the website's URL.



Note that the reaction depicted with blue and green arrows can also be represented as follows:

$$\text{Gal-1P} + \text{UDP-Glu} \rightleftharpoons \text{Glu-1P} + \text{UDP-Gal}$$

Figure 1. Galactose (Gal) metabolism in yeast

1) The Galactose metabolism pathway in yeast can be represented as shown above.

a) Label each of the internal fluxes by v_i and each of the external fluxes by b_i. (4 pts)

b) Write out the flux balance equations that specify the steady state condition of the pathway.(4 pts)

c) Derive the stoichiometric matrix for the above set of equations and write the equation in the following matrix form:

$$\mathbf{S}.\mathbf{V}=0$$

where **S** is the stochiometric matrix and **V** is the vector representing the fluxes. (4 pts)

d) Solve the linear equations above to derive the set of basis vectors that can be used to span the entire pathway. (Hint: You have to consider the null space of the matrix S. Use the function **NullSpace[]** in Mathematica) (2 pts)

e) Draw the reaction that corresponds to each solution found in part d.(2 pts)

f) What do the solutions from part d represent? Do they characterize the entire pathway as shown above?(2 pts)

2) Assume that UDP-Glu and UDP-Gal are freely available in the system.

a) Modify the pathway in Figure 1 to accommodate this (Hint: Add necessary exchange fluxes).(2 pts)

b) Repeat steps b-d in section 1 on the new pathway.(5 pts)

c) Compare the solutions found to the null space in 2b to those obtained in the last section.(2 pts)

d) Do you get biochemically possible solutions? If not, generate such solutions (Hint: Follow Schilling and Edwards 1998). (3 pts)

3) For the pathway derived in section II, using Linear Programming find a solution that maximizes the UDP-Glu flux out of the system. Interpret the result that you obtain in terms of the biochemical reactions involved. Assume the following flux limits.

- All internal fluxes are (0,Infinity)
- The Gal external flux into the system is (0,100)
- The Glu 6P flux out of the system is (0,Infinity)
- All other exchange fluxes are in the range (0,10)

See http://www.wolfram.com/products/mathematica/newin42/kernel.html for more information on using Mathematica 4.2 for Linear Programming. Note that if you do not have version 4.2 of Mathematica installed, you can either download it from http://genome.dfci.harvard.edu/~zucker (recall that the installation process takes at least 24 hours) or you can use a web applet located at http://algos.inesc.pt/lp/ . (5 pts)

Partial credit will be given for the following:

Correct identification and statement of the cost function (1 pt)
Correct statement of the relevant constraints (1 pt)
Code for optimization of this problem (1 pt)

Solution in vector form (1 pt)
Interpretation of the solution in terms of the corresponding reactions (1 pt)