Constructing a Homology-based Model of a Protein Structure

CGS701 Foundations of Biomedical Science

Protein Unit

Flextime September 3, 2004

Dr. Starr Hazard

email hazards@musc.edu

phone 2-0715

Room 443 Library Administration Building

Note: This is fairly complicated exercise. The aim is to introduce you to the concept of data mining and to give you experience of using one method to produce a homology based model of a protein. If it doesn't all work to plan do not worry ! It is the experience that matters not necessarily the end result. Try to answer as many of the questions as possible during this Flextime session.

Summary of this Flextime Session

I. Setup Netscape and Copy/Paste the query sequences into a BLAST search Form and launch the search.

II. Examine the CDD results and determine the Identity of the parent protein of your peptide.

   - Follow the CDD results and examine the peptide in an alignment with known proteins possessing the conserved domain.
   - Examine the peptide in context of a 3D structure with Cn3D

III. Examine the BLAST results and follow LocusLink to get at the relational databases regarding your protein.

IV. Use Swiss PDBViewer to independently locate a template for the model, and set up a Homology Model.

V. Submit the alignment to Swiss PDBViewer Server for model refinement.

VI. Repeat the flextime exercise on another peptide. Submit via email a summary assessment. This is due
**September 10 at 6pm.**

**Background:**

We will begin with a peptide fragment’s sequence. We will assume that there are no errors in the peptide sequence. In any real world situation, there might indeed be errors arising from various analytical methods that produce the sequence. Next we will identify the MOST LIKELY parent sequence. This will be a statistical inference based on the BLAST database search algorithm. There will be some possibility that our identification is incorrect but it will be a small chance. With the parent sequence identified we will then have access to (almost) all known information about the protein. We will then look for a suitable template molecule from which we will construct a three dimensional model. Location of a suitable template will also be based on a BLAST search and our choice will again be based on a statistical inference. The reliability of this inference has limits. For instance if no currently available 3-dimensional structure has a sequence that is more than 20-25% identical with our sequence to be modeled, then our similarity based modeling procedure will not work.

At present there are no truly reliable modeling methods which can predict protein folds for sequences that have no similarity in sequence. Check the [Critical Assessment of Structure Prediction (CASP)](http://predictioncenter.org/) site for progress in this field. However, where there is a level of similarity above the 20-25% identity level, there are a host of methods which CAN produce a suitable model based on a template. Here is a list of such homology modeling programs available at MUSC.

Composer, Matchmaker and GeneFold are licensed with our [SYBYL molecular modeling](http://www.accelrys.com/) software.

- Look/GeneMine
- Modeller
- SeqFold, Profiles3D and Homology are licensed with our [InsightII molecular modeling](http://www.accelrys.com/) software.

SwissModel accessed through the [SwissPDBViewer](http://www.expasy.ch/spdbv/) which we will use today.

Modeling of a three-dimensional structure is a worthwhile and worthy project.

It is not a substitute for evidence based determination of structure. If you have NMR or crystallography based evidence you will probably want to trust that evidence rather than your model. If there is a known structure use it. Once you have the pure protein and crystals determining the structure can go relatively quickly.

However, sequence information is much easier to come by than either NMR or x-ray data. If your sequence is similar to a known crystal structure, you can quickly make a reliable model in a few hours and really refine it in a week. You then have a tool to further your studies.

If your sequence is not like any known structure, you can try the ab initio methods but you may
have to settle for multiple alignments only.

---

**Today's Exercise**

A scenario (the main point of which is that you get a peptide sequence by some means or another):

As a result of a suspicious cluster of bacterial infections in first year graduate students an alert physician filed a report and a team of CDC investigators was dispatched. Some students were quite ill while others only exhibited minor symptoms. Samples from both groups were obtained. The CDC team quickly isolated a peptide elevated in affected compared to non-affected students. After analysis the teams have determined the mass and the sequences of CnBr digests of patient samples. Your job is to attempt to identify the origin of the peptide. That is, is it human or perhaps bacterial in origin. Then because of time constraints, you have to make a 3D model and attempt to discern the active site from a multiple alignment. In a nutshell your task is to identify the peptide, and proceed to make a homology model of the complete protein and guess its function in the affected people.

Optional section. If you want to save some of your results, the following paragraph helps you set up MS Word. Doing this also complicates the computer desktop and can be confusing.

Begin the exercise by mapping your personal Homeroom directory to your lab computer. On a Windows machine, right-click on the "MyComputer" icon. Select "Map Network Drive..." from the menu. Select a drive and enter \\homeroom.musc.edu\yourMNA". When prompted enter your password. On an Apple machine go to the **Apple Icon** in the upper left. Select **CHOOSER**. In the popup window Select **AppleShare** in the upper left and then type "homeroom" onto the Select IP address window on the right. Enter your MNA name and password. Select the folder with your MNA. Your Homeroom directory will now be mapped to the MAC and appear as a folder on the desktop. It will be useful to create a document to store your results and to act as an electronic notepad. Under the **CORE Applications** folder open **MS Office** and launch **MS Word**. Save the blank file to your desktop. Before you leave be sure to drag this file to your Homeroom folder.

The unknown peptide sequence is:

PGCKIRVTSKALELVKQEGLRFLEQELETITIPDLRGKE

You can type or copy/paste this sequence into the MS Word Notes file window and save the file to the desktop.

Next, we need to identify the protein this sequence codes for, identify the organism in which it is found, search to see whether a three-dimensional structure exists which can be used for structure-based drug design.

**Step 1: BLAST search to identify the protein's parent sequence.**

Open **Netscape** from the Communications Folder.
Go to the NCBI BLAST Server:

Select the standard protein-protein BLAST search and select the "SwissProt" database (its smaller than NR).

Copy the sequence (PGCKIRVSKALELKVQEGLRFLQELETITIPDLRGKE) and Paste the sequence into the form and launch a BLAST search.

**Step 2: Conserved Domain Analysis**

Unless you specifically request that this not be done, the query sequence will be compared to the Conserved Domain Database CDD

**Q1. Did the CDD search with our query show any hits? Click on the domain icon.**

This opens the Conserved Domain Database window. In this window you will see an alignment of our query peptide with the sequences of other proteins which also possess this domain. We can see precisely where our peptide aligns.

Click on the View 3D button and wait a moment while the molecular viewer Cn3D opens a 3D structure which has this domain.

In the Cn3D window, note that most of the structure is gray. The colored region in the structure window is the only part of the multiple alignment where all sequences are represented. This shows us where the peptide we isolated would most likely be in an actual 3D structure of the entire protein. What is the secondary structure that our peptide adopts? Not sure? Find the "style" menu and in that menu locate "coloring shortcuts" then select secondary structure. Gold colored regions are beta sheets, green colored regions are alpha helices and cyan regions are loops (ie neither sheets nor helices).

There should be a CDD Descriptive Items window. Click on the Show Annotations Panel button. On the left side of the new window, select "apolar binding pocket" then click "Highlight". In the structure and the alignment windows a number of residues are now colored contrasting yellow to indicate the residues which form the apolar binding pocket. Note as well the presence of a small molecule in the structure. This is a bound lipid (aka "apolar" ligand).

At this point we know the following about our peptide.

1. It matches a known conserved domain.
2. It will likely have a particular secondary structure.
3. There are conserved residues which might be candidates for site directed mutagenesis as we probe function.
4. We also know that our peptide is likely to form a part of a functional pocket which binds apolar ligands.

We know these things in just a few seconds. Meanwhile lets proceed to examine the complete BLAST search result.

**Step 3: Examining the BLAST results and linking to**

http://people.musc.edu/~hazards/FlexProtein2004.htm
LocusLink

Hit FORMAT! key to see results

Note: As you examine the BLAST results take a moment to fix in your mind the relationship between the E value (Expectation Score) and any alignment length and quality. In general the smaller the E value (closer to zero) the better the match between query and database entry. Beyond that a rule of thumb would be that a long range match (say over 200 residues) even though imperfect may be better than a short perfect match for identifying significant relationships between query and database entries.

You should see the following first line in the BLAST results page

gi|1709662|sp|P55058|PLTP_HUMAN Phospholipid transfer prote...

To the right of this first entry is a little blue box with an "L" in it. Click the little blue box.

This opens a LocusLink page. This is an entry from a curated (ie not archive only) database. As such it contains a wealth of information about the particular sequence we have been working on. Scan through the LocusLink page and answer the following questions.

Q2. What is the official gene symbol for this protein?

Q3. What does the Overview/RefSeq paragraph say about the protein's function?

Q4. Is there an OMIM entry for this gene/protein? Are there heritable diseases associated with this gene?

Q5. What's the most recent PubMed entry which refers to this gene/protein?

This could go on and on but you get the idea. The LocusLink entry is a treasure trove of links to known information about this gene/protein.

As a result of our data mining efforts, we are now well aware of what we are working with but the goal of today is to build an actual 3D model.

Step 4: Building a homology model

We need the entire sequence of our peptide. Scroll down the LocusLink page to the NCBI RefSeq section and click on the protein link NP_006218.

This opens the refseq entry for the protein. In the upper left next to the "DISPLAY" button is a rollover menu. From that menu select "fasta". The click on DISPLAY. Copy the resulting text to MS word. Save the document to the desktop AS TEXT ONLY not as a Word document. This "fasta" format is a sequence format that our homology modeling program can recognize.

For this we will use the SwissPD13 Viewer (SPDV). SPDV is a great tool and more complex than the tasks which we will ask of it today. Here is the URL for the online user

http://people.musc.edu/~hazards/FlexProtein2004.htm
SPDBV is a freeware program which you may obtain and install on your personal PC or MAC. It has been installed for your use in the Library MAC and PC labs. You will find it in the Special Applications Folder.

Do NOT download it to these machines.

Open the application SwissPDB Viewer.

Under menu SwissModel - click "load raw sequence to model" and select the file from the desktop in which you saved the BLAST hit in FASTA format.

SwissPDBV will read the file and build a protein. The structure appears as a long alpha helix.

Close the input log window. Locate the Control Panel (CP) window on the right and the Sequences Alignment window at the bottom. If neither appears go to the WIND menu and highlight the appropriate items.

Selecting residues within SPDBV is a bit awkward, requiring several steps and the use of information from several separate windows. As an introduction we will turn off the side chains, color the entire backbone light blue and then select the residues of the original peptide and color them red.

This next section is optional. The goal of these lines is to color the protein residues to make them easier to see on screen as a group.

- From the SELECT menu highlight "None". Next go to the Control Panel (CP) and click the mouse just under the "side" or sidechain column. All sidechains should disappear leaving only backbone atoms visible.
- From the SELECT menu highlight "all". Next go to the CP and click the mouse on the "col" or color column. From the popup window select a light blue color. All backbone atoms should appear as light blue.
- From the EDIT menu select "Find sequence". Enter the text of the original search sequence. Note that in the Sequence Alignment window these residues are now highlighted. Note as well that in the CP the text for these residues is red. At this point, use the mouse to select the COLOR column of the CP and select a contrasting color for these residues.
- At this point your display should show a light blue tube with a contrasting stripe and you should feel comfortable using the Alignment and CP windows to select residues. SPDBV may recolor the non-selected sequence to a dull gray after you use the "Find sequence" procedure. If this happens, select "Inverse Selection" recolor again the main protein. At this point the main part of the sequence will have one color and the original query peptide, a contrasting color.

With or without a colorized protein you need to continue the modeling from here on.

Now with the sequence loaded, we need to find a template to build the structure. From the SwissModel menu select "Find Appropriate EXPDB templates". This SHOULD open a new Netscape window. Double check that the entry corresponds to your sequence and click

http://people.musc.edu/~hazards/FlexProtein2004.htm
From the results window click on the hyperlink to 1BP1 (that's "one BP one"; note that this is the same entry as we found in the CDD search) and save it to the desktop.

Back in SPDBV, from the File menu select the Open PDB entry and read PDB file you just saved.

Repeat the procedures you performed above to color the Template entry some contrasting color and to turn off the side chains.

Under the WIND menu select Alignment. This displays the default alignment between the template PDB sequence and the sequence we are modeling.

Be sure that the query sequence is highlighted and then under the Fit menu hit Magic Fit. This performs an automatic alignment and threads our sequence to the template structure. This is step one in making a homology model. Accept the Magic Fit alignment. [Can't see it? - click top left button of display (the one with the three red arrows). This centers the structurally aligned protein to the template structure.]

A structure is generated - At this point you have "threaded" the sequence of the unknown structure onto the template structure. This is a crude first approximation of the final model.

In the Alignment window above the sequences is a quantitative plot of the “goodness of fit" between the template and the unknown (if this is not visible click the mouse pointer on the small triangle next to the query sequence name in the Alignment window and the window should expand to show the"fit" plot). Use the scroll bar to move down the sequence. Note the identities and similarities. Note also, as you examine the Alignment window you may highlight either the 1BPI template structure or the modeled structure. As you do this the display in the CP window changes and the colors change in the molecular display window. You can use the second, third and fourth buttons at the top to move the molecular display to the left/right, zoom or rotate respectively. When you highlight residues in the Alignment window, these residues will flash in the molecular display.

Go back to the SwissViewer and highlight the residues of the initial peptide. You can use the Edit Menu entry "find sequence" to locate the regions in the Alignment window. Simultaneously these same residues will "flash" in the molecular view window. Or you could click on the Side chains of these residues in the Control Panel. Complicating the issue is that the numbering of the structure will be different from the numbering of the sequence.

Q4. According to the Alignment Window plot are the residues of the initial peptide well modeled with respect to the template? Are there any "local" areas where the current model is not optimal?

It turns out that in making homology models, the alignment is a critical issue. The SwissPDB "Magic Fit" which we used is a good first try. The GI for the template is 2554693 while that of the PLTP sequence is 5453914. Go to the BLAST 2 Sequences site, enter the GI numbers, be sure to select BLASTP for protein comparison and make a comparison between the BLAST2 alignment and the SWPDBV "Magic Fit" alignment. You might decide to alter the alignment based on your BLAST2 results.
The main point is that quick and dirty alignments MAY NOT be the most reliable.

**Step 5 Submitting an alignment to SwissPDB server for model refinement.**

When you are satisfied with the alignment you can complete the modeling process with a few more steps. Go to the Swiss Model Menu item of the SPDBV program and select Submit Modeling request. Enter your name and your e-mail address, name the project, identify the project to the Netscape window, verify the information and press submit. Within 24 hours (often much less) you will receive a reply from SwissModel which contains an energy refined (as opposed to a roughly threaded) model of the protein. You can view this structure very conveniently using Swiss PDBViewer or a program like RASMOL (there are many viewers out there).

Final comments:

**BPI binds apolar lipids** from gram negative bacteria while **PLTP** binds and transfers phospholipids from lipoproteins to HDL particles. The conserved domains common to both proteins imply a common function even if we had not read the LocusLink/RefSeq/SwissProt annotations. The sequence similarity by itself is a very strong indication that the overall fold of the proteins is similar. These factors suggest the proteins are related (ie homologous) but specialized to perform similar but distinct functions. These two proteins probably arose through an ancient gene duplication. They are therefore paralogous.

By examining the putative lipid binding domains in the model and comparing the similar region in the template, we might be able to make informed guesses about how the sequence differences create the specialized lipid binding pockets. This sort of speculation would not be possible without the model (ie it does not follow in any straightforward way from the sequence). Until a structure is determined via an experimental method this model is our best bet for developing testable hypotheses about how the substrate specificity is determined.

**Step 6. Your take home assignment is to perform a similar analysis entirely on your own the following protein sequence, analyze the results and turn in via email an assessment of the structure.** E-mail this assignment by September 10 at 6pm.

From an organism you obtain this peptide sequence:

**TRVKTTGIVETHFSFKNLFRLFDVGGQRSERKKWIHCEDVTAIIFCVAMSGYDQMLHEDET**

Retrieve the most probable complete sequence by performing a BLAST search. Identify the organism.

Note any conserved domains. If any conserved domains are present, note conserved residues and within the peptide note secondary structure where present.

Use the entire sequence to create a homology model via the same methods used during the flextime session.

Based on the structure you have built propose a function for the peptide based on the chemical
properties of the component amino acids and their positions in the larger structure. Suggest residues which might be critical for function and defend why you think they might be critical. For example, "the ser at position 14 is conserved (based on CDD) and may have a catalytic function based on the properties of the the hydroxyl group and based on its proximity to...."

Would these residues be candidates for mutation experiments to explore the full protein function?

Write a short paragraph addressing the above questions about your modeled protein and email this to Dr. Starr Hazard (hazards@musc.edu).

This e-mailed analysis will be graded as part of the protein section. A written paragraph is required. You may append your model and or images which illustrate your points. This should take no more than about an hour once you understand the process and the issues. E-mail this assignment by September 10 at 6pm.

In addition to Chapters 3, 4 of Stryer here are links to some information, manuals and wider explanations of the exercise.

Amino Acid Properties A

Tutorial for using Cn3D

BLAST searches explained

BLAST search Glossary

SwissPDBViewer and Cn3D used together. A review of the Flextime in-class exercise.

SwissPDBViewer On-line tutorial

SwissPDBViewer User Guide

Some Other Molecular Viewers RASMOL ,CHIME and Protein Explorer Any of these freely available and related programs display molecules.

Pymol and VMD are two more molecular viewers with useful features which, in particular, are also available for the MAC OS X operating system.

Conservation of amino acids. Consider this simple multiple alignment:

<table>
<thead>
<tr>
<th></th>
<th>C. elegans (worm)</th>
<th>X. laevis (frog)</th>
<th>G. gallus (chicken)</th>
<th>B. taurus (cow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly(G)</td>
<td>GLASYTE</td>
<td>GVGSPSE</td>
<td>GLASWSE</td>
<td>GIASITE</td>
</tr>
<tr>
<td>Ser(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu (E)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gly(G), Ser(S) and Glu (E) are conserved from worms through humans in this alignment. We infer therefore that there is some constraint on the size, polarity or charge at those three positions in the parent protein. By examining these residues in the three-dimensional space of say a homology model we might be able to further infer function for these residues in the
modified by ESH September 2, 2004 Questions/Comments to ESH