BSC4933/ISC5224: Introduction to Bioinformatics

Laboratory Section: Tuesdays from 2:00 to 5:00 PM in Dirac 152.

Protein Homology Modelling

Lab Ten, Wednesday, March 18, 2009

Author and Instructor: Steven M. Thompson

Protein tertiary structural inference through homology modelling and visualization:

Homology modelling combines sequence analysis and molecular modelling to predict three-dimensional structure. You will choose a remote homologue of your Project protein that has not yet had its structure solved, and use the SwissModel WWW resource to model the molecule. The theoretical structure will then be visualized with the SwissPDBViewer and RasMol to gain insight into the way in which its structure relates to its function. Color coding different physical attributes such as residue charge, hydrophobicity, and secondary structure elements; different representations, such as alpha-carbon traces, ‘cartoon’ graphics, and space-filling models; and super-positioning of the model with an actual structure all assist in the interpretation.
Introduction — protein tertiary structure homology inference —
What can I learn about my protein’s tertiary structure based on its homology with solved structures?

Grateful acknowledgement of Susan J. Johns (University of California, San Francisco) for contributing much to this tutorial from our days together at Washington State University (1990–1998).

NMR and X-ray protein crystal structures are classified into families based on a limited set of folding patterns. See for example SCOP (Structural Classification of Proteins http://scop.berkeley.edu/) and CATH (Class, Architecture, Topology, and Homologous superfamily http://cathww.biochem.ucl.ac.uk/latest/). In general, functional sites maintain identical structural folds. And proteins of the same function generally have similar structure, (notable exceptions of molecular convergence, such as trypsin and subtilisin, outstanding). Such observations became the foundation of homology modelling. The key to its success is a thorough understanding of your dataset.

You need at least a pair of sequences for homology modelling. Multiple sequence alignments make the technique even more powerful. One member of the dataset is modeled; the others need to include at least one reference sequence that has three-dimensional coordinate data associated with it, upon which the model is based. The reference sequences are found through database similarity searching. A FastA or BLAST search of a large protein database, such as UniProt, can provide potential candidates. However, you need to figure out which entries have solved structures afterwards, pretty easy in a graphical display sequence editor like SeqLab. Another way would be a BLAST search of the MMDB resources at NCBI. That would simplify post-search processing, but requires separate downloads. Or you could create your own custom database of solved structures and use it. Regardless, even sequences whose statistical similarity are quite low to solved reference structures can sometimes still be modeled, if they share common disulfide bridge pattern(s), or perhaps functional domains, or if a reliable multiple sequence alignment can be built with a broad range of similar reference sequences.

Homology modelling combines and requires the techniques from two sets of biocomputing toolkits — sequence analysis and molecular modelling. To be successful you need to thoroughly analyze your query protein sequence — know as much about it as possible. What other proteins is it homologous to? What is the best alignment possible with these related sequences? What are those relationships, i.e. what is the inferred phylogeny? How are regions of insertions or deletions, indels, in the alignment dealt with? Which of these homologues have solved structures? If so, what are the coordinates? What structural elements are they composed of? The reference proteins’ database entries document feature locations such as functional sites and established motifs. Does your protein have all the same features? What are these proteins’ physical characteristics? Do you know your protein’s function? Is it the same as the reference sequences’? The list goes on and on . . .

Running sequence analysis programs, such as those from the GCG package, on the primary sequences can generate tons of information on known motifs and domains, and hydrophobicity and amphilicity patterns. The
secondary structure assignments for the reference proteins will be given in the record, but were those assignments made by the author manually, or with a computer program such as Define_S or DSSP (see references) based on the coordinate data, and if so, do they agree with manual assignments? If not, where do they disagree and why? What sort of secondary structure predictions can be made on your unknown protein by the more reliable secondary structure prediction programs such as PredictProtein and nnpredict? This will provide ideas of where structural elements exist. That information can be used to assist in the alignment process. And knowing the most reasonable phylogeny for the sequences provides valuable insight into both function and structure, much more so than mere similarity. This is known as “phylogenomics.” In other words, you need to learn all the same sort of stuff about your unknown protein as we’ve been learning so far this entire semester!

And reference tertiary structure can provide clues as well. What does visual examination of the structure reveal about the interplay of secondary structural elements with one another? Are certain parts of the molecule occupied with functional aspects of the protein, while others serve as spatial elements to properly position the functional portions so that interactions can occur? Are there complex folding patterns present in the structure representative of a given family of proteins that need to be maintained? Are there any structural elements in the protein that don’t belong to this folding pattern? Have substrates been determined in the dataset and how do they effect the conformation of the given functional site? Is the protein a composite of different types of functional domains? How do they interact with one another? Are disulfide bridges important to the integrity of the structure? If so, where are they and what are the residues connected in this way?

The basis of the entire process is the alignment of the primary sequences. Sometimes alignments will need to be forced, e.g., disulfide bonds are extremely important to protein structure so cysteine residues may have to be manually aligned to make disulphide bridge placement correspond. Functional and physical sites need to align between sequences. Discrepancies between a reference protein’s mature form, as usually represented in its X-ray structure, and its immature form will need to be reconciled. Any leader peptide portion of your unknown protein will often need to be removed before you can build a successful model. The overall quality of the alignment determines the quality of the final model. Therefore, as in phylogenetic inference, it is extremely important to use the best alignment possible.

After alignment the unknown sequence is overlaid on the backbone sequence of the coordinate data according to the correspondence of residues in the alignment. Indels are always trouble. Carefully evaluate any areas where gaps occur in the unknown sequence and not in any of the coordinate data. Is it a loop section that could easily be clipped out? Could the gap be removed by rotating one or two existing residues on either side close enough to one another to form a normal peptide bond? If so, go ahead and make the change in the coordinate structure data set prior to the overlay. When an insert is called for, look at the area where it would occur. Is it on the surface of the molecule? Is the region hydrophobic or hydrophilic? What are the rest of the secondary structural elements in the area? Could a similar type of structural unit be created that would be consistent with the rest of the area and that would match the hydrophobicity
requirements of the insert? Almost any configuration that meets those requirements can be created as long as the insert is outside of any functional sites.

The structural elements involved in the functional sites must be kept in the same general spatial positions. Particularly suspect are any models that modify structural members of functional sites. Changes in these areas need to be coupled with changes elsewhere in the molecule that will allow the functional site to remain basically intact, even if component helices or sheets are now longer or shorter than in the original. If these changes can be made to keep truly vital residues in the same spatial locations, then the modifications can be tried and tested visually for realism. The results should then be minimized or at least passed through some sort of distance geometry program to get the side chains into more realistic positions.

All of this is a huge, labor-intensive task. Fortunately, an automated homology modelling tool is available on Bairoch’s ExPASy Web server. Supported by GlaxoSmithKline and the Swiss Institute of Bioinformatics (SIB), it is called SwissModel (http://swissmodel.expasy.org/, see e.g. Guex, et al. [1999] and Guex and Peitsch [1997]). It has dramatically changed the homology modelling process. It is a relatively painless way to get a theoretical model of a protein structure. While not always successful in generating a homology model for your sequence, depending on how similar the closest sequence with an experimentally solved structure is to it, the minimal amount of effort involved in making the attempt makes it an excellent time investment. It is a very reasonable first approach, and will often lead to remarkably accurate representations. An extensive menu and help system is provided on all SWISS-MODEL Web pages. As with PredictProtein, you can submit an individual sequence and the server will perform a database search, in this case against all of the sequences from the three-dimensional Protein Data Bank, and then create a multiple alignment of the significant hits, and then provide a structural inference. This is “First Approach mode,” or you can submit your own customized and carefully scrutinized multiple sequence alignment using the “Alignment Interface” or the “Project (optimise) mode.” The “Alignment Interface” has a list of supported formats with FastA being the most basic, and there are some tricks to using project mode. Naturally, your template sequences must have solved structures in both cases; however, the SwissPDBViewer must be used to format and submit your data in project mode. SwissPDBViewer is an interactive molecular structure viewer and editor, also developed at GlaxoSmithKline and SIB, that allows super-positioning of structures and their corresponding sequences. It has versions for most major operating systems.

Conformation of the request submission and your results (whether they are good or bad) were formerly sent back via e-mail. Results are now returned in the SWISS-MODEL Web Workspace itself. The PDB format coordinates for the model, an interactive Java driven graphical model, a complete log of all the server actions, and the results of model evaluations are all available. You can also save results as a SwissPDBViewer (DeepView) project file containing PDB formatted coordinates for the model and templates superimposed
I submitted a *Giardia lamblia* Elongation Factor 1α sequence to SWISS-MODEL in “First Approach mode.” The results were available in less than five minutes.

The figure to the right here displays a RasMac (http://www.umass.edu/microbio/rasmol/ but only available pre-OS X [see e.g. Sayle and Milner-White, 1995]) “Strands” graphic of the *Giardia* EF-1α structural model from SWISS-MODEL superimposed over the eight most similar solved structural templates:

**Standard disclaimer**

For the last time this semester: I write these tutorials from a ‘lowest-common-denominator’ biologist’s perspective. That is, I only assume that you have fundamental molecular biology knowledge, but are relatively inexperienced regarding computers. As a consequence of this they are written quite explicitly. Therefore, if you do exactly what is written, it will work. However, this requires two things: 1) you must read very carefully and not skim over vital steps, and 2) you mustn’t take offense if you already know what I’m discussing. I’m not insulting your intelligence.

I use three writing conventions in the tutorials, besides my casual style. I use **bold** type for those commands and keystrokes that you are to type in at your keyboard or for buttons or menus that you are to click in a GUI. I also use bold type for **section headings**. Screen traces are shown in a ‘typewriter’ style *Courier* font and “///////////” indicates abridged data. The dollar sign ($) indicates the system prompt and should not be typed as a part of commands. Really important statements may be underlined.

As you’ve learned in the previous labs, specialized X-server graphics communications software is required to use GCG’s SeqLab. X11 is also required by RasMol and the SwissPDBViewer when they are installed on UNIX systems and you wish to display local windows from them. I’ll remind you of a few things while using X: X Windows are only active when the mouse cursor is in that window, and always close X Windows when you are through with them to conserve system memory. Furthermore, to activate X items, just <click> on them, rather than holding your mouse button down. Also, X buttons are turned on when they are pushed in and shaded. Finally, don’t close X Windows with the X-server software’s close icon in the upper right- or left-hand
window corner, rather, always, if available, use the window’s own “File” menu “Exit” choice, or “Close,” or “Cancel,” or “OK” button.

The project molecules (one last time)

Here are some representative PDB entries and their corresponding SwissProt entry for the project molecules for your reference. Which other ones have you found in your searches? All have several solved structures.

1) Primitive plant small subunit of ribulose bisphosphate carboxylase/oxygenase — 1GK8 (complex with four large and four small subunits from Chlamydomonas reinhardtii), SW: RBS1_CHLRE;
2) Vertebrate ras P21 transforming protein — 6Q21, SW: RASH_HUMAN;
3) Vertebrate basic fibroblast growth factor — 4FGF, SW: FGF2_HUMAN;
4) Fungal superoxide dismustase — 1SDY, SW: SODC_YEAST.

A ‘Real-Life’ Project Oriented Approach — Protein Homology Modelling

As you’ve done all semester long, activate and log on to the computing workstation you are sitting at and then log onto HPC with an X-tunneled ssh session. From an X-aware terminal window issue the following command (the X has to be capitalized and replace “user” with your account name):

$ ssh -X user@submit.hpc.fsu.edu

Preliminary preparations

Change your directory from ‘home’ to last week’s subdirectory. List that directory (ls) and check out the files left over from last week’s tutorial. Look through them and remove any that you don’t want to save. Next, change directory back to your home directory, create a subdirectory for this week’s tutorial data, and then change directory into it. After you’ve taken care of these file maintenance chores launch SeqLab in the usual manner, “seqlab &.”

Next, verify that you are in SeqLab’s “Main List” “Mode:” The list file in the main window should be that from last week. Select your RSF file used in that tutorial, the one created and refined in Lab Seven. This should be an aligned dataset containing fifty or so of your selected project molecular system sequences, annotated with database and program-generated feature descriptions. Switch “Mode:” to “Editor.” Select the same sequence that you used last week for secondary structure inference. This sequence should not have had its three-dimensional structure solved (as described in its database annotation), and should be one of the least similar of those sequences in the alignment to your original query sequence. <Double-click> its name (or use the “INFO” icon) to note the sequence’s UniProt accession code, not its identifier name. This sequence should be duplicated in your dataset; you’ll have the original ‘gapped’ sequence, as well as the ‘ungapped’ copy that you used with most of last week’s secondary structure inference programs. Select and then “CUT” the ungapped copy away. Also be sure there are no text annotation or mask lines in the dataset. If there are “CUT,” them away too. Then go to the “File” “Export” menu and export a “FastA” copy of it. This “Export”
command does not send output to the output manager, rather you'll need to verify that the file is now in your current directory using your terminal window. Sometimes FASTA format will work where nothing else will; it's a very basic format. You'll need this file in today's homework. Leave SeqLab now, as we won't be using it anymore today. Exit SeqLab with the "File" menu "Exit" choice. There's no need to save any changes; SeqLab will close.

Using SwissModel

You can do the rest of this week's tutorial on one of the classroom workstations or on your own personal machine, the requirement being RasMol and SwissPDBViewer need to be installed on the machine. These tools are not installed on HPC, as that platform's mission is not visualization. Both of these programs have been installed on the Dirac 152 classroom workstations. I'll explain how to do it on that platform, but if you prefer, you are welcome to use your own machine. The concepts are the same. Cut-and-paste operations are notoriously problematic though, so you will have to use some form of secure file transfer to copy your sequence alignment file from HPC to your local machine, however you do it. If you are logged onto a classroom workstation, the syntax for the command to move the file from HPC to your local machine follows:

$ scp user@submit.hpc.fsu.edu:path/file path/file

Launch whichever Web browser you like to use on your local computer, and connect to the SwissModel Web site (http://swissmodel.expasy.org/). This site provides access to a form driven “Workspace” for submitting homology modelling requests. It is pictured below:
Find the section of the page entitled “Modelling requests.” Select “First Approach mode.” This will move you to another page. Once there fill in all the necessary fields: your personal e-mail address, a descriptive title for the requested modelling job, and a SwissProt accession code (not ID name). Remember to use your own project molecule’s SwissProt accession code, not mine. After you’ve gotten results back from SwissModel visualize them on one of the classroom computers with RasMol and SwissPDBViewer. Another alternative is to install RasMol and SwissPDBViewer on your own personal machine.

I’ll show a screen snapshot of the form below:

![Screen snapshot of the form](image_url)

Press the “Submit Modelling Request” button. You’ll get a new screen informing you that your request has been sent off. A screenshot graphic of this next window follows below:

![Screenshot graphic of the next window](image_url)
In the past, if the requested model is one that the server has in its repository, then the request is routed to the repository instead of repeating the modelling task, and you are asked if you want to download the coordinates. We'll see if that happens to any of you this semester. Otherwise, SwissModel will present all of the results through the link on the above page. In my case those greatly abridged results follow:

SWISS MODEL WORKSPACE
An Automated Comparative Protein Modelling Server
Workunit: P000001
Title: EF1a test

------------------------------------------
Target: modelled residue
range: 1 to 396
Alignment
TARGET  1  STLTGHLIY KCGGIDQRTI DEYERKRAM
1f60A  2  gkekshinv vighvdsgk- -sttgthly kcggidkrti ekekeaael
TARGET  30  hhhhhhh h hhh hhhhhhh
1f60A  50  gkgsfkyaww ldklkaerar gitidialwk fetpkyqtv idaphrdfi
TARGET  80  sssssss sssssss sssssss sssssss ssssss sssss
1f60A  100 hhhhhhhhh hhhhhhhhh hhhhhhhhh hhhhhhhhh
TARGET  130  gdnimekSDK MPYEG---- -------PC LIDAIDGLKA PKRPTDKPLR
1f60A  150 vavnmdsvk --waresrfge ivketsnfik kvgyout--kt vpfvpisgw
1
TARGET  179  sssssss sssssss sssssss sssssss sssssss sssssss
1f60A  196 gdnmleatatn apwykgweke tkagvvgkt lleaidaleq psrptdkplr
Anolea / Gromos / Verify3D

Modelling log
3.70 (SP3)
Loading Template: 1f60A.pdb
Loading Raw Sequence
Loading Alignment: ./user.align.submit.fasta.FF
Removing HET groups from template structure
Refining Raw Sequence Alignment
ProModII: doing simple assignment of backbone
ProModII: adding blocking groups
Adding Missing Sidechains
AddPolar H
BuildDeletetedLoopsModel
Trying Ligating with anchor residues GLU 193 and CYS 196
Trying Ligating with anchor residues TYR 192 and CYS 196
Number of Ligations found: 47
ACCEPTING loop 35: clash= 0 FF= 35.4 PP= -1.00

Trying Ligating with anchor residues LEU 353 and GLU 356
Trying Ligating with anchor residues THR 352 and GLU 356
Trying Ligating with anchor residues ARG 351 and GLU 356
Trying Ligating with anchor residues LY 350 and GLU 356
Number of Ligations found: 9

Number of Ligations found: 4

Number of Ligations found: 18

Number of Ligations found: 345

Number of Ligations found: 500

ACCEPTING loop 132: clash= 0 FF= 378.0 PP= 0.00

Number of Ligations found: 2

Number of Ligations found: 92

Number of Ligations found: 91: clash= 0 FF= 1236.5 PP= -5.00
Optimizing Sidechains

Adding Hydrogens

Optimizing loops and OXT (nb = 26)

Final Total Energy:  -12826.456 KJ/mol

Dumping Sequence Alignment

Download the PDB coordinate file from the Web page, and then display it with “more” to get a feel for what they look like. The following is a screen trace of the first several lines from my example:

HEADER    SWISS-MODEL (Automated Protein Modelling Server)
TITLE     Elongation factor 1-alpha (EF-1-alpha)
ACCODE    P02993_C00001
COMPN    MOL_ID: 1;
COMPN    2 MOLECULE: Elongation factor 1-alpha (EF-1-alpha);
SOURCE    MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: Artemia salina ;
SOURCE    3 ORGANISM_COMMON: Brine shrimp;
KEYWDS    Elongation factor; Protein biosynthesis; GTP-binding; Methylation.
EXPDTA    THEORETICAL MODEL (SWISS-MODEL SERVER)
AUTHOR    ProMod (SEE REFERENCE IN JRNL Records)
REVDAT    1  12-MAR-2003
JRNL    1 AUTH   M.C.PEITSCH
JRNL    1 TITL   PROTEIN MODELLING BY EMAIL
JRNL    1 REF    BIO/TECHNOLOGY V.  13  658 1995
JRNL    1 REFN   ISSN 0733-222X
JRNL    2 AUTH   M.C.PEITSCH
JRNL    2 TITL   PROMOD AND SWISS-MODEL: INTERNET-BASED TOOLS FOR
JRNL    2 TITL 2 AUTOMATED COMPARATIVE PROTEIN MODELLING.
JRNL    2 REF    BIOCHEM. SOC. TRANS. V.  24  274 1996
JRNL    3 AUTH   M.C.PEITSCH,N.GUEX
JRNL    3 TITL   LARGE-SCALE COMPARATIVE PROTEIN MODELLING.
JRNL    3 REF    PROTEOME RESEARCH: NEW FRONTIERS IN FUNCTIONAL
JRNL    3 REF  2 GENOMICS. 177 1997
JRNL    4 AUTH   M.C.PEITSCH,N.GUEX
JRNL    4 TITL   SWISS-MODEL AND THE SWISS-PDBVIEWER: AN
JRNL    4 TITL 2 ENVIRONMENT FOR COMPARATIVE PROTEIN MODELLING
JRNL    4 REF    ELECTROPHORESIS V.  18  2714 1997
JRNL    5 AUTH   N.GUEX,A.DIEMAND,M.C.PEITSCH
JRNL    5 TITL   PROTEIN MODELLING FOR ALL
JRNL    5 REF    TIBS V.  24  364 1999
JRNL    6 AUTH   T.SCHWEDE,A.DIEMAND,N.GUEX,M.C.PEITSCH
JRNL    6 TITL   PROTEIN STRUCTURE COMPUTING IN THE GENOMIC ERA
JRNL    6 REF    RESEARCH IN MICROBIOLOGY V. 151 107 2000
REMARK    1
REMARK    2 RESOLUTION. NOT APPLICABLE. SEE REMARK 4.
REMARK    3
REMARK    3 REFINEMENT. NONE.
REMARK    3
REMARK    3 MODELLING METHOD.
REMARK    3 PROGRAM : ProModII
REMARK    3 AUTHORS : N.GUEX,M.C.PEITSCH
REMARK    3
REMARK    3 ENERGY MINIMISATION.
REMARK    3 PROGRAM : GROMOS96
REMARK    3 AUTHORS : VAN GUNSTEREN
REMARK    3 PARAMETER : IFP43B1
REMARK    3 TOPOLOGY FILE : MTB43B1
REMARK    3 METHOD 1 : STEEPEST DESCENT
REMARK    3 CYCLES 1 : 200
REMARK    3 CONSTRAINTS 1 : 25 / C-FACTORS
REMARK    3 METHOD 2 : CONJUGATE GRADIENT
REMARK    3 CYCLES 2 : 300
REMARK 3 CONSTRAINTS 2 : 2500 / C-FACTORS
REMARK 3
REMARK 4
REMARK 4 THE COORDINATES IN THIS ENTRY REPRESENT A MODEL STRUCTURE.
REMARK 4 PROTEIN DATA BANK CONVENTIONS REQUIRE THAT *CRYST1* AND
REMARK 4 *SCALE* RECORDS BE INCLUDED, BUT THE VALUES ON THESE
REMARK 4 RECORDS ARE MEANINGLESS.
REMARK 5
REMARK 5 THIS MODEL IS BASED UPON THE COORDINATES OF:
REMARK 5
REMARK 5 PDB ENTRY 1f60.pdb
REMARK 5 TRANSLATION 19-JUN-00
REMARK 5 CRYSTAL STRUCTURE OF THE YEAST ELONGATION FACTOR COMPLEX
REMARK 5 EEF1A:EEF1BA
REMARK 5 X-RAY DIFFRACTION
REMARK 5 RESOLUTION. 1.67 ANGSTROMS.
REMARK 5
REMARK 5 PDB ENTRY 1ijf.pdb
REMARK 5 TRANSLATION 26-APR-01
REMARK 5 NUCLEOTIDE EXCHANGE MECHANISMS IN THE EEF1A-EEF1BA COMPLEX
REMARK 5 X-RAY DIFFRACTION
REMARK 5 RESOLUTION. 3.00 ANGSTROMS.
REMARK 5
REMARK 5 PDB ENTRY 1ije.pdb
REMARK 5 TRANSLATION 26-APR-01
REMARK 5 NUCLEOTIDE EXCHANGE INTERMEDIATES IN THE EEF1A-EEF1BA
REMARK 5 COMPLEX
REMARK 5 X-RAY DIFFRACTION
REMARK 5 RESOLUTION. 2.40 ANGSTROMS.
REMARK 5
REMARK 5 PDB ENTRY 1g7c.pdb
REMARK 5 TRANSLATION 10-NOV-00
REMARK 5 YEAST EEF1A:EEF1BA IN COMPLEX WITH GDPNP
REMARK 5 X-RAY DIFFRACTION
REMARK 5 RESOLUTION. 2.05 ANGSTROMS.
REMARK 5
REMARK 5 PDB ENTRY 1jny.pdb
REMARK 5 TRANSLATION 26-JUL-01
REMARK 5 CRYSTAL STRUCTURE OF SULFOLOBUS SOLFATARICUS ELONGATION
REMARK 5 FACTOR 1 ALPHA IN COMPLEX WITH GDP
REMARK 5 X-RAY DIFFRACTION
REMARK 5 RESOLUTION. 1.80 ANGSTROMS.
REMARK 5
DBREF P02993_C00001          SWS P02993 1 442
IPREF P02993_C00001 P02993 1 442 885E6E95E0779687
SEQALI
SEQALI P02993 1 GKEKIHIN IVVIGHVDSG KSTTTGHLIY KCGGIDKRTI EKFEKEAQEM
SEQALI 1f60A 2 GKEKSHIN VVVIGHVDSG KSTTTGHLIY KCGGIDKRTI EKFEKEAAEL

Not all requests are successful. If the process is unsuccessful, you may get a message with the words “No Success” stating that the attempt was unsuccessful and suggesting that you cut down your sequence to only those areas of good alignment before attempting to do any further modelling on the protein. The similarity quality requirements should be given in the message:

Please look at the other messages issued by the server.
The degree of similarity of your sequence with proteins of known 3D structure may be too low.
At present, Swiss-Model will generate models for sequences which respond to these criteria:
BLAST search P value : < 0.0001
FASTA search standard deviations above mean : > 9.0
Global degree of sequence identity (SIM) : > 25 % spread of > 40% of the submitted sequence.
Molecular Visualization

Use your Web browser to connect to the “World Index of Molecular Visualization Resources” (http://molvis.sdsc.edu/visres/index.html). This is a fantastic resource for information, tutorials, free tools, and commercial software in the broad field of molecular visualization and modelling. I heartily recommend spending some time perusing its contents.

Three major pieces of molecular visualization software have been installed on the SCS classroom computers: Cn3D, RasMol, and SwissPDBViewer. The first, Cn3D, requires ASN.1 structural data as formatted by NCBI, rather than PDB format data as sent by SwissModel, and therefore, is not appropriate for this tutorial, though it works great to visualize the results of Entrez Structure searches performed at NCBI, which I encourage you to check out. The latter two do read PDB format and will be used today. I will only provide minimal guidance in their use and want you to explore their many facets on your own.

Connect to the RasMol (http://openrasmol.org/) and SwissPDBViewer WWW home pages (http://us.expasy.org/spdbv/). Go over the documentation found there to familiarize yourself with these programs’ use and capabilities. These sites contain links to program manuals, tutorials, FAQ sheets, and other vital documentation, as well as links to the program downloads themselves.

Be sure that you saved your Web data from SwissModel as a PDB file, and that you saved (and then secure transferred, if necessary) the PDB coordinate file to whichever machine you plan on using for visualization. To run RasMol on a classroom workstation type the command “rasmol filename.pdb &” in your terminal window, where “filename.pdb” is the name of the PDB format file that you got from SwissModel. Explore the various visualization alternatives offered by RasMol. “Quit” RasMol afterwards with the appropriate menu command. The X11 window should go away and return you to your terminal.

SwissPDBViewer (also known as DeepView) is also available on the SCS classroom computers. Launch SwissPDBViewer on a classroom computer from your terminal window with the command “spdbv &.” Use the “File” menu to specify your SwissModel PDB file to load it into the program. As with RasMol, just explore the software to gain some comfort with its use. SwissPDBViewer is way more powerful and much more complicated than RasMol, and is capable of all sorts of visualization and analysis techniques, including limited minimization. Spend some time checking out its many features.

Things to investigate in both programs include the color coding of different physical attributes such as residue charge, versus hydrophobicity, versus secondary structure elements; different representation models, such as alpha-carbon traces, versus space-filling, versus ‘cartoons;’ and superpositioning of the model with an actual structure. Notice that it also includes a multiple sequence alignment visualization tool that tracks the superpositioning from the structure onto the sequence. “Quit” SwissPDBViewer after exploring for a while.

Log off your current session on the workstation that you are using.
Homework assignment

Use the “SwissModel Alignment Interface” to submit your entire alignment to SwissModel. The Report Form will ask you to discuss your experience, pros and cons, and note any significant differences between the results of this submission mode and the previous “First Approach mode.”

Conclusion

Gunnar von Heijne in his quite readable but dated treatise, *Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit* (1987), provides an appropriate conclusion to this tutorial series:

“Think about what you’re doing; use your knowledge of the molecular system involved to guide both your interpretation of results and your direction of inquiry; use as much information as possible; and do not blindly accept everything the computer offers you . . . if any lesson is to be drawn . . . it surely is that to be able to make a useful contribution one must first and foremost be a biologist, and only second a theoretician . . . We have to develop better algorithms, we have to find ways to cope with the massive amounts of data, and above all we have to become better biologists. But that’s all it takes.”

Resources and References


