CS444 (001) Homework4

George Mason University
Instructor: Amarda Shehu

Out: November 13, 2012
Due: December 06, 2012

Problem 1. Obtaining a target sequence for knowledge-based modeling [25 pt]

Your goal here is to choose an amino-acid sequence for a human enzyme with sequence identity in the [30% 70%] range with other proteins of known structure. Follow these steps:

1. Perform an advanced search in Uniprot (http://www.uniprot.org/) with the following options:
   enzyme AND length:[100 to 300] AND organism: Homo Sapiens [9606]"
   This should give > 800 human enzymes with length between 100 and 300 amino acids.
   The "Download" option should provide you with a FASTA file for all entries.
   Let's assume that you named this file uniprot_results.fasta

2. You now need to narrow this list to enzymes who have identity in the [30% 70%] range.
   To do so, download first a local copy of Fasta_vr.35, which is the most recent version, at (http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml)
   There are detailed instructions on compiling and usage at the above website
   In order to know the sequences of all proteins with known structures, you need to obtain a FASTA file for all PDB entries. Obtain this file and name it pdb_seqres.fasta
   Using Fasta35, retrieve the top 10 matches of each Uniprot query sequence from the PDB using the following command line input:
   fasta35 -d 10 -H -q uniprot_results.fasta pdb_seqres.fasta > results.out

3. Scroll through the results to identify an entry from Uniprot whose sequence identity with everything in the PDB is in the [30% 70%] range. You will see that, for instance, P31941 (Probable DNA dc->dU editing enzyme APOBEC-3A) is one such entry. One its top matches in the PDB has sequence identity 65.3% (PDB Entry 2KBO:A). Choose another protein sequence as your query/target sequence. Confirm that its has indeed sequence identity as specified by pasting it in NCBI's BLAST searching against entire PDB.

Check the end of this document for what you need to submit as deliverables.

Problem 2. Comparative Modeling of Structure for Chosen Target Sequence [20 pt]

You will use SWISS-MODEL to obtain a model structure for the protein sequence you chose above. SWISS_MODEl (http://swissmodel.expasy.org/) is a comparative modeling webserver. While it exists in software version, I suggest you use the webserver. I also suggest that you use the "Alignment Mode" rather than "Automated," as it tends to be more accurate.

SWISS-MODEL needs sequences for the alignment, and you need to provide these. As before, go to Uniprot, select the "BLAST" tab, paste the FASTA sequence of your query/target protein, and perform a search using default settings. Download the results in FASTA format using the "Download" option. The file contains proteins that are similar to your target protein. You need to obtain a multiple sequence alignment of these for SWISS-MODEL. To do so, load the FASTA file into Clustal_Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) and run a multiple sequence alignment with default options.

Download the alignment file and submit it to SWISS-Model Alignment Mode.

Make sure to select the chosen query sequence (obtained in Problem 1.) as your target sequence from the drop-down boxes on the next presented page. You have to choose a protein as your template sequence, as well. Choose the one that is the closet match according to FASTA35 in Problem 1. Results will be sent to you by SWISS-MODEL over email. They typically take no more than 30 minutes.

SWISS-MODEL also provides some analysis of the quality of the model it has obtained. It uses 3 programs to evaluate models: Anolea, Gromos, and QMEAN. Anolea, for instance, evaluates atomic empirical mean force potential and highlights for users areas of high and low energy scores.

Check the end of this document for what you need to submit as deliverables.

Problem 3. Threading-based Modeling of Structure for Chosen Target Sequence [20 pt]

You will now use the Phyre (**P**rotein **H**omology/analog**Y** **R**ecognition **E**ngine) webserver to obtain a model structure for the protein sequence you chose above in Problem 1. Phyre 2.0 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) is a threading-based modeling webserver with decent performance in CASP. The webserver has a limited set of features compared to the version used in the competition. However, it is also very easy to use, as all it requires is an email address, job identified, and a protein sequence, which you paste into the “Amino Acid Sequence” box.

Check the end of this document for what you need to submit as deliverables.

Problem 4. Model Comparison [20 pt]

Here you will compare the two different models you obtained through the comparative modeling and threading-based techniques in problems 3. and 4. above. Both SWISS-MODEL and Phyre come with a variety of different metrics for the model they produce. You will use here a standard measure to evaluate the quality of the two models and compare. You will use PDB Sum (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html) to obtain structural analysis data for each model. Submit the pdb file and provide an email address where you obtain this data.

You will use PROCHECK(http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/) to analyze Ramachandran angles of the obtained models. PROCHECK will create a Ramachandran plot and show you residues with favorable an unfavorable angles for each model. You have the option of downloading the plots as PDFs. PROCHECK notes that over 90% of amino acids should fall in the "most favored regions" in "good quality" models. You will use the information to determine from the obtained plots whether the SWISS-MODEL model or the Phyre one is better.

Check the end of this document for what you need to submit as deliverables.

Problem 5. Ab-initio Modeling [15 pt]

Repeat the steps in Problem 1. to identify a human protein with length between 100 and 150 amino acid

s whose sequence identity to any protein the PDB is no higher than 15%. Submit this sequence to the ROBETTA structure prediction server (http://robetta.bakerlab.org/). The software is also available, but I would suggest you use the webserver to obtain a model.
Check the end of this document for what you need to submit as deliverables.

Extra Credit: Modeling Dimeric Structure with RosettaDock [10 pt]

The D3 dopamine receptor structure has just been made available. Isolate a chain of this protein from its PDB fie. Create another file with the same chain, as you will use RosettaDock (http://rosettaserver.graylab.jhu.edu/docking/submit) to dock the two chains on top of each-other and obtain a dimeric structure.

Check the end of this document for what you need to submit as deliverables.

Deliverables:

Your deliverable for this homework can be a webpage with various links and info on it. Create a webpage under your personal gmu webpage, titled CS444_Hw4_Fall2012.

For problem 1: Under the specified webpage, provide links so I can see: uniprot_results.fasta (5 pts), *pdb_seqres.fast (5 pts),* and the fasta file of the sequence (10 pts) you end up choosing in problem 1. Provide another link (5 pts) that shows me the results of BLASTing the chosen sequence to all known proteins in the PDB. The link an be an image, so I can see the top entries returned by BLAST and their similarity to your chosen target sequence.

For problem 2: Provide links so I can see: the FASTA file obtained from Uniprot (2 pts), an image that shows the result of BLAST through Uniprot (2 pts), the alignment file obtained from Clustal_Omega (3 pts), an image that shows the results of running Clustal_Omega (2 pts), the PDB id of the chosen template (2 pts), an image that shows you selecting the right query sequence and the desired template protein in SWISS-MODEL (2 pts), a file that shows the coordinates of the model returned by SWISS-MODEL (3 pts), entire output file provided by SWISS-MODEL (2 pts), description of where are the regions of low energy scores highlighted by ANOLEA (2 pts).

For problem 3: Provide coordinates of the model (3 pts), an image that shows the model in 3-dimensions (2 pts), an image that shows the secondary structure prediction consensus used by Phyre (5 pts), a partial image of the results section where I can see the top 3 folds of the however many used for creating the model, together with the “Estimated Precision” percents (5 pts). What is the lowest percent in the folds used by Phyre to build the model (3 pts)? To which class of proteins do the majority of the used folds belong (2 pts)?

For problem 4: Provide a link with the results of PDB SUM for the SWISS-MODEL and Phyre models (2 pts), another link with the results of PROCHEK for each model (3 pts), PROCHECK Ramachandran Statistics plots for each model (5 pts), PROCHECK G-Factor Scores plot (5 pts), a description of what the plots show for each model, and whether you can draw a conclusion s to which model is higher

quality based on the shown statistics (5 pts).

For problem 5: Provide a link to the FASTA file of the Uniprot entry used as target sequence for ROBETTA (3 pts), a link to the that shows me the results of BLASTing the chosen sequence to all known proteins in the PDB (2 pts) (link an be an image, so I can see the top entries returned by BLAST and their similarity to your chosen target sequence), a link to the output provided by ROBETTA (5 pts), a link to the top model (3 pts), a description of whether the model is of good quality (2 pts).

*For extra credit:*Provide a link to the input pdb files you used for RosettaDock (2 pts), a link to the model returned by Rosetta (4 pts), an image of the dimeric structure in 3-dimensions (2 pts), and a description of whether the dimeric structure is of good quality (2 pts).

*If the results returned by a sever are in html format, print the page in pdf format and provide a link to that pdf from your website.