

# CHEM 498Q / CHEM 630Q: Molecular Modelling of Proteins

## TUTORIAL #1b: Comparison and validation of alignments

### INTRODUCTION

In the first part of Tutorial #1, you have successfully identified a protein sequence and have found a structural template you could use to model its structure.

In the second part, you will first learn how to predict the secondary structure of a sequence and use that information to increase your confidence in the validity of a sequence alignment. You will then learn how to use the PSI-BLAST algorithm to construct Position-Specific Scoring Matrices (PSSMs) and to perform more sensitive searches on sequence databases. You will use these PSI-BLAST searches to confirm your choice of a structural template.

### REQUIRED PRE-LAB READING

*Iterated Profile Searches with PSI-BLAST*

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-2.html>

### PRE-LAB REPORT

None.

### READING

#### On sequence alignment statistics:

*The Statistics of Sequence Similarity Scores*

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>

*The Statistics of PSI-BLAST Scores*

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html>

#### On protein structure:

Chapter 2 of Zvelebil & Baum (“Protein Structure”): All sections

### PROCEDURE

#### STEP 6: Predict the secondary structures of your query and template sequences

Using the PSIPRED server (<http://bioinf.cs.ucl.ac.uk/psipred>), predict the secondary structure of both your query sequence and the template sequence you have chosen.

- ◆ How do the predicted secondary structures of the two sequences compare? Compare the sequences according to the BLAST alignment obtained at STEP 4. Report any significant difference between the two—especially if they are predicted with a high level of confidence.
- ◆ For the template sequence, compare the PSIPRED results to the known secondary structure as reported in the PDB. How accurate are they? Where are the prediction errors located (if any)?

#### STEP 7: Visualize the PSSM of the conserved domain

Visualize the PSSM corresponding to the conserved domain you found at STEP 1 using NCBI's PSSM Viewer ([http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm\\_viewer.cgi](http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi)). Make sure you align your own protein sequence to the PSSM. (You can select either a “Stacked Bar View” where

the sequence is presented vertically, or a “Matrix View” where it is presented horizontally. Note that for the “Stacked Bar View” you may have to click on the “Next >>” button to see the entire length of the domain.)

◆ How is this PSSM related to the functionally important amino acids you found at STEP 2?

### **STEP 8: Perform a PSI-BLAST search for your query sequence**

Using the PSI-BLAST algorithm from the NCBI website (<http://blast.ncbi.nlm.nih.gov>), iterate a PSI-BLAST search on the “nr” database until no new sequences are found above the *E*-value threshold.

For the iterated PSI-BLAST search, use an *E*-value threshold of 1.0e-10 (instead of the default 0.005) and a maximum number of target sequences large enough that you get to see some sequences with *E*-values worse than the threshold.

◆ How many iterations are needed to converge the list of sequences? (If the list is still not converged after 10 iterations, use a threshold of 1.0e-20.)

Once the list of sequences above the threshold is “stable”, download the corresponding PSSM file. (Expand the “Download” section at the top of the results page, and click on the “PssmWithParameters” link. This will download a “scoremat” file on your computer.)

View this PSSM by loading the “scoremat” file you just downloaded into the PSSM Viewer.

◆ How does it compare to the PSSM from STEP 7?

### **STEP 9: Find homologous structures with lower sequence similarities**

Load the PSSM you saved in STEP 8 and perform a single PSI-BLAST iteration on the “pdb” database.

◆ Compare this list of high-scoring PDB sequences with the one from STEP 4. Where is the template sequence you chose in STEP 4 falling in this new list?

◆ Does the new list contain sequences that were not in the list from STEP 4? Would you consider using any of them as a structural template?

## **INSTRUCTIONS FOR THE LAB REPORT**

### **STEP 6**

Present the secondary structure diagrams for both sequences and identify how the two should be juxtaposed according to your BLAST alignment.

### **STEP 7 and STEP 8**

PSSMs can be fairly long and, although the PSSM Viewer offers many options, they can be a little overwhelming to look at. For the lab report, your primary concern should be to confirm that the functionally important residues you have identified from the literature correspond to high scores on the consensus sequences of the PSSMs.

### **STEP 9**

Whichever bit score you use to separate the “high-scoring” from the “low-scoring” PDB sequences, make sure your report all “high-scoring” sequences.