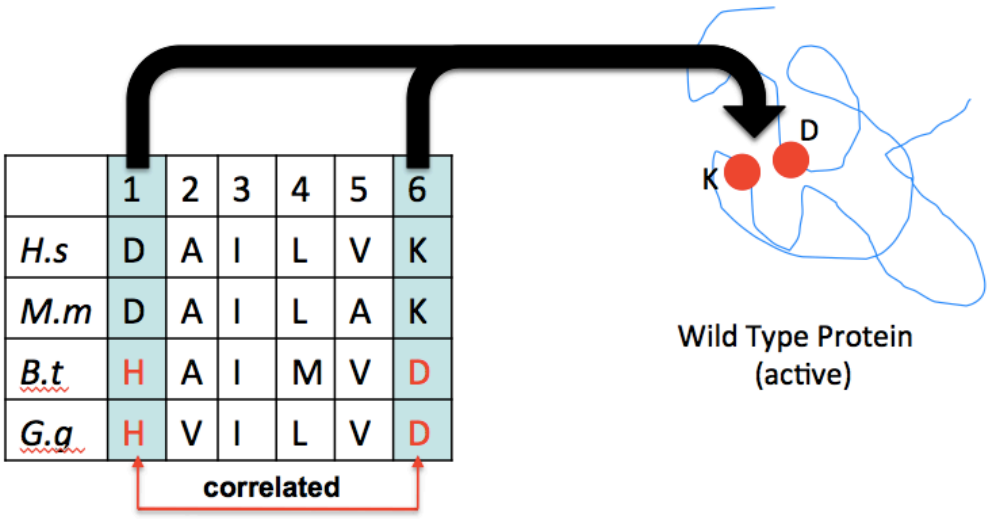
Structural Biology 2017
Practical 2

Learning models from protein sequences to predict protein structure.



March 13th 2017

Introduction

This tutorial is designed to guide you through carrying out a covariance analysis, using DCA (direct coupling analysis) for a protein family alignment that you have been provided with. To run the analysis, you will need to use MATLAB, which should be installed on your workstation. This tutorial will focus on the enzyme Dihydrofolate reductase, or DHFR.

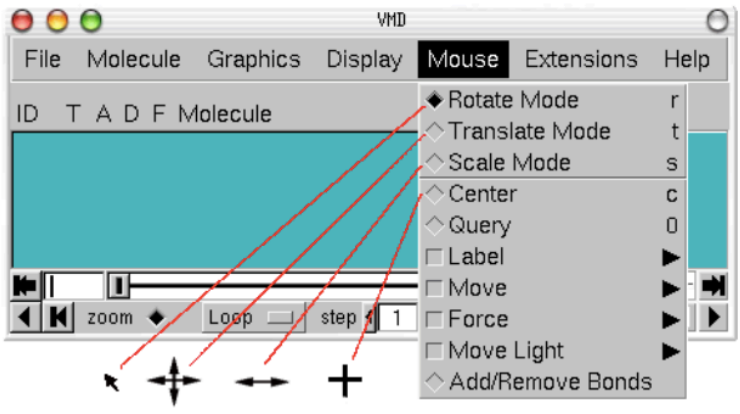
You should first read through the Wikipedia page on this protein: https://en.wikipedia.org/wiki/Dihydrofolate_reductase to learn a bit about how it works. Some information about drugs that target DHFR can be found here: http://pdb101.rcsb.org/motm/34 . DHFR is a key enzyme in folate metabolism. It catalyzes an essential reaction for de novo glycine and purine synthesis, and for DNA precursor synthesis.

Preliminaries

To run the code that you are provided with, you will need to have a working installation of matlab on the machine that you are using. To get started, go to the course moodle page and download the folder of files that has been provided for Practical 2. You should find that you have a sequence alignment file (this ends in .fas) which is a formatted text file that you can open in any text editing programme.

You will also find a few different matlab files, these end in '.m' and should be placed in your working directory. Inside matlab you can change the directory to wherever you choose – we will call this your 'working directory' for the rest of the tutorial. First make a folder in your working directory called 'OutputFiles'. Inside that, make another folder called 'DYR_ECOLI'. Place the sequence alignment file, and the files ending in .indextable and .indextableplus into this folder.

Next, download a file containing the coordinates of our protein of interest, in this case the DHFR protein. Go to the RCSB protein data bank website (http://www.rcsb.org) and download the relevant X-ray crystallographic structure PDB file, 1RX2. Recall that you may visualize the structure file using VMD, by typing '**vmd 1RX2.pdb**'.



To manipulate your view of the graphical image, you can use the Mouse menu, or more efficiently, the shortcut keys '**r**' (rotate), '**t**' (translate) or '**s**' (scale), followed by using the left mouse button to change the view. Explore the structure that you have downloaded, in tandem with the pdb101 article. Check that the first sequence in the alignment file matches the sequence of 1RX2. Then identify those sequence positions in DHFR that make contact with each of the ligands, and make an annotated version of the first sequence that indicates those sequence positions.

Next, open the **Representations** interface (from the main menu, **Graphics→Representations**), and change the Selected Atoms from '**all**' to '**protein**'. Choose '**New Cartoon**' as the Drawing Method and choose '**Chain**' for the Coloring Method. Follow the protocol that you learnt in the previous tutorial to investigate and visualize any non-protein molecules that may be present. Again following the protocol provided last time, visualize the secondary structure of the molecule in VMD.

Another view of a protein structure is provided by a protein contact map. This is basically a binary matrix that indicates whether each pair of sequence positions is 'in contact' or not. Read the Wikipedia page: https://en.wikipedia.org/wiki/Protein_contact_map to find out more about these objects. How do the different types of protein secondary structure appear in a protein contact map? Make some notes of your thoughts on this point - we will discuss this in class.

Covariance Analysis Basics

The goal of protein covariance analysis is to use the available sequence data to learn a model that includes pairwise interactions between positions in the protein sequence. However, there are a number of possible evolutionary pressures that might lead two sequence positions to coevolve – proximity in tertiary structure is just one of these. Think of other possible causes and make a list – we will discuss this during the session.

Next open matlab, and open the three main programme scripts that you have been provided with. These are

covariance_analysis.m

get_constraints_PC.m

test_constraints_PC.m

The structure of the code is that the first function, covariance_analysis.m, analyses the available sequence alignment to calculate the covariance between every pair of sequence positions in the alignment. Look at this piece of code. There are some brief comments to help you understand what the different sections of the code do. We will go through these together in class, to understand the basic structure of the analysis. You should feel free to add comments to your

copy of the code, to add a comment you simply start the line with a % sign. This tells matlab to ignore the contents of this line.

What are the different figures produced by this code? What do they tell you about the input sequence alignment?

The output of the first programme, 'Frob' is a matrix of pairwise scores, one for every pair of sequence positions in the alignment. Using the command included in the script file 'Script_mf_PC.m' that involves the 'covariance_analysis' function, run just this first piece of code to return the variable 'Frob'. Then open this variable in matlab. Can you tell anything from this data? What might you expect to be the case if the protein covariance analysis works? How might the analysis be confounded by other evolutionary pressures?

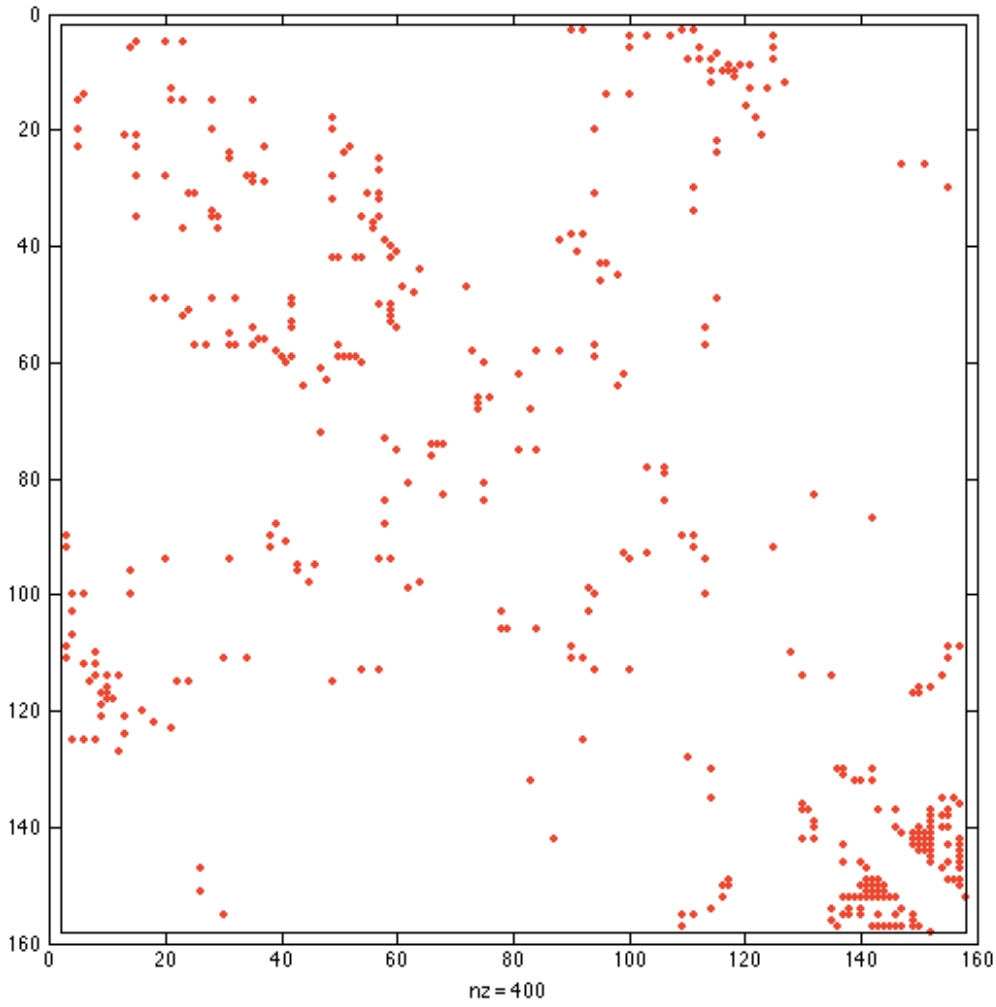
Translating between protein sequence and structure – numbering.

To test whether the statistical model learnt from covariance data provides any information about protein structure, we need to map the pairwise scores to the protein structure. Look again at the first sequence in the alignment, and the sequence in the protein structure file. Are these identical in terms of how they are numbered?

There can be significant differences between the numbering of the sequence, or sequence alignment, and the numbering used in the PDB for the protein structure. To account for this, we constructed the files that end in '.indextable' and '.indextableplus'. Open these two files in any text editor now. What do you notice about these files? What are the different columns? How do you think this might be helpful for the interpretation of the protein covariance analysis scores.

Now that we have the matrix of pairwise scores, and the table that provides the mapping between the sequence numbering and the structure numbering, we can translate the protein covariance scores into a set of predicted contacts in tertiary structure. This requires the second matlab function that you have been provided with, 'get_constraints.m'. We will go through this piece of code in class, to understand the different sections. If you read through the code you will notice that various pairs of sequence positions are removed from consideration before a set of predicted contacts is extracted from the sequence data. Think about why these might be extracted, we will discuss this in class.

Run the code for DHFR, using the command provided in the file 'Script_mf_PC.m'. What are the figures produced by matlab? You should see a figure like that shown below, what is this telling you? How can you adjust the parameters given in the command line to change what is shown on these figures?



Comparing the output to protein structure data

The final step of this analysis is to compare the predictions made by the covariance analysis to the experimentally derived coordinates of the 3D protein structure. You should recall from the first few lectures that a great deal of time and effort goes into the crystallization of each protein or protein complex. Although there are now pipelines for X-ray crystallography, many proteins are still highly challenging and conditions for successful crystallization may be difficult to identify and require the presence of other 'chaperone' proteins to increase stability.

To compare those pairs that are identified by the covariance analysis as strongly interacting with the set of structural contacts found experimentally, you will need to use the final piece of code, 'test_constraints_PC.m'. Try running this and then interpret the output figures and files that the code creates.