Andrew Torda
Björn Hansen

Universität Hamburg

Zentrum für Bioinformatik

Übung zur Vorlesung
Grundlagen der Strukturanalyse

Wintersemester 2011/2012

08.11.2011

Übung 2: Protein Classifications

1. Contents

2. Introduction and Aims

Classifying proteins is a popular pastime and the classifications are used for problems ranging from structure to function prediction. Sometimes it is well automated, but some of the classifications are largely built by human beings. Sometimes it is treated as a conventional clustering problem, but sometimes even the clusters are largely defined by humans. Here, we will look at some of the classifications which are very heavily based on human decisions.

The classifications to be used are

CATH: This defines 4 levels of hierarchy

Class, architecture, topology, homologous proteins

At the lowest level (homology) the members are very similar to each other and have easily detected sequence homology. This "H" level is often further divided up into different levels of sequence similarity (S35, S60, ...). Within a “topology” family, the proteins have a similar shape, but their sequences may be rather different. For example, the globin family includes haemoglobins, but also includes a domain from diphtheria toxin which has a very similar shape, but no obvious sequence or functional similarity.

SCOP: This usually defines 4 hierarchical levels, but uses different names

Class, fold, superfamily, family

As with CATH, proteins clustered together at the lowest level are sequence similar and usually have the same function. By the "fold" level, proteins will have a similar shape, but maybe no detectable sequence or functional similarity.

PFAM: is a classification based on sequences using hidden markov models. It does not impose a hierarchy on the proteins

Aims

- To understand the principles of structure classifications.
- To become familiar with the common classifications.
- To obtain and interpret the structural annotation for a protein.

3. Addresses

PDB	**www.rcsb.org**
SCOP	**scop.mrc-lmb.cam.ac.uk/scop**
CATH	**www.cathdb.info**
PFAM	**pfam.janelia.org**
Tools for Protein Structure Comparison	**cl.sdsc.edu/**

4. Tasks

a) Chimera: looking at the structures

Fetch the coordinates for

1ftt

Unfortunately, this coordinate file has 20 models (it was solved by NMR). The easiest way to look at this structure is to delete most of them:

Favourites

model panel

then select everything except the first model and click on "close" on the right hand side.

Fetch the second coordinates, for

1apl

To make the picture clearer, you might try

- select all coordinates (in the Model Panel)
- trace chains
- Actions -> ribbons -> show

To look for similarity (at the sequence level):
In the model panel

- Select both sets of coordinates
- match

In the "MatchMaker" dialog

- Select one protein on the left and the second protein on the right. You may play with the other options, but the defaults should work fine.
- apply

You should see the structural similarity between the proteins. You should also look in the panel "MultAlignViewer". Look under

- Tools
- Percent Identity

You should find an estimate of sequence similarity at the bottom of the MultAlignViewer.

A more interesting case.
Close the existing coordinates or restart chimera.
Load the coordinates for

1gxw
1i1i

To simplify the picture, select both coordinates, trace chains and show ribbons.
In the Model Panel, select both coordinates and then the "match" button. At this point, is there any point to think these coordinates are similar?
Note down the names of the files and whether they look at all similar to you.

On the web page *http://cl.sdsc.edu/*, calculate the structural alignment for two chains (follow the link “TWO CHAIN"). Enter the names (PDB-ID & chain identifier) of the coordinates: *1gxw:A* and *1i1i:P*. Click on the “Calculate Alignment” button. This will start a calculation to structurally align the two molecules. After some seconds, you should see a page describing the structural alignment. Follow the link “Download alignment as a PDB file” and save this file as any_name.pdb. This file contains the coordinates of *1gxw* (chain A) and *1i1i* (chain P) with the coordinates of one protein rotated and translated on to the other.

Go back to chimera, open the pdb file. Make sure you have a command line (Favourites -> command line). Colour the chains differently with a command like

color red :.P

color blue :.A

This superposition should look very different to the one that you originally made in chimera for *1gxw* and 1i1i. Try to follow the secondary structure elements (strands and helices) in the smaller structure and see if there is a corresponding element in the second structure.

Stop and make notes:

- Could you see similarity between *1i1i* and *1gxw* using chimera's built in sequence alignment?
- Using the superposition from SSAP (the file, super.pdb) was the similarity clear?

b) Homeodomain: 1ftt and 1apl

In both SCOP and CATH, search for the proteins *1ftt* and *1apl*.

In each classification, try to move around the classification tree (level) by going up from the leaf representing each protein.

Which classification level is common to both *1ftt* and *1apl*?

How many different homeodomain types are listed in SCOP? Now, compare against CATH. This database splits homeodomains into “sequence families”. How many are there and is the number comparable to the value from SCOP?

c) Immunoglobulins

Use this as a keyword to search both classifications. Both SCOP and CATH dedicate a branch of the classification tree to these domains. Try to match the terminology (family, superfamily, fold) that SCOP uses, to the names used for CATH's topology and homology levels.

d) Differences of opinion (*1gxw* and *1i1i*)

There is a group of proteins known as metalloproteases or neutral proteases. In both SCOP and CATH, find the entry for *1gxw*. Note the number of domains that each classification claims to find. Note down the broad hierarchical description (hydrolase, ß-roll, …). Where there is more than one domain, note down the domain boundaries.

What is the biggest difference between SCOP and CATH viewpoints?

For the classification which claims to recognise more than one domain, note down the boundaries (start and end residues) of the domains.

Given two different opinions, one may look for another point of view.

Visit the PFAM web site and look up 1gxw. How many domains does it find? Note down the domain boundaries.

Now, repeat the steps for 1i1i. You should have enough information to fill out a table like.

		SCOP	CATH	PFAM
1gxw	Number of domains Name type of each domain Boundaries of domains		note the CATH codes as well	
1i1i	Number of domains Name type of each domain Boundaries of domains		note the CATH codes as well	

5. Assignment

Please answer the following questions in a brief written report (1page), and hand it in not later than November 22, 2011 with your name.

- Which classification level is common to both *1ftt* and *1apl* in CATH and SCOP?
- For the immunoglobulins, describe the hierarchy in CATH and SCOP that leads to a protein such as *1bww*.
- Fill out a table comparing the *1i1i* and *1gxw* as outlined above.
- Note down whether you could see any similarity between the two proteins when using the initial sequence-based alignment (chimera) and the superposition based on structure.
- For the classifications which believe in more than one domain, are the domain boundaries similar?