Andrew Torda
Björn Hansen

Zentrum für Bioinformatik

Übung zur Vorlesung
Grundlagen der Strukturanalyse

Wintersemester 2013/2014

18.11.2013

Übung 4: 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** **v3-4.cathdb.info**
PFAM	**pfam.janelia.org**
Tools for Protein Structure Comparison	**cl.sdsc.edu**

4. Tasks

Chimera: looking at the structures

Start *chimera* with

```
> /usr/local/zbhtools/chimera-1.8.1/bin/chimera &
```

Afterwards, 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. For this purpose, open the **Model Panel** from the favourites menu and use the “group/ungroup” function to expand the listing to the individual submodels. Afterwards, select and close everything except the first model.

Close the Model Panel and fetch the second coordinates, for

1apl

The protein of interest is bound to a DNA double helix. To make the picture clearer, delete everything from this model except chain C.

- Select → Chain → C
- Select → Invert (selected models)
- Actions → Atoms/Bonds → delete

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

- select both sets of coordinates
- match

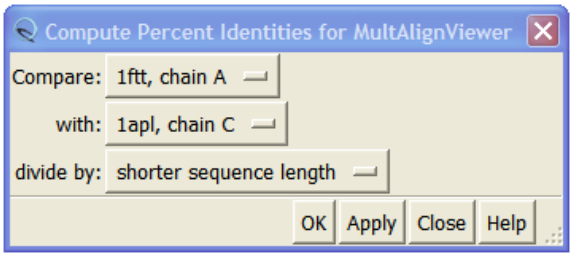
In the "MatchMaker" dialog

- select “Show pairwise alignment(s)” (we want to have a look at the sequence alignment as well)
- 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. Now let’s have a look at the similarity on the sequence level. The new window, which has just opened, is the the MultAlignViewer.
Look under

- Info → Percent Identity
- Press “OK”

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

At first, we want to calculate a structural superposition purely based on the sequence alignment without considering any structural information. In the Model Panel, select both coordinates and press the "match" button. In the MatchMaker menu, chose Smith-Waterman as alignment algorithm, deselect the option “Include secondary structure score (30%)” and press “Apply”. At this point, is there any point to think these coordinates are similar?

Now we want to include structural information for calculating the superposition of the two structures. Redo the structural alignment as described above, but this time select "Include secondary structure score (30%)". Can you see a difference?

To have a second opinion, we can use external tools to calculate the structural alignment. As an example, we will use the DaliLite server for pairwise comparisons. Go to

http://ekhidna.biocenter.helsinki.fi/dali_lite/start

There, you can specify PDB IDs (e.g. *1gxw* and *1i1i*) for the two structures you are trying to align. Optionally, you can specify certain chains you would like to compare. A few minutes after submitting your query, you are presented with a results page. In the section "Summary" you find a list of possible matches. Download the one with the highest Z-score[1] via the links in column "PDB". Give the downloaded file a proper name and add the suffix .pdb to indicate that it is a pdb file. This pdb file contains only the structure of your second molecule with the coordinates rotated and translated on the first one. Since the second molecule is placed on top of the first one, this is sufficient. In order to view the superimposed coordinates of the two proteins, you need to open both the original structure (*1gxw*) and the downloaded pdb file in *chimera*.

This superposition should look very different to the one that you originally made in *chimera* for *1gxw* and *1i1i* without considering the secondary structure information. 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.

5. Assignment

Please perform the given tasks and answer all questions in a brief written report. Bring this report with you on **December 2**, **2013**. The students to present the answers will be selected randomly.

a) Note down whether you could see any similarity between *1i1i* and *1gxw*

- using *chimera*'s sequence-based alignment?
- after enabling the option "Include secondary structure score (30%)" in the MatchMaker menu?
- using the superposition from DaliLite server?

Why is a structural superposition based on the sequence information alone rather unreliable for distantly related sequences. (6 P)

b) In both SCOP and CATH, search for the proteins *1ftt* and *1apl*.
In each classification, try to move around the classification tree by going up from the leaf representing each protein.
Which classification level is common to both *1ftt* and *1apl* in CATH and SCOP?
How many different homeodomain types are listed in SCOP? (6 P)

c) Use "immunoglobulin" as keyword for a search in SCOP and CATH. Both classifications dedicate a branch of the classification tree to these domains. Describe the hierarchy in CATH and SCOP that leads to a protein such as *1bww*. Try to match the terminology (family, superfamily, fold) that SCOP uses, to the names used for CATH's topology and homology levels. (6 P)

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, …). For the classification that claims to recognise more than one domain, note down the boundaries (start and end residues) of the domains. What is the biggest difference between SCOP and CATH viewpoints? 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? Again, note down the domain boundaries. For the classifications which believe in more than one domain, are the domain boundaries similar? Now, repeat the steps for *1i1i*. With the information you got, fill out the table below comparing *1i1i* and *1gxw*. (12 P)

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

[1] The Dali-Z score reflects the similarity between two molecules based on intramolecular distances. Structures that have significant similarities have a Z-score above 2, and usually have similar folds.

	Pfam			