

Exercises to the Lecture ``Protein Prediction``
Summer Term 2010

Sheet 4

General information

- There will be no exercise on Friday, June 4th
- Due date for this exercise sheet will be on Friday, June 11th

Exercise 7: Evolutionary robustness of protein structures (18 points)

Referring to the paper found at:

/mnt/home/rost/schaefer/exercise4/protein_robustness.pdf

- Explain in your own words: What does the author mean by close and remote structural homologues? How does this relate to the homology threshold established in the paper from the last exercise?
- What is meant by 'convergent' and 'divergent' evolution? For both cases, what would be the implication for sequence and structure similarity of two proteins?
- Try to explain, why we would expect a distribution of sequence identities between structural homologues as shown in figure 4.
- Try to discuss: How could the distribution have changed over time, starting at very early moments in evolution?
- In which ways does the expected distribution differ from the observed one (figure 3)?
- Explain in your own words: What possible reasons could be responsible for what we see in figure 3 in the left panel? How does the author explain the peaks found in the right panel?
- What does the author mean by 'anchor residues'? How many are there?

Exercise 8: List of Interfaces (15 points)

In Exercise 6 of the last sheet, you determined which PDB chain corresponds to which protein in

SwissProt. Here, we want to use this mapping to find all complexes containing chains from different proteins in the PDB. Two protein are considered different if they have different SwissProt identifiers.

Using the fasta file of all protein sequences in the PDB and the mapping described above:

- 1.) Find all PDB entries which have
 - a) at least two chains corresponding to different proteins.
 - b) at least one pair of different proteins where both chains are larger than 30 residues.
- 2.) Any pair of two different protein chains in a PDB file found in 1.) possibly has a common interface. List all the possible interfaces in a single file called interfaces.txt in the format also used in Sheet 2 Exercise 2 and sort the lines alphabetically.

Exercise 9: Pairwise Similarities (15 points)

To reduce redundancy within the complexes found in Exercise 8, we need to know the exact pairwise sequence identities between all similar chains. We will use Blast to calculate them.

Using the fasta file of all protein sequences in the PDB and the mapping described above:

- 1.) Collect all .f files corresponding to chains contained in the pdb entries found in Exercise 8 Task 1 and concatenate their content. This should create one big fasta file.
- 2.) Using this fasta file, build a binary blast database using formatdb (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html). Use "-i" as the only non-default parameter.
- 3.) Blast all the files of 1.) against the database of 2.) with the following command:

```
blastall -p blastp -i <PDB_ID>_<CHAIN_ID>.f -d <database> -o  
<PDB_ID>_<CHAIN_ID>.blast -m 8 -b <total number of chains in the database>  
-v <total number of chains in the database> -e 10e-3
```

Hint:

We only need blast output files of non-identical chains, as the output for identical chains will obviously be the same. Thus, in order to reduce computing time, you can first search for all identical chains in the set of sequences, reduce the set to only contain different chains, do step 3) and re-include the identical chains by copying and editing the output for the one chain which took part in step 3).

Exercise 10: PISA (15 points)

Complexes found in the PDB may not represent interactions as found in vivo. In this context, PISA (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) helps you identify the biologically functional units.

- 1.) Describe the major principles and functions of PISA. What is the input, what is the output?
- 2.) Write a tool which sends a given PDB entry to PISA and returns the set of the complexes found in the most likely set of assemblies in pdb file format. Do not apply the tool for the dataset created in Exercise 8 yet.