

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

Sheet III

General information

- Send an email (one per group!) to hamp@rostlab.org, schaefer@rostlab.org including the paths to your results (answers, scripts, datasets) until **Friday May 28 9:00 am**. Scripts should be executable for us so that we can reproduce your results!

Exercise 5: Significant alignments (12 points)

Referring to the paper found at:

</mnt/home/rost/schaefer/exercise3/hssp.pdf>

- a) What is the general aim the authors pursue in their paper?
- b) Give an example each for both cases where
 - two aligned protein sequences share low sequence similarity but high similarity in structure, and
 - two short similar sequences have very different structures.

What is the implication of that observation?

- c) Which three measures do the authors put into mutual relationship to get a safe structural homology threshold?
- d) Give a short explanation on how the authors determine the homology threshold. How do they measure structural similarity? Why do they use protein sequences from the PDB database for their sequence alignments?
- e) What are the two thresholds where two structures are thought to be *structurally identical*?
- f) Explain what you see in figure 4. What distinguishes the upper region from the lower one? Which conclusion can you draw from an alignment of length 110 with 20% sequence identity? What about an alignment length of 30 and sequence identity of 70%?

Exercise 6: PDB - Swissprot Mapping (20 points)

With this exercise, we want to start a new project. In the end we hopefully come up with an improved prediction method for protein-protein-interaction hotspots. The very first step to be taken is to find interactions between different protein chains within one PDB structure. Two protein chains are potential interesting interactions, if they map to different Swissprot identifiers (and therefore should be from different proteins).

- Split the fasta file found at `/mnt/home/rost/schaefer/exercise3/pdb.f` into multiple files so that one file contains exactly one chain sequence and is named `<PDB_ID>_<CHAIN_ID>.f` .
- Now, we do the actual mapping to the Swissprot database. For each `.f` file, execute the following command:

```
blastpgp -i <PDB_ID>_<CHAIN_ID>.f -d  
/mnt/opt/blast_db/sp/uniprot_sprot.fasta -o  
<PDB_ID>_<CHAIN_ID>.blast -m 8 -e 10e-10 -j 1
```
- The blast files created above are in tabular format:
The 2. column lists the identifier of the hit.
The 3. column represents sequence identity.
The 7. column gives the position of the start of the alignment in the query sequence.
The 8. column gives the position of the end of the alignment in the query sequence.

For each `.blast` file, determine the set of hits which fulfill both of the following criteria:

- 90% of the residues in the query sequence aligned
- 90% sequence identity

If there is no such hit, discard the whole chain.

Otherwise, take the one with the lowest e-value. This is the hit mapping the query PDB protein chain to the Swissprot entry representing the protein which codes for this PDB chain.

- For all `.blast` files having a Swissprot hit according to the criteria above, create a line in the following format:

```
<PDB_ID> <CHAIN_ID> <Swissprot ID>
```

Sort all lines alphabetically and store them in a single file named `pdb_sp_mapping.txt`

Hint: As you already know the blast runs take some time, especially since we are dealing with a huge amount of sequences to be tested.