

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

**Sheet II
Part 1**

General information

- New time for the exercise is Friday, **14:45 – 16:15 (s.t.)**
- There will be no exercise on Friday May 14, but exercise sheet II part 2 will be made available online on that day.
- Due date for the entire exercise sheet II will be on Friday May 21.
- Send an email (one per group!) to hampt@rostlab.org, schaefer@rostlab.org including the paths to your results (answers, scripts, datasets) until **Friday May 21 9:00 am**. Scripts should be executable for us so that we can reproduce your results!

Exercise 1: Redundancy Reduction of Protein Complexes (8 points)

Referring to the paper found at:
</mnt/home/rost/schaefer/exercise2/ppi.pdf>

- a) What is an Atomic Contact Vector (ACV)?
- b) How many dimensions does an ACV have?
- c) What does one dimension represent?
- d) Why was 6Å found to be a reasonable cut-off value for contacting residues?
- e) Give the major steps that led to the creation of the non-redundant set of 209 transient recognition complexes.

Exercise 2: Dataset creation (20 points)

In this exercise we want to consider the transient protein-protein interactions found in the paper of exercise 1. You will find the list of noncovalent transient protein complexes within the following file:

</mnt/home/rost/schaefer/exercise2/chains.txt>

where

- 1A4Y A:B means that within pdb structure 1A4Y chains A and B interact transiently.

- 1CIC AB:CD means that within structure 1CIC chains A and B may transiently interact with chains C and D, but there is no transient interaction between A and B or C and D.

Two residues X and Y of two transiently interacting protein chains are said to be part of an interface if at least one atom of X is within 6 Å (1Å = 10⁻¹⁰ m) distance to at least one other atom of residue Y.

Your task:

For each chain in chains.txt, determine the residues which are part of an interface in at least one transient interaction. Store your results in the following format:

- **One file per chain**
- **Filename: <PDB_ID>_<CHAIN_ID>.ref**
- **File content:**
 - **Line 1: Chain sequence in 1-Letter-Code**
 - **Line 2: Annotation of interface residues**

Example:

File 1XYZ_A.ref:

<Begin of file>

ASLIDUHEFUIHWEFUIHUEIHF

--P---PPP-----P---

<End of file>

You will find all pdb structures listed in chains.txt in /mnt/home/rost/schaefer/exercise2/pdb

We are specifically interested in the ATOM records which define the coordinates for each atom found in a residue (beside some other information). One ATOM record looks like this:

```
ATOM 70 CA SER L 10 -5.325 21.962 15.292 1.00 38.11 N
```

Of special interest are the following fields:

- **70** denotes a running number over all atoms found in the structure
- **CA** denotes the atom type, in this case (alpha) carbon
- **SER** denotes the 3-letter code for the residue (or amino acid name), in this case serine
- **L** denotes the chain identifier
- **10** denotes a running number over all residues within a chain
- **-5.325, 21.962, 15.292** denote the x, y, z coordinates of that atom within the structure (measured in Å)

A complete specification of how an ATOM record is defined could be found at

<http://www.wwpdb.org/documentation/format32/sect9.html>

Overall workflow:

- For each entry in chains.txt determine all possible transiently interacting chain pairs. E.g. entry 1CIC AB:CD gives you the chain pairs A-C, A-D, B-C, B-D
- For each such chain pair, determine the residues in each chain that are in contact with residues from the other chain (according to the definition above). For that, you will need the three-dimensional coordinates for each atom found in the specific pdb file.
- In the cases where a chain is involved in more than one transient interaction, only provide one file for this chain by merging all interface residues in one annotation.
- **Hint:** Write your pdb parser in a generic way such you can re-use it for arbitrary pdb chains, since we will need it for future exercises.