**Technische Universität München**
**Lehrstuhl XII: Bioinformatik**

**Prof. Burkhard Rost**
**Christian Schäfer**
**Tobias Hamp**

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

**Sheet II**
Part 2

## General information

- Due date for the entire exercise sheet II (both parts 1 and 2) 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 3: ISIS Original Paper (9 points)

Referring to the paper found at `/mnt/home/rost/schaefer/exercise2/isis.pdf`:

a) What exactly made the authors believe that the prediction of transiently interacting residues based on sequence alone could be basically possible?

b) Which specific features did the authors use for their prediction method? Why? How were these obtained?

c) Give a short wrap-up of the training of the method.

d) Where lies the potential in using not only evolutionary information but also predicted structural features? How could the ISIS predictions be improved even further?

e) The random prediction technique described in Section 2.3 accounted for both the number of predicted interface residues and heterogeneous distribution. Explain how not considering any one of these two criteria could bias the evaluation given in Figure 1.

## Exercise 4: Evaluating ISIS (20 points)

In this exercise you are supposed to use ISIS to predict the interface residues found in exercise 2 of exercise sheet II (part 1) and give a short evaluation.

*Generation of the ISIS input*

1.) For each `<PDB_ID>_<CHAIN_ID>.ref` file, create a `<PDB_ID>_<CHAIN_ID>.f` file in the following format:

Line 1: `'>' <PDB_ID> <CHAIN_ID>`
Line 2: `<Sequence of the chain>`

2.) For each .f file, execute the following blast query on one of the practical machines:
```
blastpgp -i <PDB_ID>_<CHAIN_ID>.f -d /mnt/opt/blast_db/big80/big_80 \
-j 3 -o <PDB_ID>_<CHAIN_ID>.blast -e 0.001 -h 0.001
```

3.) For each .f and .blast file, execute the following script:
```
/usr/share/librg-utils-perl/blast2saf.pl <PDB_ID>_<CHAIN_ID>.blast \
fasta=<PDB_ID>_<CHAIN_ID>.f saf=<PDB_ID>_<CHAIN_ID>.saf
```

4.) For each .saf file, execute the following script :
```
/usr/share/librg-utils-perl/copf.pl <PDB_ID>_<CHAIN_ID>.saf \
<PDB_ID>_<CHAIN_ID>.hssp
```

5.) For each .hssp file, execute PROFphd (it will create an .rdbProf file):
```
prof <PDB_ID>_<CHAIN_ID>.hssp best
```

Applying ISIS

1.) With the .f, .hssp and .rdbProf files for each chain as input, let ISIS predict the interface residues using default parameters (see `man profisis`). Store the result file of ISIS as `<PDB_ID>_<CHAIN_ID>.crd.isis`

2.) Repeat step 1.) but use the non-default option `--nocrd-restriction` and store the result file as `<PDB_ID>_<CHAIN_ID>.nocrd.isis`

3.) Convert all .isis files into the following format:
Line 1: `<Sequence of the chain>`
Line 2: `<Interface annotation>`
Line 3:
Line 4 to EOF: `<Complete original output file>`

*Evaluating ISIS*

1.) Calculate the following residue-based performance measures for all chains seperately using the `.crd.isis` files: Sensitivity, Selectivity, Positive Predictive Value (PPV), Negative Predictive Value (NPV)

2.) Sum up the TPs, FPs, TNs and FNs of all the chains and calculate the measures of step 1.)

3.) Repeat steps 1.) and 2.) with the `.nocrd.isis` files

4.) Extract Figure 1 from the ISIS Original Paper (see exercise 3) and mark the two points calculated in the two repetitions of step 2.)

*Definition note*
Positive Prediction: An interface residue according to the .(no)crd.isis file
Negative Prediction: An non-interface residue according to the .(no)crd.isis file
Positive Reference: An interface residue according to the .ref file
Negative Reference: An non-interface residue according to the .ref file

True Positives (TPs): All Positive Predictions which are also Positive References
True Negatives (TNs): All Negative Predictions which are also Negative References
False Positives (FPs): All Positive Predictions which are not Positive References
False Negatives (FNs): All Negative Predictions which are not Negative References