

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

Sheet V

General information

- Grading For successful participation (i.e. „Schein“) you will have to fulfill all of the following conditions:
 - At least 30% of available exercise points and at least 30% of available points in the final exam.
 - The overall points are made up by the achieved exercise points (weight 40%) and the exam points (weight 60%).
 - Pass: At least 40% of overall points; for achieving highest grade you do not have to gain 100% of points.
- Please send an email (one per group) to hampt@rostlab.org **and** schaefer@rostlab.org including the paths to your results (answers, scripts, datasets) until **Friday June 18 9:00 am for Exercise 11** and **Friday June 25 9:00 am for the programming part (Ex. 12 and 13)**. Scripts should be executable for us so that we can reproduce your results!

Exercise 11: Homology Modeling (20 points)

Referring to the lecture and the paper found at

/mnt/home/rost/schaefer/exercise5/homology_modeling.pdf:

- a) Explain in your own words the two assumptions that build the basis for homology modeling. What would be the relation to the hssp curve discussed in exercise 5?
- b) Where lies the limit in homology modeling? What could be alternative approaches for structure predictions in these scenarios?
- c) Explain the significance of the alignment correction step. In which way could multiple alignments and the knowledge of the template structure help?
- d) In your own words, what are the difficulties with loop modeling? What could be possible reasons for that and what approaches exist? Explain briefly what is meant by Ramachandran plot.
- e) What are rotamers? What is meant by combinatorial explosion in that case and what could be done to prevent it?
- f) Give a short wrap-up of the model optimization step. Specifically address the issue of force fields.
- g) Model validation is an important step in the whole process. Pick two possible aspects for model checking and explain them briefly.

Exercise 12: Application of PISA (10 points)

In this exercise, you apply the PISA tool you wrote in Exercise 10 to the set of complexes which contain chains from different proteins.

Apply PISA to each PDB ID in the list of possible interfaces created in Exercise 8 (Obviously only once per PDB ID). Use the list found at `/mnt/home/rost/hampt/interfaces.txt`.

For each input PDB ID, parse all the complexes in the most likely set of assemblies and find out which chains really physically interact, i.e. have at least one pair of interacting residues. Two residues are said to be interacting if at least one pair of their atoms is closer than 6Å.

Create a new list of interfaces which only contains these between physically interacting chains.

Use chain identifiers as output by PISA. In cases where two chains in one PISA structure have the same ID (e.g. in very large complexes), substitute the chain ID as given by PISA with another, so far unused, alphanumeric character (A-Z;a-z;0-9).

Exercise 13: Redundancy Reduction (20 points)

Now we want to reduce redundancy, i.e. filter out interactions which are too similar to each other to justify having both in the set.

Use the following files:

- `/mnt/home/rost/hampt/interfaces.txt`
- A reference set of blast files created in Exercise 9 of the last sheet (to be found at `/mnt/home/rost/hampt/blast/`)
- The reference PDB - SwissProt mapping file, available at `/mnt/home/rost/hampt/pdb_sp_mapping.txt`

- 1) Considering only chains in the list of interfaces: Find all the nodes (see *Definition Note*).
- 2) For each node, find all neighboring nodes.
- 3) For each node n: For all neighbors of n: As long as any pair of neighbors has an HVAL of greater than 30, remove one of the two interfaces with node n from the list of protein interfaces. (Algorithms preserving the highest amount of interfaces might get a bonus).
- 4) Save the final version of the list of interfaces.

Definition Note:

Node: *A set of protein chains which meets all of the following criteria:*

- 1) *It contains only chains having the same SwissProt ID*
- 2) *There is no chain outside the node which has the same SwissProt ID and a PIDE (see definitions below) of greater than 95% with a chain inside the node.*
- 3) *If the node has more than one chain: Each chain in the node shares more than 95% PIDE with at least one other chain in the node.*

Neighboring Nodes: *Considering nodes m and n: If any chain of m has an interface with a chain of n, the two nodes are neighbors.*

L(c, d): *The number of aligned residues in the best (highest ranked) blast alignment for chains c and d. (If there is no such alignment, consider c and d to have PIDE<95% and HVAL<30 in steps 1 and 3 above, respectively)*

PIDE(c, d): *Using the same blast alignment as L: The number of identical residues divided by L(c, d) times 100*

HVAL(c, d):

$$\text{HVAL} = \text{PIDE} - \begin{cases} 100 & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32\{1+\exp^{-L/1000}\}} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases}$$

(HVAL, PIDE and L all take chains c and d as parameters)

HVAL between two nodes: *The highest HVAL among all pairs of chains with both chains coming from different nodes.*