Assignment: GOR III secondary structure prediction, influence of protein family and using evolutionary information.

Information on course slides:

http://www.ulb.ac.be/di/map/tlenaert/Home_Tom_Lenaerts/INFO-F-439.html

Detailed discussion methodologies:

GOR III: Gibrat, J.F., Garnier, J., Robson, B. (1987) *J. Mol. Bio.* **198**, 425-443.

MCC : Matthews, B.W. (1975) *Biochim. Biophys. Acta* **405**, 442–451.

Introduction

The aim of this assignment is to give you insight in the data and concepts behind secondary structure prediction. You will implement the GOR III algorithm, apply the 'jack-knife'/'leave-one-out' for internal validation, investigate the influence of using evolutionary information, and finally interpret the results.

1. The provided **starting data set** is the secondary structure data per residue for a set of 498 proteins (single chains from the Protein Data Bank (PDB)). Two sets will be available, one as determined by STRIDE (*stride_info.csv*), the other by DSSP (*dssp_info.csv*). You also receive data on the CATH protein family each of these proteins belongs to (*cath_info.csv*).
2. You implement the GOR III algorithm.
3. You apply this algorithm to predict the secondary structure for all the proteins in the starting data set for both the STRIDE and DSSP sets. The protein you examine cannot be in the data you used to parameterise the GOR III algorithm, however, so you have to apply a 'jack-knife'/'leave-one-out' algorithm.
4. You compare the results from STRIDE and from DSSP with the Q_3 and MCC quality scores to look at the variation in the success of prediction, both overall (the whole set) and by protein family (subset per actual CATH protein family).
5. You use the secondary structure prediction to predict the protein family. You have to determine your own criteria for this step, and compare against the actual protein family as determined by CATH.
6. The final part of this assignment is to explore the improvement in the GOR III approach by combining it with a sequence alignment search from UniProt. This should improve the reliability of the prediction, as described for GOR V and many other secondary structure prediction methods.

Details

1. The STRIDE (*stride_info.csv*), and DSSP (*dssp_info.csv*) files are tab delimited and contain the following information per column:

PDB_code PDB_chain_code PDB_seq_code residue_name secondary_structure

Only use the data if *residue_name* is one of the 20 natural amino acids. The *secondary_structure* field is **Helix**, **Beta**, **Other** or **Coil**. For the purposes of this assignment, classify residues with **Other** secondary structure as **Coil**.

The file with the protein family information (*cath_info.csv*) is also tab delimited and contains the following information per column:

PDB_code PDB_chain_code protein_family

Where protein_family is **Alpha** (helical), **Beta** (sheet), **Alpha/beta** (mix of both) or **None**. There will be small inconsistencies in the input data, which you will have to deal with.

2. In the original GOR III implementation the authors used 'dummy' frequencies because they did not have enough data. The dataset provided here is much larger so you will not have to do this; just implement with the data as is.

   The short one-letter codes to be used for the secondary structure are H for alpha helix, E for beta sheet and C for coil.

3. Efficiency is necessary for this step; think about how to do this 'leave-one-out' as quickly as possible without having to recount all data.

4. The formulas for the Q_3 and MCC scores are:

$$Q_3 = \frac{N_{residues\_correctly\_predicted}}{N_{residues\_total}}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5. In this step you can show some creativity and understanding; what criteria would you use to determine the protein family from the secondary structure prediction?

6. The basic procedure is that you get a set of similar sequences from Uniprot, run your GOR III predictor on each of them, and then combine the results as described in the last lesson. You should check the 6 protein sequences below; first the amino acid sequence is given, then in green the corresponding secondary structure assignment (for validation purposes).

   A. PDB code 1arl, alpha/beta

```
>1arl_A; molId:1; molType:protein; unp:P00730; molName:APO-CARBOXYPEPTID...
ARSTNTFNYATYHTLDEIYDFMDLLVAEHPQLVSKLQIGRSYEGRPIYVLKFSTGGSNRPAIWIDLGIHSREWIT
```

```
QATGVWFAKKFTEDYGQDPSFTAILDSMDIFLEIVTNPDGFAFTHSQNRLWRKTRSVTSSSLCVGVDANRNWDAG
FGKAGASSSPCSETYHGKYANSEVEVKSIVDFVKDHGNFKAFLSIHSYSQLLLYPYGYTTQSIPDKTELNQVAKS
AVAALKSLYGTSYKYGSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPTAQETWLGVLTI
MEHTVNN
```

```
>1arl A
CCCCCCCCCCCCCCHHHHHHHHHHHHHHHCCCCEEEEEEEECCCCCEEEEEEECCCCCCCC
EEEEEECCCCCCHHHHHHHHHHHHHHHHHCCCCHHHHHHHHCEEEEECCCCHHHHHHHH
HCCCCCCCCCCCCCCCCCCCCCHHHCCCCCCCCCCCCECCCCCCCECCCCCCCCHHHHHHH
HHHHHHCCEEEEEEEEECCCEEEECCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCEE
EEHHHHCCCCCCCHHHHHHHCCCCEEEEEEECCCCCCHHHCCHHHHHHHHHHHHHHHHH
HHHHHHC
```

B. PDB code 1ava, chain C, all beta

```
>1ava_C; molId:2; molType:protein; unp:P07596; molName:BARLEY ALPHA-AMYL...
ADPPPVHDTDGHELRADANYYVLSANRAHGGGLTMAPGHGRHCPLFVSQDPNGQHDGFPVRITPYGVAPSDKIIR
LSTDVRISFRAYTTCLQSTEWHIDSELAAGRRHVITGPVKDPSPSGRENAFRIEKYSGAEVHEYKLMSCGDWCQD
LGVFRDLKGGAWFLGATEPYHVVVFKKAPPA
```

```
>1ava C
CCCCECECCCCCECECCCEEEEEECCHHHCCCEEEEEECCEEEEEEEECCCCCCCCCCE
EEEECCCCCCCCECECCCCEEEEECCCCCCCCCCCECEECCCCECCECEEECCCCCCCCCC
CHHHCEEEEECECCCCCCEEEEEECCCEEECEEECCCCCCCCCEEECCCCECCEEEEECC
C
```

C. PDB code 1avm, alpha + beta

```
>1avm_A; molId:1; molType:protein; unp:P80293; molName:SUPEROXIDE DISMUT...
AVYTLPELPYDYSALEPYISGEIMELHHDKHHKAYVDGANTALDKLAEARDKADFGAINKLEKDLAFNLAGHVNH
SVFWKNMAPKGSAPERPTDELGAAIDEFFGSFDNMKAQFTAAATGIQGSGWASLVWDPLGKRINTLQFYDHQNNL
PAGSIPLLQLDMWEHAFYLQYKNVKGDYVKSWWNVVNWDDVALRFSEARVA
```

```
>1avm A
CCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCHHHHHHHHHHHHHHHHHHHHHHCCCCCHHH
HHHHHHHHHHHHHHHHHHHHCECCCCCCCCCCCCHHHHHHHHHHHCCHHHHHHHHHHHHC
CCCCEEEEEEEECCCCEEEEEEEECCCECCCCCCEEEEEEECCHHHCHHHHCCCHHHHHH
HHHHHECHHHHHHHHHHCCC
```

D. PDB code 1hge, chain B, coiled coil

```
>1hge_B; molId:2; molType:protein; unp:P03438; molName:HEMAGGLUTININ, (G...
GLFGAIAGFIENGWEGMIDGWYGFRHQNSEGTGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEVEG
RIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQLRENAEEMGNGCFKIYHKCDNACIE
SIRNGTYDHDVYRDEALNNRFQIKG
```

```
>1hge B
CCCCCECCCECCCECCCCCCCEEEEEEECCEEEEEEEHHHHHHHHHHHHHHHHHHHHCCCC
EECCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHEEECCCCEEEECCCCCHHHHHHHHCCCCCCHHHHHHHHHHHCCCCC
```

E. PDB 1hmo, all alpha

```
>1hmo_A; molId:1; molType:protein; unp:P02246; molName:HEMERYTHRIN;
GFPIPDPYCWDISFRTFYTIIDDEHKTLFNGILLLSQADNADHLNELRRCTGKHFLNEQQLMQSSQYAGYAEHKK
AHDDFIHKLDTWDGDVTYAKNWLVNHIKTIDFKYRGKI
```

```
>1hmo A
CCCCCCCCCCCHHHCCCCHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHH
HHHHCCCCCHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHCHHHHCCCC
```

F. PDB 1jsu, only secondary structure when it binds a partner

```
>1jsu_C; molId:3; molType:protein; unp:P46527; molName:P27;
HPKPSACRNLFGPVDHEELTRDLEKHCRDMEEASQRKWNFDFQNHKPLEGKYEWQEVEKGSLPEFYYRPPRPPKG
ACKVPAQES
```

```
>1jsu C
CCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCHHHHHHHCEECCCCEECCCCCCCCEEEECCCC
CHHHHCCCCCCCCCCCCCCCC
```

You do **not** have to implement a procedure to do a multiple sequence alignment (MSA) to get sequences to compare to. You can just go to UniProt to download the alignment and start from there: I will describe that procedure here, but you may choose a different one to get your MSA.

1. Go to http://www.uniprot.org/
2. Click on the top left '**Blast**' tab
3. Enter your one-letter amino acid sequence in the '**Sequence or UniProt identifier**' box
4. Press the '**Blast**' button next to this box
5. Wait for the results…
6. After the results show up, there will be an orange '**Download**' button on the right-hand side of the browser screen, just under the query area. Click this button.
7. In the '**List**' box you can now click '**Download**' to get the list of UNIPROT IDs that match this sequence.
8. Click on the '**Align'** tab at the top.
9. Enter the list of UniProt IDs in the '**Sequence or UniProt identifier**' box.
10. Press the '**Align** button next to this box
11. Wait for the results…
12. You can now click the orange boxes on the right to get the multiple sequence alignment.

Evaluation

You should implement your work on the Jupyter platform, and comment your code and analysis results. On there we should find:

1. The code, the secondary structure prediction per protein (as done from the DSSP and STRIDE data), with the Q3 and MCC scores for each, and your protein family prediction. An example line in this output file would be:

   9xyz CCHHHHHHCCEEEECCEEEEECCCHHHH 67.3 0.523 Alpha/beta

   with the first column the PDB code, the second the prediction, the third the Q3 score, the fourth the MCC score, and the fifth your protein family prediction.

2. An analysis and report on the implementation of the GOR III and the 'leave-one-out' approach, and a discussion of the results from steps 4, 5 and 6. For step 6 this report should describe the Q3 and per-secondary structure MCC quality indicators for all 6 proteins, with only GOR III (or your improved version of it) and with the combined GOR III/sequence alignment method. Use graphs to clarify your results, and indicate distribution ranges where appropriate!

Things to remember

The project is individual work. All plagiarism, copying or fraud will result in disciplinary actions.