1st Assignment: Secondary structure prediction, influence of protein family and using evolutionary information.

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** and apply the 'jack-knife'/'leave-one-out' for internal cross-validation. If you have already implement GOR III before, you may implement **another prediction approach** (*e.g.* an SVM) with 10-fold cross-validation. You will then investigate the influence of using evolutionary information on the prediction performance, and interpret the results.

1. The provided **starting data set** is the secondary structure data per residue for a set of 494 proteins (single chains from the Protein Data Bank (PDB)). Three data sets are available:

    a. The per-residue secondary structure from STRIDE (*stride_info.txt*)

    b. The per-residue secondary structure from DSSP (*dssp_info.txt*).

    c. The per-protein CATH protein family (*cath_info.txt*).

2. You implement the GOR III algorithm (or another approach), separately trained on the STRIDE and DSSP per-residue secondary structure data sets.

3. To assess the performance of your approach, you apply a 'jack-knife'/'leave-one-out' approach for GOR III (a 10-fold cross validation is also acceptable for other methods). This means that you take the protein(s) you want to predict out of your training data set, parameterise your prediction algorithm on this reduced dataset, and then predict the secondary structure for the protein(s) you removed.

4. You compare the results from STRIDE and from DSSP with the Q_3 and MCC quality scores (see below) to look at the variation in the prediction performance, both overall (for the whole set) and per protein family (make subsets per CATH protein family).

5. You then predict the protein family of each sequence, based on the secondary structure prediction results you obtained for each protein. To do so, you can determine your own criteria and approach, and then assess the performance of your method on the actual protein family as determined by CATH.

6. The final part of this assignment is **not required** if you run out of time, but is interesting to perform, and will give you 'bonus' points. The task 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.

Additional details for steps above

1. The STRIDE (*stride_info.txt*), and DSSP (*dssp_info.txt*) 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 (PDB_code, PDB_chain_code) combination is the unique identifier for each protein.

   The file with the protein family information (*cath_info.txt*) 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.

Note that amino acids can be indicated by their three-letter code (*e.g.* ALA) or by their one-letter code (*e.g.* A). You can find mappings between these online, or via Biopython.

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. The general approach you should take is to first count the frequencies for all proteins, keep track of the values for the individual proteins while you do so, and then subtract the values from the protein you are predicting from the total values (but make sure you always start from the original total values!). You then only have to recalculate the log scores.

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

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

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

Make sure you understand correctly what the meaning is of a True/False Positive and a True/False Negative! See the slides from the course.

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 for this step 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 lesson – basically the 'highest count' for the predicted secondary structure per column in the multiple sequence alignment. You should check **at least two** of the 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

CCCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCEEEEEEEECCCCCEEEEEEECCCCCCCC

EEEEEECCCCCCHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHCEEEEECCCCHHHHHHHH

HCCCCCCCCCCCCCCCCCCCCCHHHCCCCCCCCCCCECCCCCCCECCCCCCCCHHHHHHH

HHHHHHCCEEEEEEEEECCCEEEECCCCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCEE

EEHHHHCCCCCCCHHHHHHHCCCCEEEEEEECCCCCCHHHCCHHHHHHHHHHHHHHHHHH

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

CCCCECECCCCCECECCCEEEEEECCHHHCCCEEEEEECCEEEEEEEEECCCCCCCCCCE

EEEECCCCCCCCECECCCCEEEEECCCCCCCCCCECEECCCCECCECEEECCCCCCCCCC

CHHHCEEEEECECCCCCCEEEEEECCCEEECEEECCCCCCCCEEECCCCECCEEEEEECC

C
```

C. PDB code 1avm, alpha + beta

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

```
>1avm A

CCCCCCCCCCCCCCCCCCCCHHHHHHHHHCHHHHHHHHHHHHHHHHHHHHHHCCCCCHHH

HHHHHHHHHHHHHHHHHHHHHCECCCCCCCCCHHHHHHHHHHHCCHHHHHHHHHHHHHHC

CCCCEEEEEEEECCCCEEEEEEEECCCECCCCCCEEEEEEECCHHHCHHHHCCCHHHHHH

HHHHHECHHHHHHHHHHHCCC
```

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

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

```
>1hge B

CCCCCECCCECCCECCCCCCCEEEEEEECCEEEEEEEHHHHHHHHHHHHHHHHHHHCCCC

EECCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

HHHHHHHHHEEECCCCEEEECCCCCHHHHHHHHCCCCCCHHHHHHHHHHHCCCCC
```

E. PDB 1hmo, all alpha

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

```
>1hmo A
CCCCCCCCCCCHHHCCCCHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHH
HHHHCCCCCHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHCHHHHCCCC
```

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

CCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCHHHHHHHCEECCCCEECCCCCCCEEEECCCC

CHHHHCCCCCCCCCCCCCCCCC
```

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. Set the number of Hits to 100.
5. Press the '**Run blast**' button below the box

6. Wait for the results…
7. After the results show up, there will be a table of sequences under the Alignment header. Click the top left box next to 'Alignments' title. This will select all the sequences that were found by BLAST.
8. Now click on the '**Align**' button just below the 'Alignments' title.
9. Wait for the results…
10. You can now download the multiple sequence alignment (MSA) for the proteins with the 'Download' button above the 'Alignment' title.

Note that, in order to predict each sequence, you have to first remove the gaps from the sequence, then predict the secondary structure, and finally map the predictions back to the MSA. Then you count the predominant secondary structure code for each non-gap column in the original sequence!

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   A    CCHHHHHHCCEEEECCEEEEECCCHHHH    67.3    0.523   Alpha/beta
   ```

   with the first column the PDB code, the second the chain code, the third the prediction, the fourth the Q3 score, the fifth the MCC score, and the sixth 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 2 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.