Algorithms in Structural Bioinformatics
Secondary Structure Prediction Practical

TME-3

In this practical you will consider the following sequence:
`AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRLSIKEATEQSQPAA`

This sequence was one of the target sequences in the second *Critical Assessment of Techniques for Protein Structure Prediction (CASP) meeting* (`http://predictioncenter.org/casp2/`). You will try and predict the secondary structure of the corresponding protein.

Exercice - 1 Predictions based on a single sequence

In a first attempt you will rely on two secondary structure prediction methods that make use of one single query sequence only.

1- Use the *GOR* method version IV (for example from the *Network Protein Sequence @nalysis web server NPS@* `http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html`) to predict the secondary structure of the above sequence.

a- What percentages of the different secondary structure types are predicted?
b- On the basis of this prediction, in which of the CATH classes would you place the protein?

2- Use the *Predator* method (for example from the *NPS@* web server) to predict the secondary structure of the same sequence.

a- What percentages of the different secondary structure types are predicted?
b- On the basis of this prediction, in which of the CATH classes would you place the protein?

Exercice - 2 Predictions based on multiple sequence alignment (MSAs)

You will now analyse the results of more sophisticated methods that enrich their prediction by considering sequences homologous to the query sequence.

1- Go look at the predictions of the *Jpred* method obtained during *CASP2* experiment (`http://xray.bmc.uu.se/gerard/embo2001/predic/t0004.html`). The top part of the page shows the multiple-sequence alignment used in the prediction. At the very top is our mystery sequence, followed by related sequences found in the OWL sequence database. Each sequence is identified by its SWISS-PROT code. The residues highlighted in red are those that are strictly conserved across the alignment. At the bottom is given a simple conservation score (labelled "consv") for each residue position.

At the bottom of the page, below the sequence alignment, are the secondary structure predictions from the various methods used by *Jpred*:

Jpred name	Full name	Description	Reference
dsc	DSC	Multiply-aligned homologous sequences	King & Sternberg (1996)
mul	MULPRED	Combination of single sequence methods	Geoff Barton (unpublished)
nnssp	NNSSP	Combination of nearest-neighbour & MSAs	Salamov & Solovyev (1995)
phd	PHD	MSA input to a neural network	Rost & Sander (1994)
pred	PREDATOR	Database-derived statistics	Frishman & Argos (1996)
zpred	ZPRED	MSAs & GOR decision constants	Zvelebil et al. (1987)

The line labelled "cons" gives the consensus prediction derived from all the methods used, weighted primarily by the *PHD* prediction which tends to be the most accurate of all the methods.

The line labelled "PHD Rel" gives *PHD*'s relative level of confidence in its prediction. Regions associated with scores of 7-9 represent the highly-confident regions, most likely to have been correctly predicted.

a- Which CATH class does the consensus prediction suggest?
b- How much of the protein is predicted as helix?
c- How many regions of regular secondary structure might you be quite sure are correct?
d- What percentage of the sequence has been confidently predicted?

Exercice - 3 **Solvent accessibility**

The *Jpred* method has also predicted the solvent accessibility of each residue. This information is given in the line labelled "access": residues are predicted as being buried (B) or exposed (E) in the protein structure. Of particular interest are α-helices or β-strands which are either amphipathic or totally buried.

An amphipathic helix is one which, because it is on the surface of the protein, has one side consisting largely of hydrophobic residues which face the protein's hydrophobic core, and the opposite side consisting largely of polar residues which face out into the solvent. As α-helices have a periodicity of 3.6 residues per turn, the pattern of buried (B) residues will be of the form: i, i+3, i+4, i+7. Such a pattern would suggest a surface helix.

Similarly, an amphipathic strand has one side hydrophobic and the other polar. The geometry of strand residues means that the pattern of buried (B) and exposed (E) residues is a simple alternating one. In this case the pattern suggests the strand is the "edge" strand of a β-sheet, and pokes out into the solvent.

β-strands in alpha/beta proteins are often completely buried. These can be identified by a run of hydrophobic residues.

1- Can you identify any of the above features in the *Jpred* accessibility predictions for our mystery protein?

Exercice - 4 **Secondary structure assignment**

The protein you have been predicting the secondary structure is the **S1 motif of polyribonucleotide nucleotidyltransferase from** *E. coli*. The 3-dimensional structure of this protein was solved using NMR (PDB code 1SRO). Based on this tertiary structure, one can assign the secondary structure of the protein, for example using the program *DSSP* which relies on hydrogen-bonding geometry.

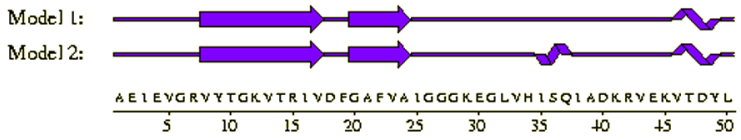
Here is the secondary structure of the protein, as calculated by *DSSP*:

```
         1         2         3         4         5         6         7
1234567890123456789012345678901234567890123456789012345678901234567890123456
.........:.........:.........:.........:.........:.........:.........:......
AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRLSIKEA
-------EEEEEEEEEE--EEEEE--------------------HHHH-----EEEEEEEE-------EEEE---
```

The first thing to note is that the sequence above is 8 residues shorter than the original sequence. Presumably the experimental data were insufficient to uniquely define the structure of the C-terminal end of the protein. Consequently you will merely ignore these 8 residues when assessing the accuracy of the prediction methods.

1- The second thing to note is that the structure was determined using NMR. To check the quality of the structure go the corresponding PDBsum page (`http://www.ebi.ac.uk/pdbsum/1sro`). What percentage of residues are in the "core" regions of the Ramachandran plot according to *procheck*? What does this tell you about the quality of the model?

2- As one can observe the 1SRO entry contains an ensemble of 20 (slightly different) models and, within the ensemble, there is some disagreement as to the secondary structure of the protein. For example, here are the first 50 residues of models 1 and 2 in the ensemble:



How can you interpret this difference?

3- Visualise the structure with *Jmol* on the PDBsum page, or with *Pymol*. Look carefully at the structure. You can observe that there is a loop which looks like it should be part of a β-sheet yet presumably lacks the appropriate H-bonding patterns. This could be due to the relatively poor quality of the NMR model. Which range of residues looks like it might belong to a "missed" strand?

Exercice - 5 Prediction accuracy

To evaluate the quality of the secondary structure predictions you have performed, you will assume that the *DSSP* assignement is correct.

1- Look again at the PDBsum page for 1SRO and note down the protein's actual fold (as given by its CATH classification). What is the protein's fold type?

2- Comparing *Jpred* predictions with the correct secondary structure, see how well the regions predicted with "high-confidence" fare. How well do you think the *GOR*, *Predator* and *Jpred* predictions have done overall? How useful are they? Any other observations?