## Secondary structure prediction

1. What are the differences between secondary and tertiary structure?

   Secondary structure refers to arrangements of peptide chain such as a-helices/beta-sheets and more. These structures are based on physical properties of aminoacid motifs and interactions between their hydrogen molecules, orientation of sidechains and other.
   Tertiary structure describes a higher order interactions within and between secondary motifs. Interactions such as hydrophobicity, disulphide bonds contribute to tertiary structure formaition.

2. Why do researchers want to predict protein secondary structures from sequences?

   3D prediction is harder computationally and more uncertain. Knowing at least secondary structure might provide some information about function (catalytic sites) and location, by comparing to 2D structures and folds of proteins with known 3D structures.

   Additionally, rate of generation of new sequence data outpaces the rate at which new 3D structures can be determined, while educated guesses may be made based on numerous existing folds.

3. What is the Swissprot accession number of the sequence?

   P02144

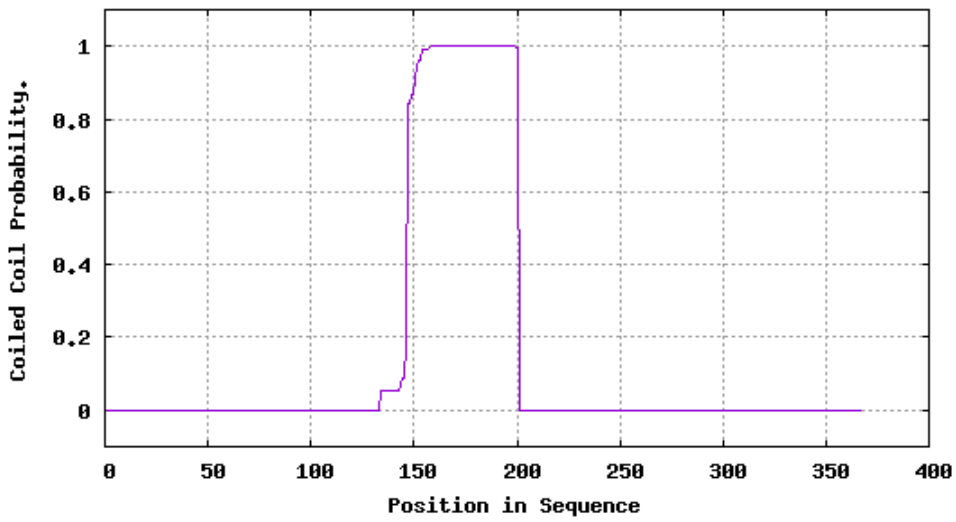
4. How many helices longer than five residues are predicted by the method?

   7

5. Find the myoglobin (human) protein structure in pdb. How many alpha helices does it have?

   8

6.

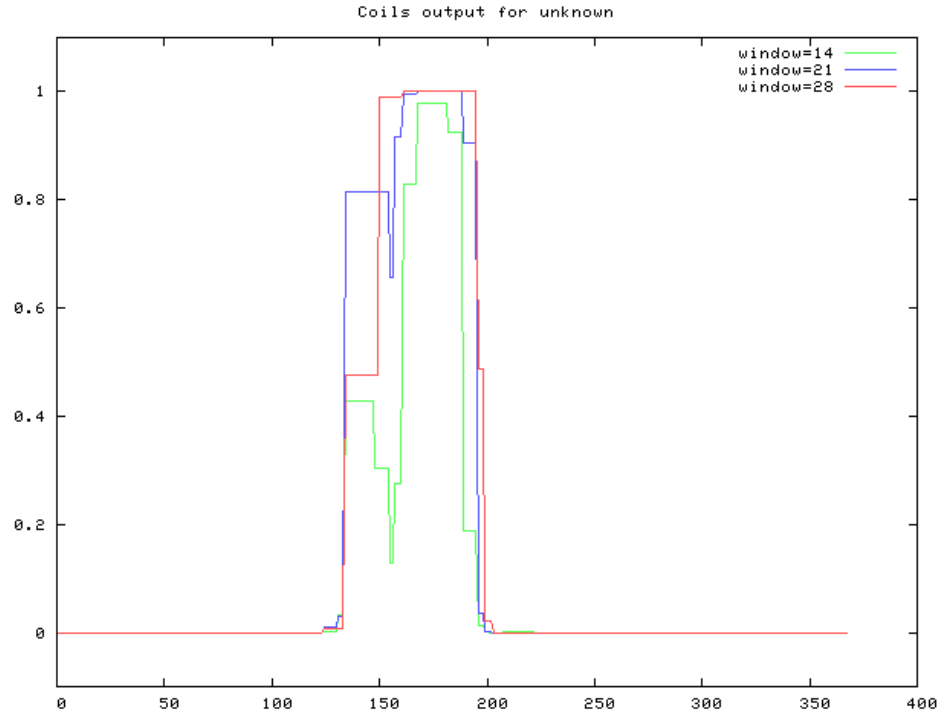



*MARCOIL*

6. Submit the sequence of chicken c-fos to MARCOIL and COILS. Compare the predictions and describe the results you get?

Marcoil output:

```
 NUMBER PREDICTED COILED-COIL DOMAINS WITH THRESHOLD 1.0 : 1
      1. from 134 to 200 (length = 67) with max = 100.0
NUMBER PREDICTED COILED-COIL DOMAINS WITH THRESHOLD 10.0 : 1
      1. from 146 to 200 (length = 55) with max = 100.0
NUMBER PREDICTED COILED-COIL DOMAINS WITH THRESHOLD 50.0 : 1
      1. from 147 to 200 (length = 54) with max = 100.0
NUMBER PREDICTED COILED-COIL DOMAINS WITH THRESHOLD 90.0 : 1
      1. from 151 to 200 (length = 50) with max = 100.0
NUMBER PREDICTED COILED-COIL DOMAINS WITH THRESHOLD 99.0 : 1
      1. from 154 to 200 (length = 47) with max = 100.0
```




*Coils prediction*

MARCOIL vs COILS:

Both prediction methods are based on MTIDK and MTK protein databases, weighted on frequency of aminoacids in coils within those protein families.

MTK: Myosin, Tropomyosin, Keratin
MTIDK: Myosin, Tropomyosin, Kinesin, Intermediate Filament, Desmosomal Proteins

COILS predict based on matrices constructed for heptad (7 residue) based reading windows. The 7 residue windows for coiled coils is important in this case, as they represent the core defining motif of coiled coils: a, d residues are hydrophobic and buried close to their hydrophobic neighbours(forming the superhelix), while others are exposed and hydrophilic. After 7 residues, the motif can repeat with the same position and orientation of starting residue.

Both matrices have strengths and weaknesses in identifying coiled segments depending on overall structure of the protein: MTK is more specific for identifying proteins countaining dimerized coiled coils,, due to being calculated based on them, while MTIDK scores higher with mixed-motif/globular proteins of other types. Both could be used simultaneously on a sequence to scrutinize the prediction.

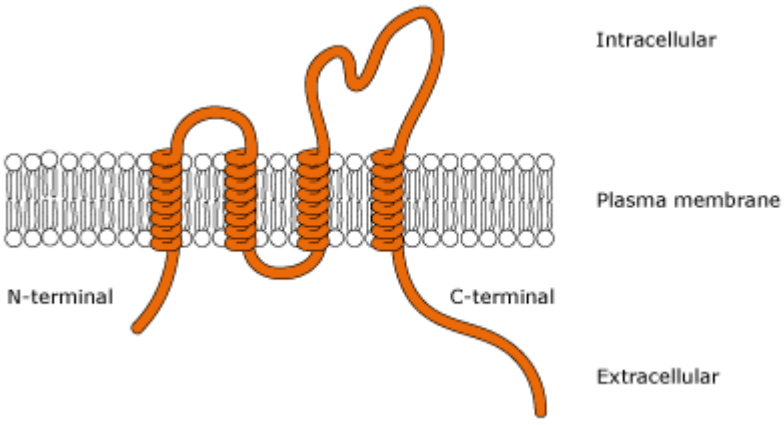
While COILS predictions are based on the matrices, MARCOIL uses HMMs trained on the same databases. MARCOIL claims to have a lower rate of false positives and more customization options, and generally higher accuracy, while the database-related specificity is the same as for COILS.

In the COILS output, we see output from scans in 3 sliding windows. These windows are based on probable lengths of coiled coil motifs within globular proteins. The conventional and most reliable window for identification of new coiled coil motifs and their terminal ends is 21, but 28 can also produce good results in some cases. 14 is there to test reading frames and investigate motif parameters in identified coils, and not for predictions themselves.

So for COILS, we look at the 21 and 28 window output, and we see a small step of 0.8 and 0.5 probability respectively, we can note that the spans of the steps are based on sliding reading frame resolutions. The coil motif is discovered, similar to MARCOIL, at around residue 137. MARCOIL output arrives at around 1 probability at a similar point as COILS, but the initial motif scores very low in comparison, only 0.2 probability of coiled coil. We can see from output data, that the first prediction at the lowest threshold occurs at residue 137, jumping to nearly 1 after 7 residues. This 7 residue jump is due to detection of a repeating motif, characteristic for coiled coils but not randomly similar short a-helical motif. The higher overall coil probability for COILS comes from averaging a larger reading window, as opposed to continuous state evaluation by the HMM.

## Membrane protein topology prediction

1. Pick your favorite transmembrane helical protein and describe based on it the typical features of transmembrane helices.



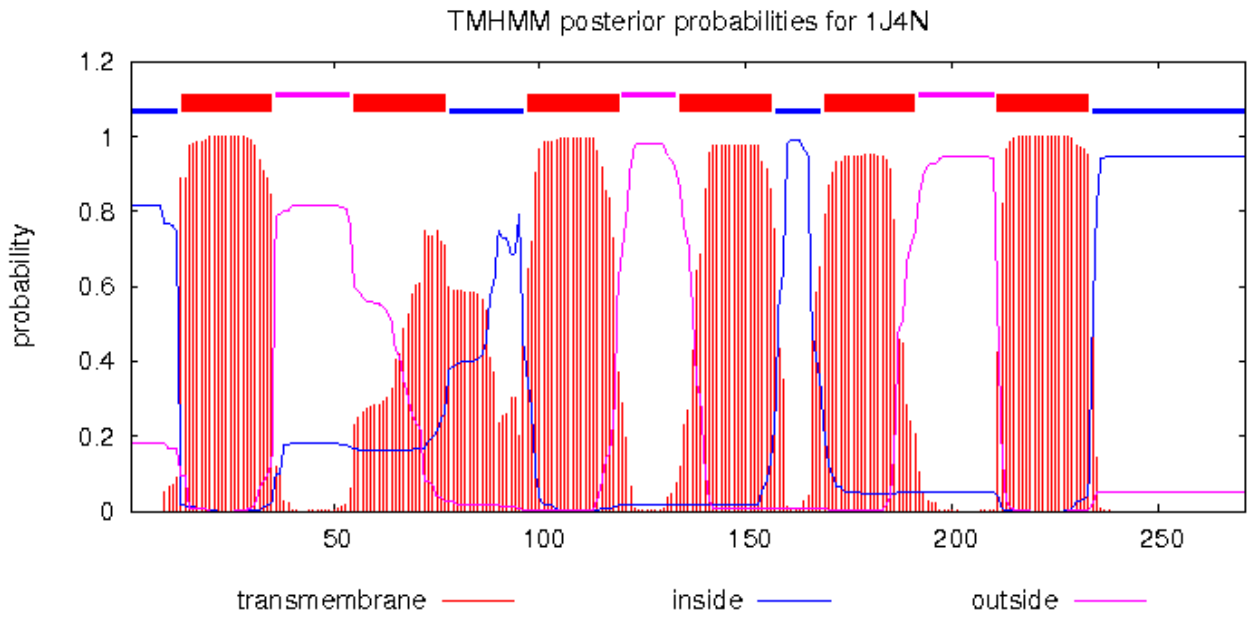

Transmembrane helical proteins are often visualized like in the picture above. The helices are not necessarily dispersed in that manner, and sometimes are bundled together.

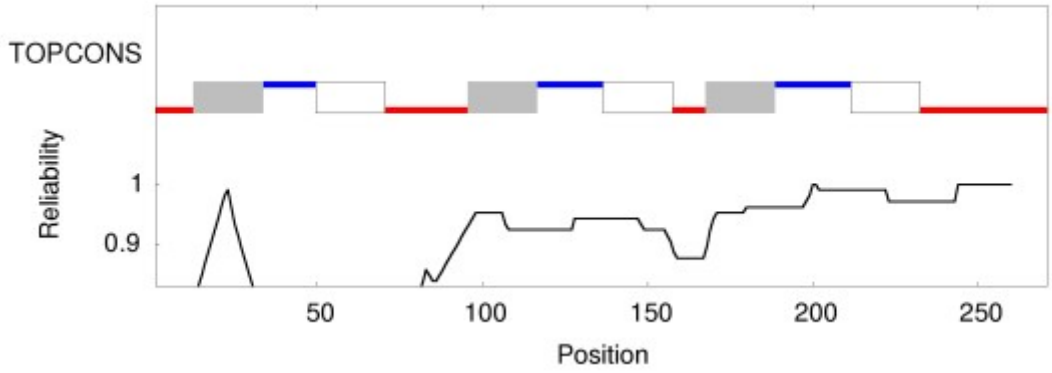
The length of the helix is usually around 20 residues, required to span the 30Å membrane, Membrane embedded helices favour hydrophobic residues and are usually a-helices. 3 10 helices are more common near the borders of the membrane, as are some amphipathic residues.

2. What does membrane protein topology mean?

Topology refers to orientation (inside/outside) of protruding residues of transmembrane proteins

3. How many transmembrane helices were predicted by the two methods for 1j4n.fa? Include screenshots of the predictions.




*Note: Colours reversed*

6 for both. However TOPCONS provides consensus using multiple tools, and 2 topology prediction tools predict 5.

4. Are the N- and C-termini of the sequence predicted to be inside or outside of the cell membrane?

Both predict inside, but 2 predictions in TOPCONS place N-terminus outside.

5. How many helices with at least 5 residues were predicted using the method?

8 with DSSP
7 with STRIDE
Difference between DSSP and Stride appears to be that Stride also considers backbone geometry, on top of hydrogen bonding predictions.

6. Explain the difference between predicting secondary structure and predicting the membrane protein topology.

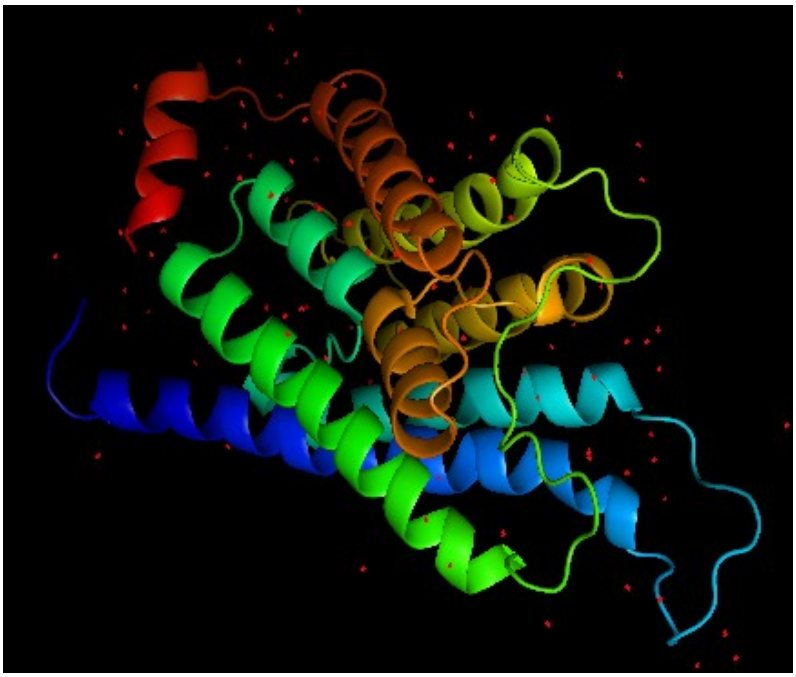
Both detect secondary structures such as a-helices and b-sheets, but topology also looks at transmembrane structure lengths, properties in transition regions, and charge of the protruding residues which determine their position inside/outside the cell.

7. Go to pdb and find the experimentally determined structure of the protein. What is the name of the protein?

Aquaporin-1

8. How many helices do you think the protein has? Include a screenshot of the image.

9 helices



9. How many transmembrane helices do you think the protein has? Motivate your answer.

6 transmembrane helices: This is revised from my previous observation of 8 helices. Technically, only 6 helices span all the way across the membrane, but the two shorter helices, while still being located in the membrane region (not protruding outside or inside), are not considered transmembrane as they dont span the length of the membrane. According to PDB, 6 helices are well within the length typical for transmembrane helices. The two short helices which were misidentified as transmembrane previously, are hemipores and contribute to permeability of the channel and are predicted to be ligand-binding, which may indicate presence of regulatory processes involving those residues.

10. How many helices do the determined structure contain according to DSSP and STRIDE? Does this result agree with your predictions and your visual inspection? Why / Why not?

DSSP 12 (9 a-helices and  3 x 3 10 helices)
Stride 11 (9 a-helices and 2 x 3 10 helices)

The discrepancy from my prediction is that I did not count the 3 10 helices separately, but instead defined an uninterrupted helical structure as one a- helix. Upon further inspection, I can identify two 3 10 helices as continuation of yellow and teal a-helices, one of which the two methods are divided on. Both methods however, agree on 3 10 helix being a part of the red a-helix, but it is hard to see only from visual inspection.