Homework VII. - Prediction in 1D

Integrated Structural Bioinformatics Course

István Szepesi-Nagy

K45SFS

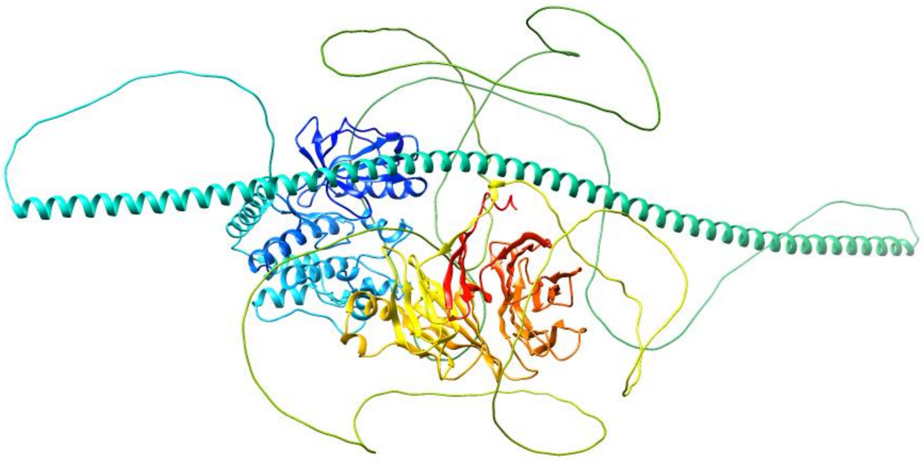
2022. 05. 03.

Content

Chosen Protein

O95819 [1]:

- **Name** Mitogen-activated protein kinase kinase kinase kinase 4
- **Number of residues**: 1239
- **Organism:** Homo sapiens



1. Figure - MAP4K4 protein - AlphaFold predicted structure

Introduction

In this task I had to predict disorder regions, coiled coil and SAH (Single α-helix) structures of the given protein sequence. These three characteristics can be identified on different levels of abstraction. Disordered regions could be identified as coiled coils and coiled coils might be SAHs. While the other direction should not be true, while these low complexity regions are densely and correctly structured.

Overall evaluation

After completing the predictions for each characteristic, I compared the results with each other. The first visible result is that the predicted regions highly overlap with on another since they have some correlation with each other.

For the chosen protein, the predictions were reliable and meaningful, because multiple tools gave similar results. After comparing the results, I have checked the output regions with the protein's 3D structure and the used tools tent to give efficient predictions.

Results

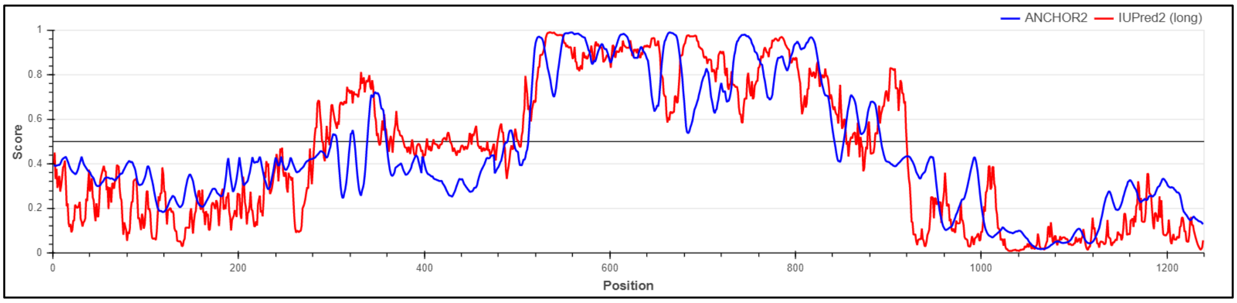
Disorders

Tools:

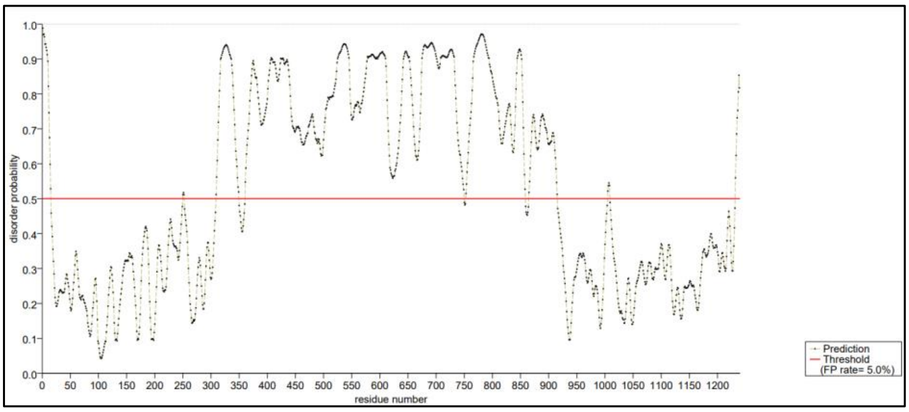
- IUPred2A [2]
- PrDOS [3]

IUPred provides disorder regions based on a simplified energy estimation method to detect parts with high interaction energies.

When I was running the predictions the FoldUnfold tool was unavailable, so I had to find another solution and found PrDOS. PrDOS identifies disorder regions based on the frequency of hydrophilic and charged residues. The higher the frequency the more disordered the region is.



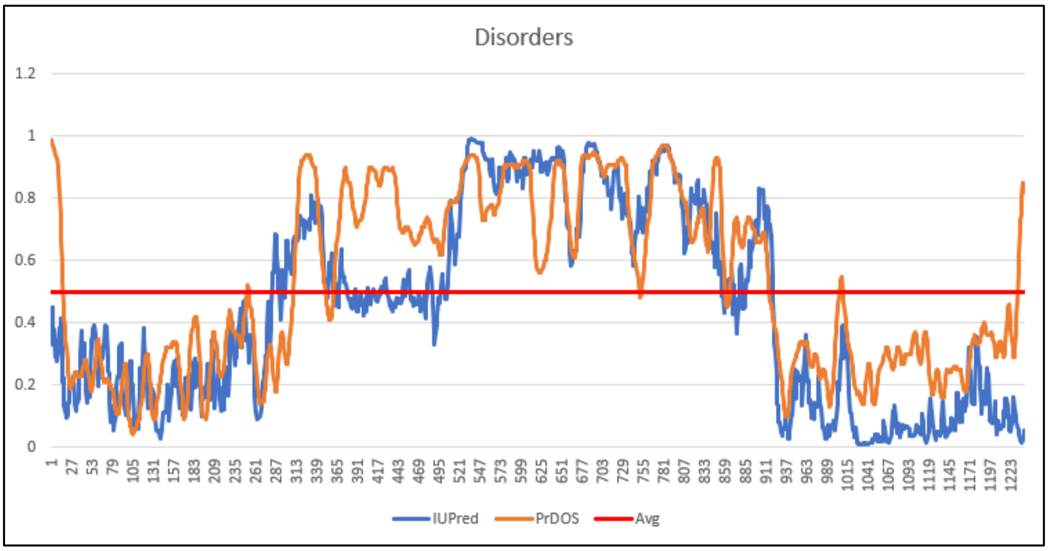
2. Figure - IUPred2A disorder prediction (red). The horizontal line represents the average value (0.5), above the line the region is more intrinsically disordered.



3. Figure - PrDOS disorder prediction. Above the horizontal line the regions are more hydrophilic / charged so more disordered.

The two diagrams show slightly different outputs, but the general segments are identified correctly. From a consensus of the two results the protein has disordered regions in the middle of the sequence (between around 300 - 900). In my opinion more detailed investigation of the predictions are not necessary, since in both cases the predicted regions are roughly in the same area, in the middle.

This could also mean, that the coiled coil and SAH structures lay in the middle of the protein sequence.



4. Figure - Disorder result comparison | IUPred (blue) and PrDOS (red).

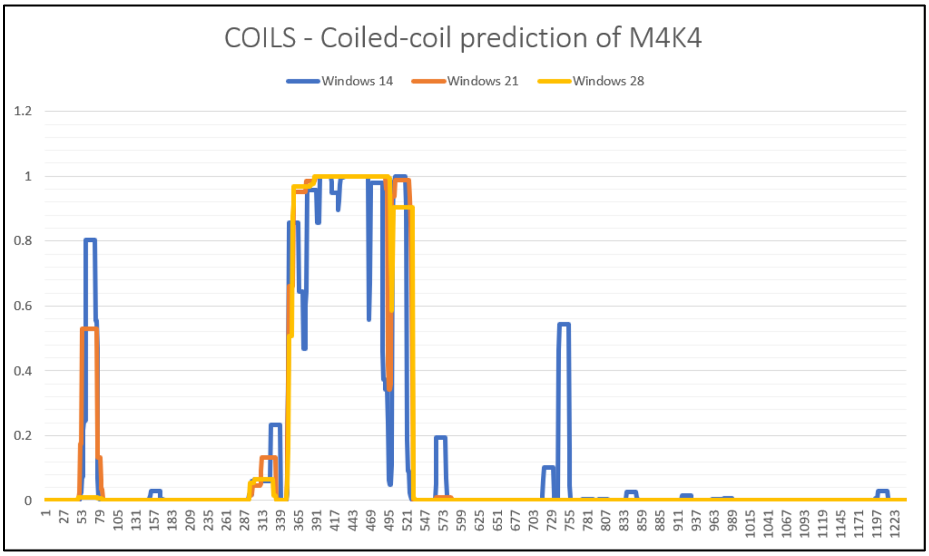
Coiled coil

Tools:

- COILS [4]
- CCHMM [5]
- Waggawagga [6]

For the coiled coil prediction, I used three tools. COILS calculates probabilities for each residue, based on known could coil proteins. CCHMM also uses known coiled coil structures, but also implements hidden Markov model for the prediction. Waggawaga calculates residue heptads from known coiled coil structures.

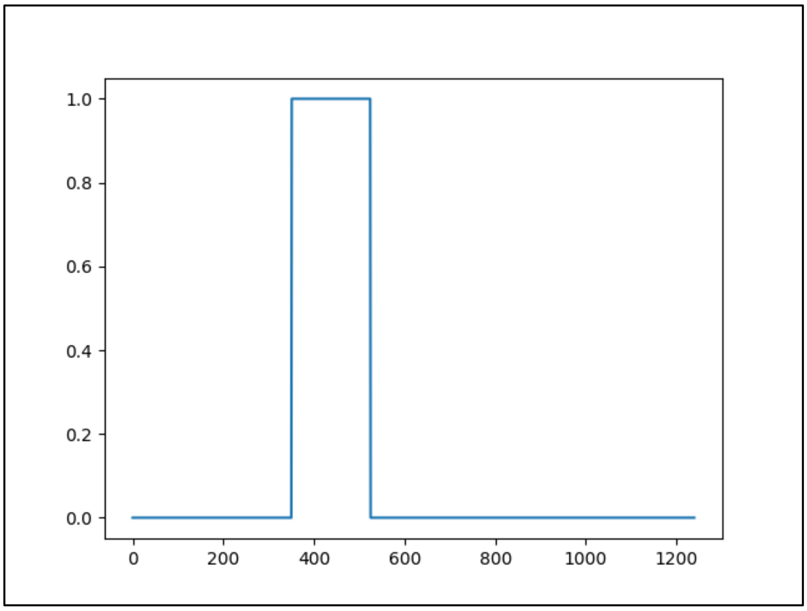
From the disorder prediction one should expect coiled coil(s) in the middle part of the protein sequence. Unsurprisingly all three tools predict a coiled coil structure around the same segment (~350 - 520).



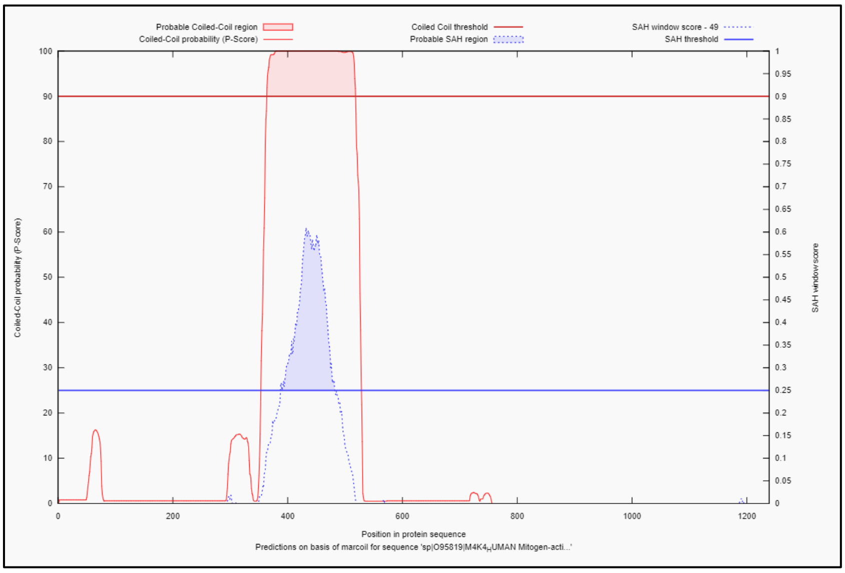
5. Figure - COILS prediction graph. Different window sizes represent the analyzed residues at a given place.

5. Figure shows a graph made with Excel from the COILS prediction data. The different window sizes represent the analyzed number of residues at a given point. Here it is visible that larger window size resulted a clear identification of the coiled coil structure.

6. Figure is a simple Pyplot output from the CCHMM output data. In this case the prediction was represented with only H-marked residues, so it was a clear boundary between the coiled coli and rest of the protein structure.



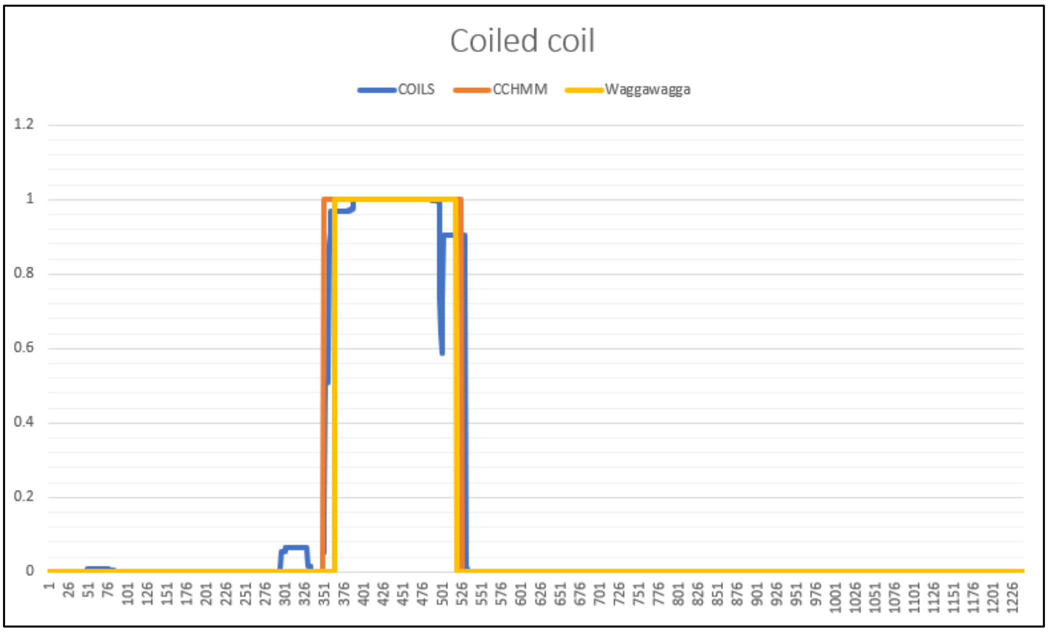
6. Figure - CCHMM prediction output using Matplotlib Pyplot library for representation.



7. Figure - Waggawagga prediction for coiled coil structure (red area).

Waggawagga shows almost the same segment and it puts a margin at 0.9 probability. Above the line in the red area, there is a very high chance of coiled coil structure.

Lastly, I compared and plotted all the results on a single chart. The common area of coiled coil prediction is around between 360th and 525th residues. Also, just for visualization I compared the 1D predictions with AlphaFold protein structure prediction on 9. Figure.



8. Figure - Coiled coil results together.

9. Figure - Thre predicted CC between 360. - 525. residues represented in 3D structure.

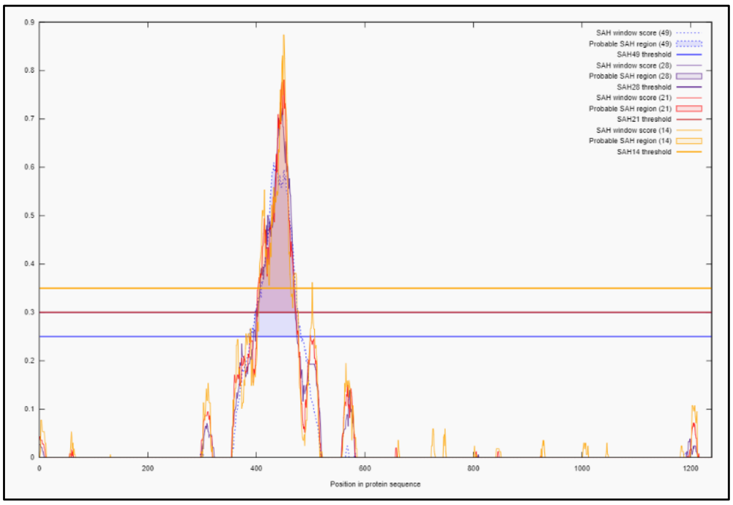
Single α-helix

Tools:

- Waggawagga [6]
- CSAHserver [7]

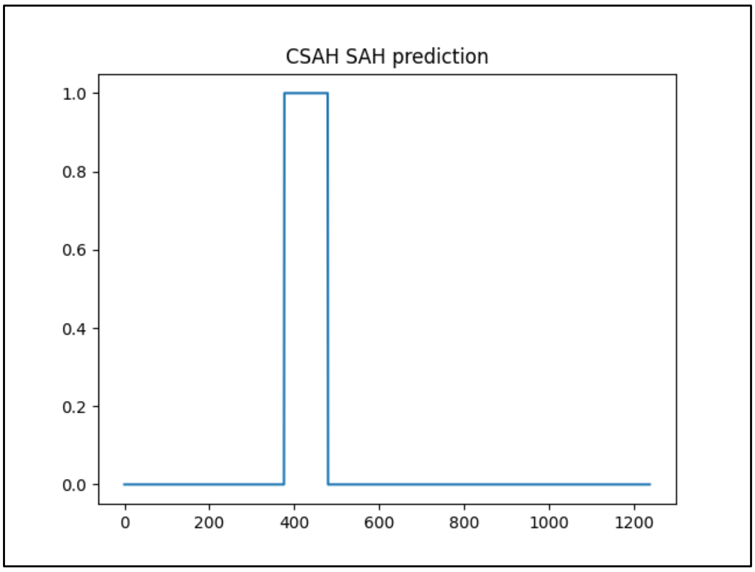
Once again, I used Waggawagga tool, but in this case for SAH prediction. It gives SAH-scores for each residue which is calculated as the sum of interactions divided by the windows size (which is set to 21 by default). SAH is usually applicable above 0,25 SAH-score.

CSAHserver gives probability values based on the ionic interactions between residues.



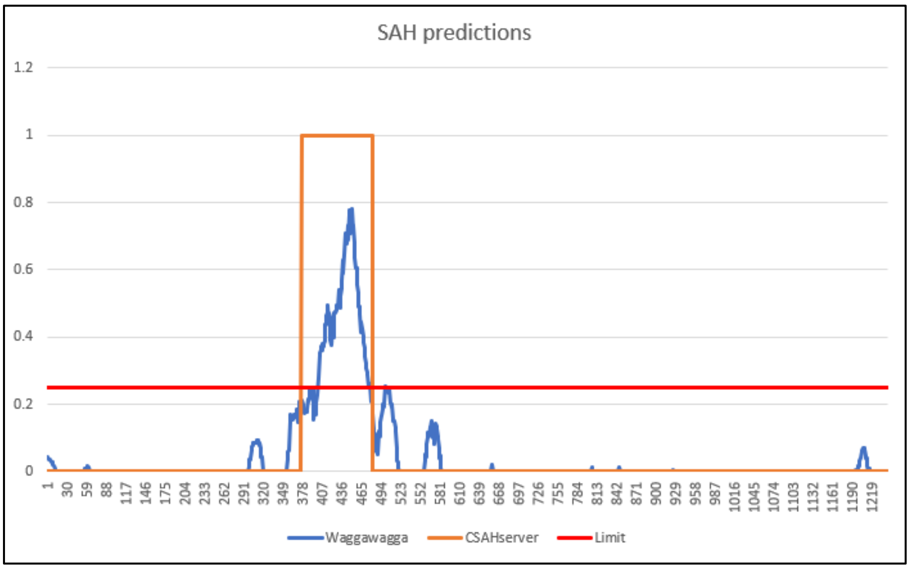
10. Figure - Waggawagga SAH prediction. Probable SAH region: blue with 26 window size, red with 21 window size, yellow with 14 window size.

For the CSAHserver output I had to use Python again for better visualization. The residues, that are probably part of the SAH structure, typed by uppercase residue letters. I plotted the results based on this differentiation.



11. Figure - CSAHserver prediction plot using Matplotlib Pyplot.

The two predictions together show similar areas, and from this one can identify SAH structures more confidently. If we consider taking the Waggawagga's 0.25 limit to count, then the common would be between residues 390. and 480.



12. Figure - SAH results together.

13. Figure - Predicted segment of the protein represented in AlphaFold's 3D prediction.

Summary

Prediction	Tool	Boundaries	Consensus
Disorder	IUPred2A	300-350 + 550-850 + 900-920	300-900
	PrDOS	300-900	
Coiled coil	COILS	350-530	360-520
	CCHMM	350-525	
	Waggawagga	360-520	
SAH	Waggawagga	390-480	390-480
	CSAHserver	375-480	

1. Table - Summary of prediction boundaries.

Resources

[1]. https://www.uniprot.org/uniprot/O95819

[2]. https://iupred2a.elte.hu/

[3]. https://prdos.hgc.jp/cgi-bin/top.cgi

[4]. https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_lupas.html

[5]. http://gpcr.biocomp.unibo.it/cgi/predictors/cc/pred_cchmm.cgi

[6]. https://waggawagga.motorprotein.de/

[7]. http://csahserver.itk.ppke.hu/cgi-bin/csahserver.cgi

Appendix

- https://github.com/sznistvan/StructBio_K45SFS/tree/main/HW7_1D