Rachael C. Kretsch
Mathematics of Big Data, Fall 2016
Harvey Mudd College
rkretsch@g.hmc.edu

Accuracy of secondary protein structure prediction tools for chromoproteins and fluorescent proteins

Anonymous Author(s)
Affiliation
Address
email

Abstract

abstract...

1 Introduction

1.1 Secondary protein structure prediction

This project aims to look at methods to predict secondary protein structure. Protein structure prediction is a major field of study and is a problem that takes massive computational power to solve. There are two main approaches looking from a biochemical point of view. The first is isolated the protein, crystallizing it, and performing crystal chromatography to figure out the structure. This structure is relaxed into its hypothesized structure via molecular dynamics. The second method is experimentally less intense. One of the most plentiful and easy to obtain biological data is DNA sequence. From the DNA sequence of a coding region there are simply rules to propose a great starting point for the protein's amino acid sequence. The problem of predicting the 3D structure from an amino acid sequence is extremely hard. I will reduce this problem to simpler features. My aim is to look at how we can use the amino acid sequence, the primary structure, to deduce secondary structure components like β-sheets, α-helices, and coils.

The field of computationally predicting secondary structures has been active since the 1970s [7]. The large majority of these methods are extremely generalizable to proteins of all classes. The training databases used for these methods are often very diverse with a goal of less than 25(%) homology between each protein [8]. This projects inspiration drew from a method created by Cambria, called Hydropathy Clustering Assisted Methods [3]. This method is specifically designed to give accurate prediction for water soluble globular proteins, which are often difficult to predict using the conventional generalizable methods because of their special hydrophobicity concerns. This project concerns itself with a group of proteins, chromoproteins and florescent proteins, whose secondary structure is predicted with below average accuracy using traditional methods, however, their secondary structure is an excellent indicator for identifying important parts of the protein for manipulation.

1.2 Fluorescent proteins and chromoproteins

Chromoproteins and fluorescent proteins are vital for biological and medical research. Chromoproteins are colored proteins and fluorescent proteins fluoresce light when the absorb a certain wavelength. They are used to target and visualize components of interest. As biological research progresses, proteins with more extreme properties such as new colors, and fluorescence able to pass through thick skin are needed. Finding these in nature is very difficult so the best method is to mutate current proteins and increase their functionality. The structure of these chromoproteins and fluorescent proteins is very important and quick conserved between proteins. The structures are mostly β-strands that wrap into β-barrels. Inside these barrels are coils that have been shown to be vital for the fluorescent properties of these proteins. Mutations to these coils cause dramatic changes in function from increasing function to changing wavelength. Obtaining structural data from experiment, either NMR or x-ray crystallography, can be very difficult and time consuming. On the other hand sequencing DNA is very simple, quick, and cheap. In fact the cost of sequencing is decreasing quicker than Moore?s law! Therefore, are computational method to convert sequence to structure would help experimentalists target areas for mutation. Because chromoproteins and fluorescent proteins have relatively predictable structural elements and these are almost wholly defined by secondary structure. This makes this data set an excellent candidate for secondary protein structure predictions.

2 Methods and materials

2.1 Logistic regression method implementation

The protein sequences were extracted from the Protein Data Base [10] from results of the searches "chromoprotein" and "florescent". Proteins less than 25 amino acids in length were removed as well as any sequences with the same ID, the first sequence was kept. The left a dataset of 739 with a total of 392,106 amino acids. The correct secondary structure to which each amino acid in these sequences belong was extracted from DSSP [1],[2], a standardized database for interpretations of NMR and crystallography results. Scoring matrices were created by analyzing the frequency that each amino acid was part of each structure. α-helices and 3-helices were identified as helical structures and were given a weight of 1. β-strands and β-bridges were identified as β structures and were given a weight of -1. The remaining structures, 5-helices, turns, and bends were identified as coil structures and given a weight of 0. Essentially this model assumes that a amino acid is either in a correct environment to form a helical structure or a β structure and if the preference for either is not strong enough, then is coils. The averaged score of an amino acid, was simply calculated by looking at all instances in the data set and averaging the described weights. This was also done for the neighboring amino acids up to 3 amino acids away. 3 was selected, because an helix can only form with 6 amino acids so we needed to be analyzing interactions at least 3 away to predict and helix. Matrices were created for each amino acid with the scores of itself and that neighbor scores for the surrounding 6 amino acids. This create a data set of 2,744,742 data points. A quarter of the data was randomly selected as a validation set. A multinomial logistic regression was performed on the remaining 75(%) of the data. L2 regression was shown to be the most precise with a regularization factor of 3.4. Accuracy was calculated according the the convention of secondary structure prediction methods, Q3.

$$Q3 = \sum_{n=1}^{k} NC(i)/NO(i)$$

where NC(i) is the number of correct amino acid predictions for protein i, and NO is the number of observed amino acids for protein i.

2.2 Literature method testing

2.2.1 GORIV [4]

GORIV is a secondary protein structure prediction method based in information theory and bayesian statistics. It uses the basis that given a data set os sufficient size we can find the probability of an amino acid of some type being a certain secondary structure. This can be extended to joint distributions involving the amino acids next door. The result in an expression for the probability of the confirmation S at a position i given the surrounding amino acid identities R_n

$$I(\delta S_i; R_1, ..., R_n) = log[\frac{P(S_i, R_1, ..., R_n)}{P(n - S_i, R_1, ..., R_n)}] + log[\frac{P(n - S)}{P(S)}]$$

Approximations:

The method has a training data base of 267 proteins. The newest update, GORV, included evolutionary data to increase accuracy [9].

The sequences prediction data was calculated from `https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html`.

2.2.2 SOPM [5]

The first step for the SOPM method is to generate a limited database. This is done by using an alignment algorithm by Needleman and Wunsch (1970) [11] to pairwise compare the input protein to the proteins in the reference database. Because, this alignment method, as with most alignment methods, is parameterized with the idea that sequences compared will be quite similar we must be very careful because we are comparing proteins that are sometimes vastly different, especially in length. The parameter that is most volatile is the gap penalty, the parameter that penalizes a large insertion not present in the other sequences. To reduce this, 5 randomized copies of each protein of the reference database were created. The averaged randomized score was taken as the noise, and the actual protein sequence similarity could then be separated from this noise. The database is selected from the group of proteins most similar to the input protein. However, only one of each highly related protein is added so as not to skew the prediction parameters by a large group of very similar structures and sequences.

The second and third step consist of a continuous loop until prediction accuracy is sufficient. Firstly, we loop through all the protein in the sub-database and individually predict their secondary structure using the data from the other proteins in the sub database. This is accomplished by breaking the protein into 17 overlapping amino acid sequences which are then compared. The scores are accumulated as each protein is compared. There are adjustment parameters for each of these scores representing the adjustment for α-helix, β-sheet, and coils. These are optimized every step by the following equation

$$f_k(i+1) = f(i) + \frac{NO(i) - NP_k(i))}{NP_k(i)}$$

where NO(i) is the total amount of amino acids observed and NP_k is the amount of amino acids predicted as k structured. When the fit is satisfactory it uses the fit to predict the input protein. The method has a training database of 239 proteins.

The sequences prediction data was calculated from `https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopm.html`. The sequence was predicted for the 3 structures, with a similarity threshold value of 8 and a window length of 17.

2.2.3 s2D [8]

This is the newest method tested and therefore predictably has the largest training set of 2671 proteins.

The sequences prediction data was calculated from `http://www-mvsoftware.ch.cam.ac.uk/index.php/s2D`.

2.3 Comparing methods

The four methods were compared for the 20 amino acids that the logistic regression performed best and worst on. The accuracies were compared for each of these proteins. The identity, function, and structure was also manually analyzed to investigate if there were any patterns that caused a method to fail.

3 Results

3.1 Logistic regression

The logistic regression parameters were chosen to be L2 regression and a regression parameter of 3.4 as seen in Figure 1. the coefficients of the fit are seen in Figure 2. The selection of the training set

Table 1: Accuracy of various secondary structure prediction methods

Method	Accuracy on data set (%)	Reported accuracy (%)
Logistic regression	54	-
GORIV	48	64
SOPM	51	69
s2D	64	85-88

was shown to not have a significant effect on the accuracy (Figure 3). The accuracy of individual protein selection was skewed towards high accuracies and centered around 54(%) (Figure 4).

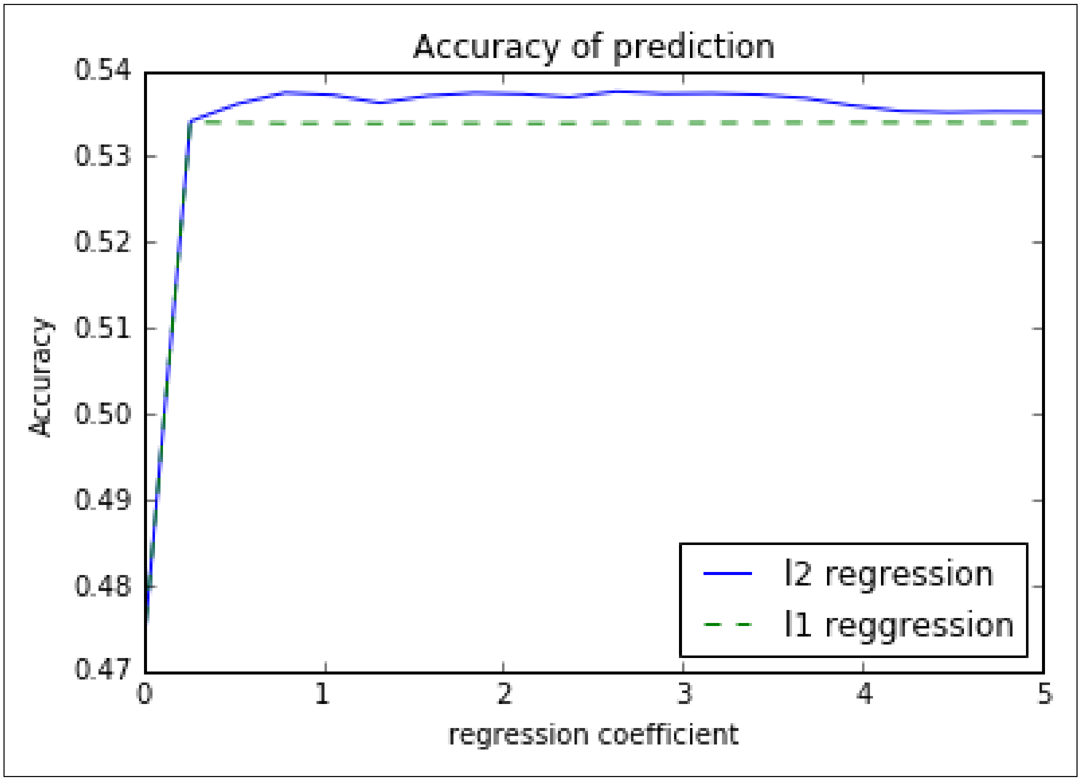

Figure 1: Selecting the logistic regression parameters.

3.2 Method comparison

4 Discussion

The method develop could be used on a chromoproteins or florescent protein whose structure is unknown. Let's say a new protein is discovered that has far-red florescence and we want to try and increase this function. We could use this method to predict were the coil structures are and pinpoint these parts of the sequence for mutations. This targeted approach would be a vast improvement over random mutations.

The accuracy measurement could be adjusted to give a more precise picture. Each method gives the probability that the given amino acid is each secondary structure. So these confidence intervals could be used to improve the accuracy measurements.

More generalized methods need to be investigated to conclude that chromoproteins and florescent protein secondary structure is not accurately predicted by these and warrants another method. One such method is YASSP, which focuses on choosing optimal kernels in a cascaded model constructed from two SVM-based models [6].

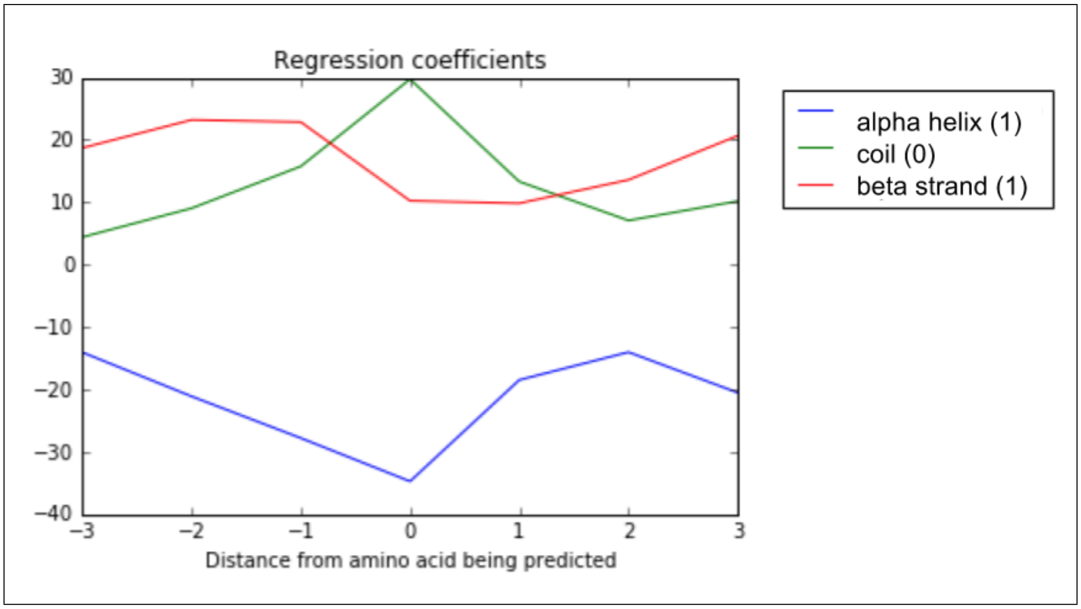

Figure 2: Logistic regression fit coefficients.

Table 2: Accuracy of methods for specific proteins

	Accuracies (%)			
Protein ID	Logistic regression	GORIV	SOPM	s2D
1bgp	25	64	62	64
4q7t	25	44	42	67
4qgw	25	69	61	83
5h88	26	37	37	57
4l1s	26	64	52	70
5h89	27	37	39	60
3s0f	27	49	58	70
4q9w	27	51	55	70
3rwt	27	37	35	48
5hzo	28	37	46	64
1bfp	60	48	60	77
3ekh	60	54	61	55
3ned	60	40	46	73
4k3g	60	49	54	59
3cfc	60	58	60	61
1xkh	60	50	48	47
2wht	60	37	58	70
4w6b	60	44	54	69
4xvp	60	44	49	52
3dqh	60	42	56	73

An interesting future problem could be to implement a machine learning algorithm to address post-transcriptional modifications. Essentially, the DNA sequence does not map perfectly to the

Availability

All data, source code, and text from this project can be found at this git hub repo: `https://github.com/hmc-cs-rkretsch/Secondary-Protein-Structure`

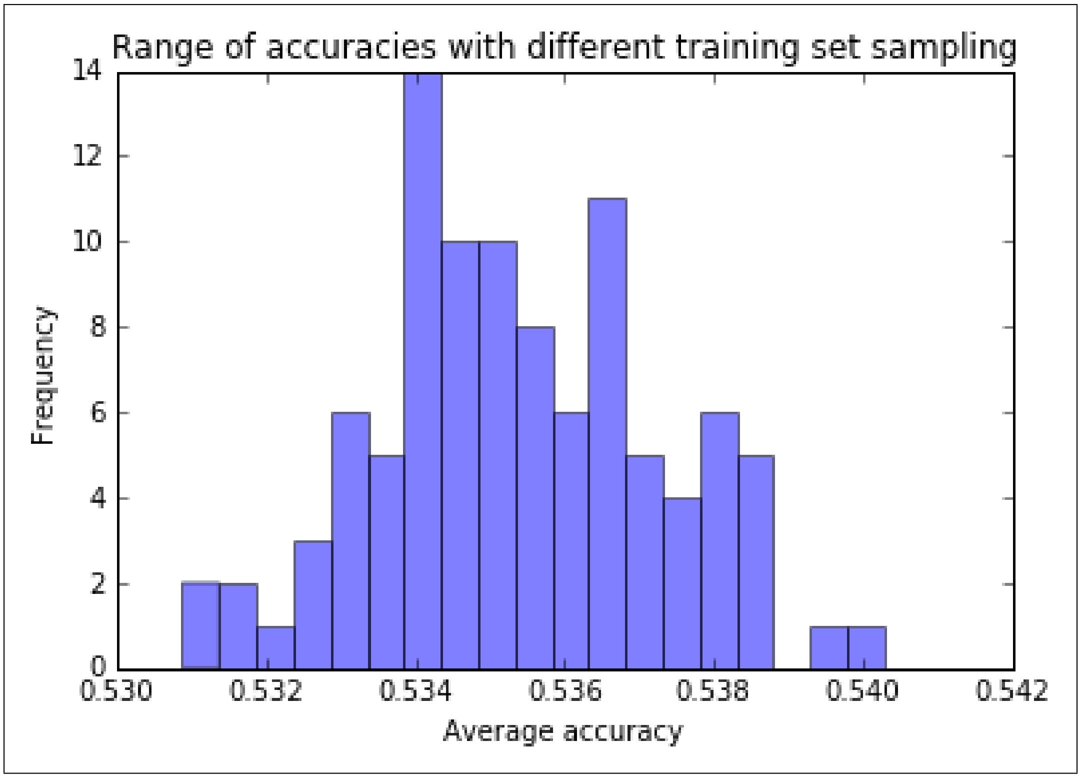

Figure 3: Accuracy dependence on training set sampling.

Acknowledgments

I would like to thank Professor Gu for teaching me the basics that allowed me to better understand these complex and cool methods. Thank you to all the graders for making this course reliable. And finally, thank you to my past research advisors and professors for helping me find my areas of interests.

References

[1] Touw, W.G. & Baakman, C. & Black, B. & te Beek, T.AH. & Kreiger, E. & Joosten, R.P. & Vriend, G. (2015) A series of PBD related databases for everyday needs. for connectionist rule extraction. *Nucleic Acids Research*, 43(Database issue): D364-D368.

[2] Kabsch W. & Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features *Biopolymers*, 983 22 2577-2637. PMID: 6667333; UI: 84128824.

[3] Cambria, A. (2009) Hidropathy Clustering Assisted Methods. http://www.acbrc.org/hcam.html

[4] Garnier, J. & Gibrat, J.F. & Robson, B. (1996) GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. *Method in Enzymology*. 266:540-53.

[5] Geourjon C. & Deleage G. (1994) SOPM: a self-optimimization mehtod for protein secondary structure prediction. *Protein Engineering.* 7(2):157-64.

[6] Karypis, G. (2006) YASSPP: Better Kernels and Coding Schemes Lead to Improvements in Protein Secondary Structure Prediction. *Proteins.* 64:575-86.

[7] Singh, M. (2001) Predicting Protein Secondary and Supersecondary Structure. Princeton University. CRC Press.

[8] Sormanni, P. & Camilloni, C. & Fariselli, P. & Vendruscolo, M. (2015) The s2D Method: Simultaneous Sequence-Based Prediction of the Statistical Populations of Ordered and Disordered Regions in Proteins. *J. Mol. Biol.* 427: 982-96.

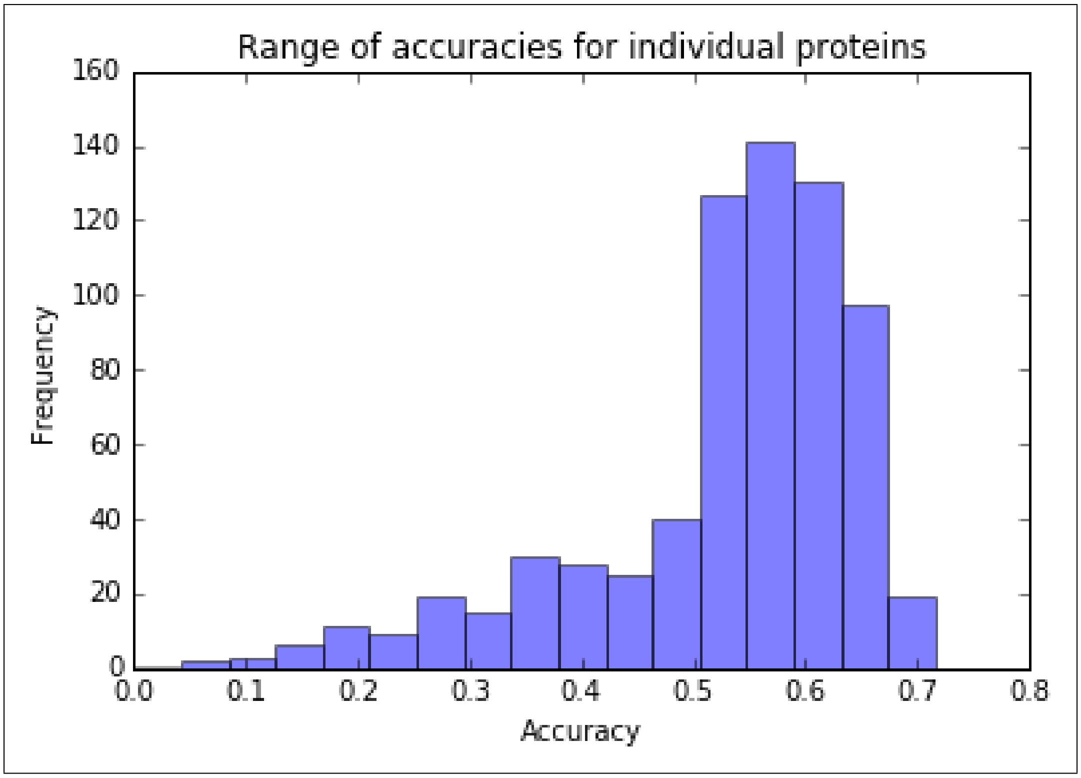

Figure 4: Range of accuracies in the data set.

[9] Sen, T.Z. & Jernigan, R.L. & Garnier, J. & Kloczkowski, A. (2005) GOR V server for protein secondary structure prediction. *Bioinformatics* Jun 1; 21(11): 2787?2788.

[10] RCSB Protein Data Bank. An Information Portal to 124928 Biological Structures. 739 proteins used, please see additional resources for these proteins and acknowledgments to all the scientists to whom these 739 proteins structures are acknowledged.

[11] Needleman, S.B. & Wunsch, C. (1970) *J. Mol. Biol.* 48, 443-453.

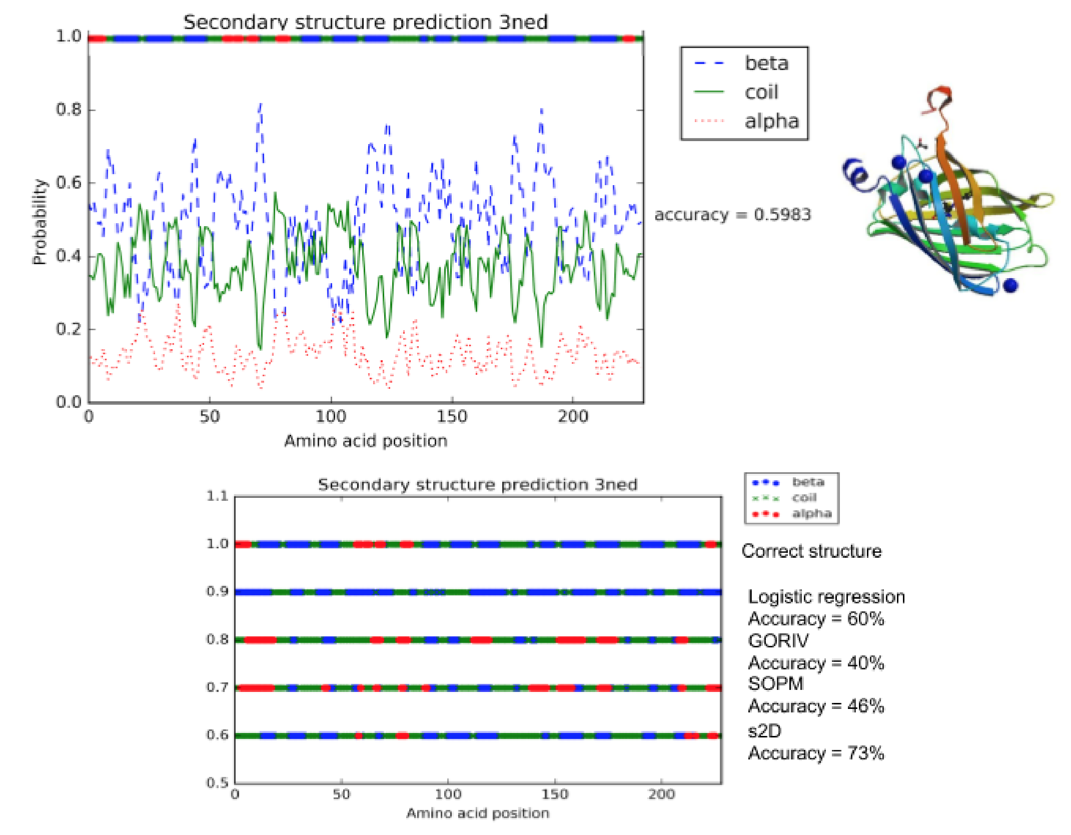

Figure 5: Method comparison for 3ned.

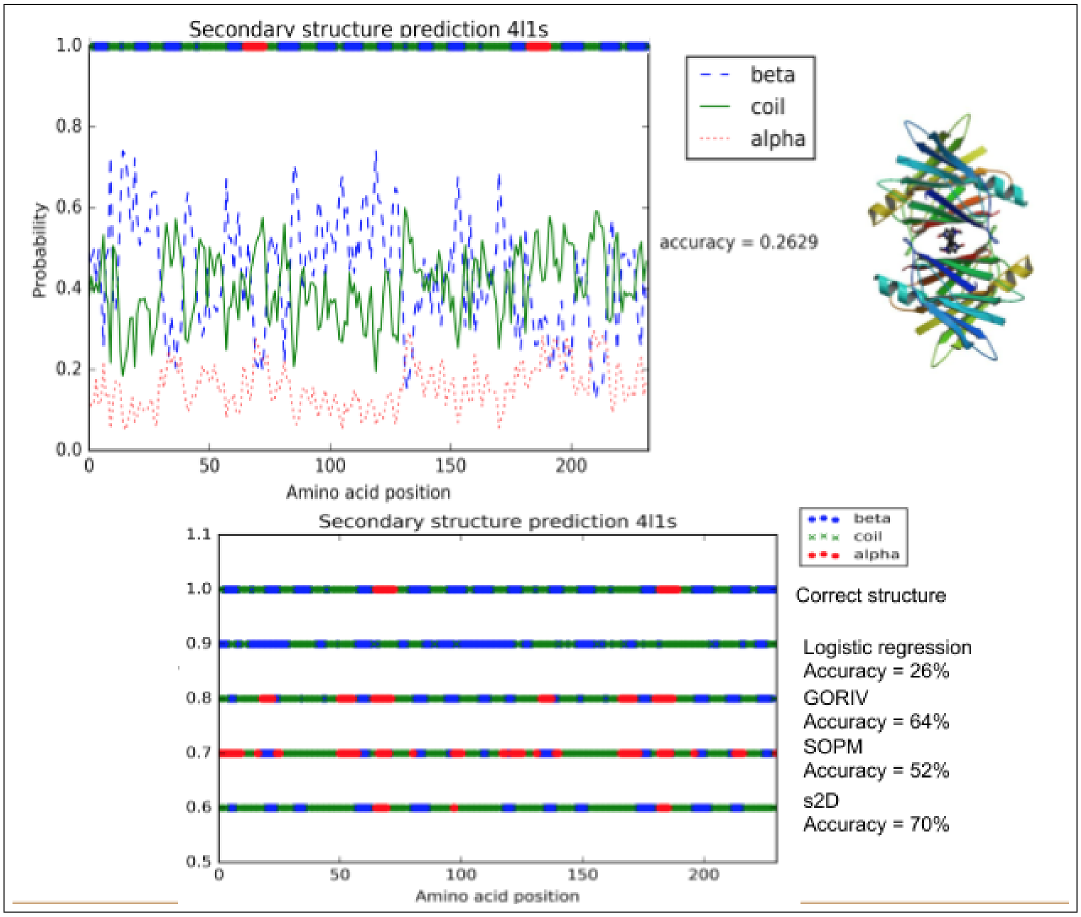

Figure 6: Method comparison for 4l1s.