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. I have previously done work on molecular dynamic methods, but now I would like to look at it from the other direction. 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 beta sheets, alpha helices, and coils.

2 Methods and materials

2.1 Logistic regression method implementation

2.2 Literature method testing

2.2.1 GORIV

2.2.2 SOPM

2.2.3 s2D

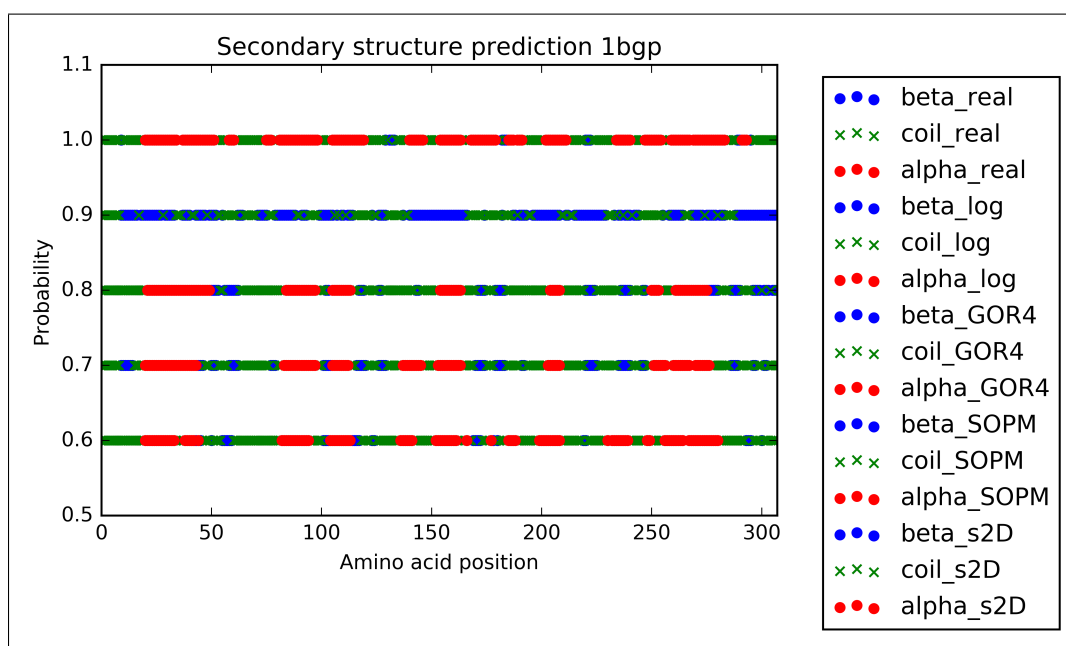# 3 Results

3.1 Footnotes

3.2 Figures

Figure 1: Sample figure caption.

3.3 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

                    https://www.ctan.org/pkg/booktabs

This package was used to typeset Table 1.

Table 1: Sample table title

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

4 Discussion

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

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`

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.

[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.