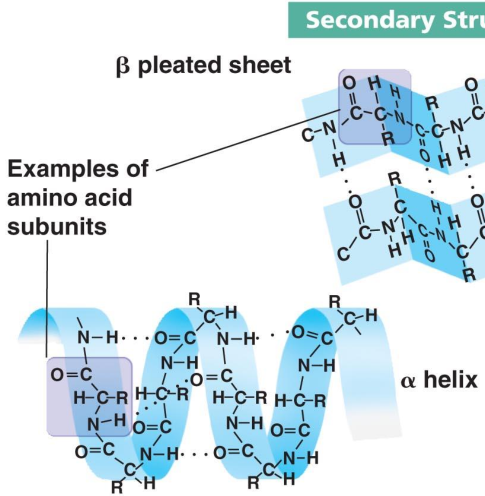
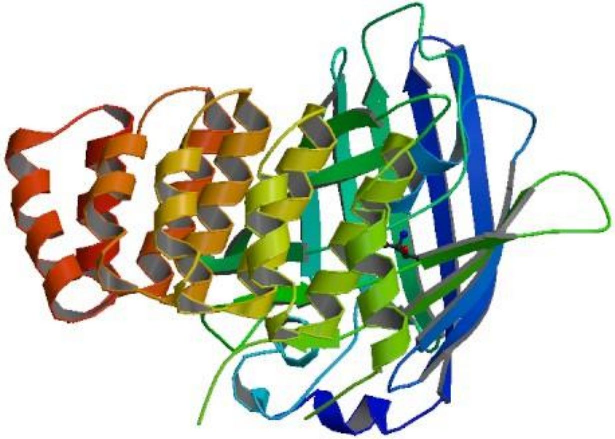
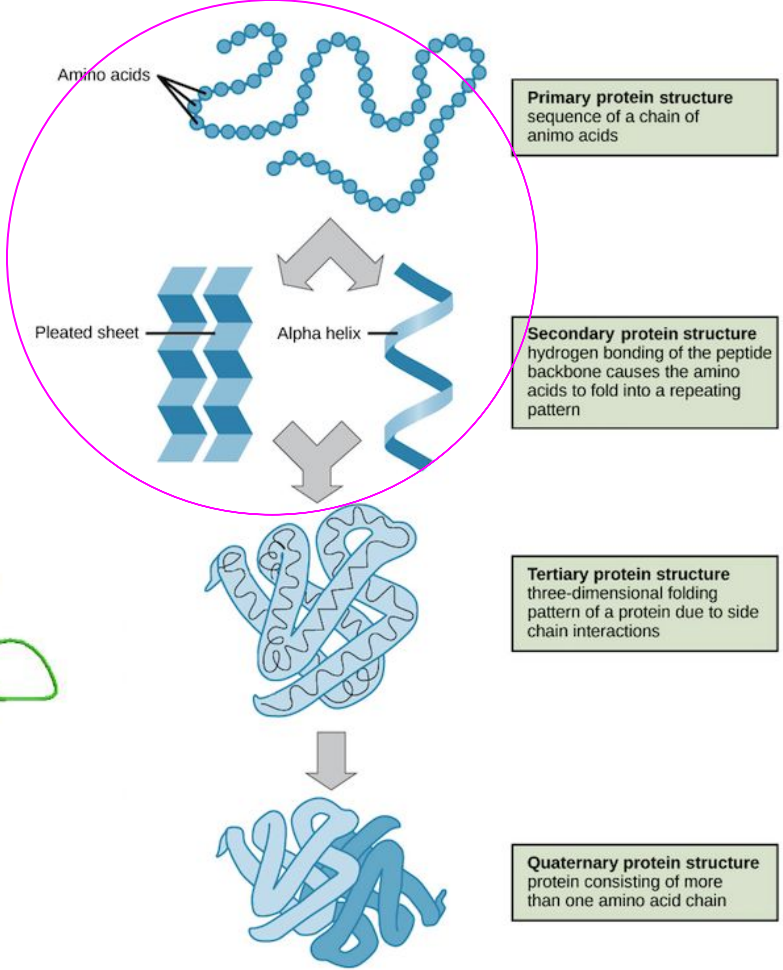
Secondary Protein Structure Determination

Big Data Final Project
Fall 2016
Rachael Kretsch

Protein Structure







Primary protein structure
sequence of a chain of animo acids

Secondary protein structure
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern

Tertiary protein structure
three-dimensional folding pattern of a protein due to side chain interactions

Quaternary protein structure
protein consisting of more than one amino acid chain

Chromoproteins and Fluorescent Proteins

“Chromoprotein” and “Fluorescent” PDB

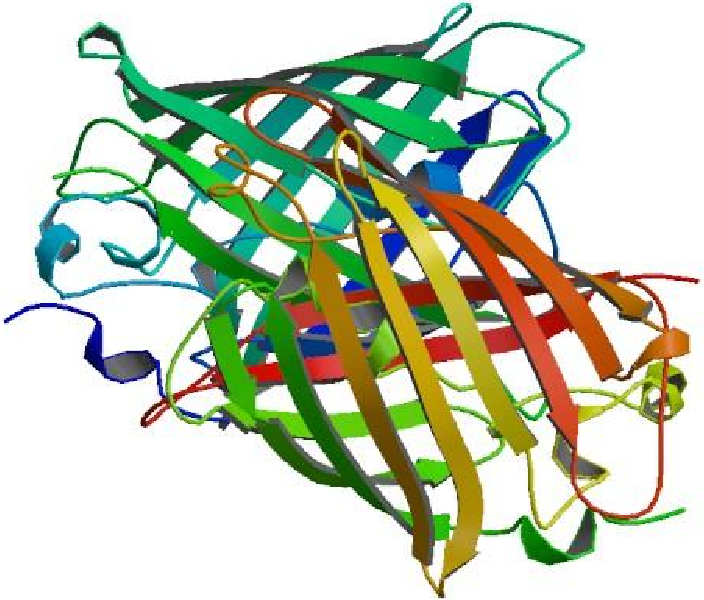
5 data points removed <25

Only took one form of each protein

739 total proteins

392,106 amino acids

2,744,742 values



SOPM 239 proteins (1994) , 267 proteins GORIV (1996), s2d 2671 proteins (2014)

Data Base

124,928 proteins

RCSB PDB

PROTEIN DATA BANK

My algorithm

Interpret fasta	Ignore unusual amino acids Ignore duplicates of same sequence ID Delete short sequences (<25)
Get structure	Dssp: standardized secondary structure assignment library for PBD from NMR and/or crystallography data
Score	Helix (1): alpha-helices, 3-helices Coil (0): turns, coils, bends, 5-helices Beta (1): strands, bridges
Get matrices	7 scores per amino acid
Logistic regression	L2, reg=3.4
Assess	Accuracy: Q3 $Q3 = \sum_{n=1}^{k} NC(i)/NO(i)$

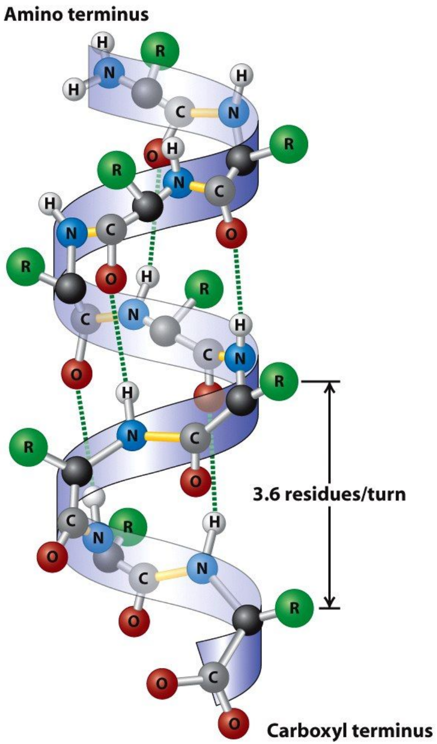
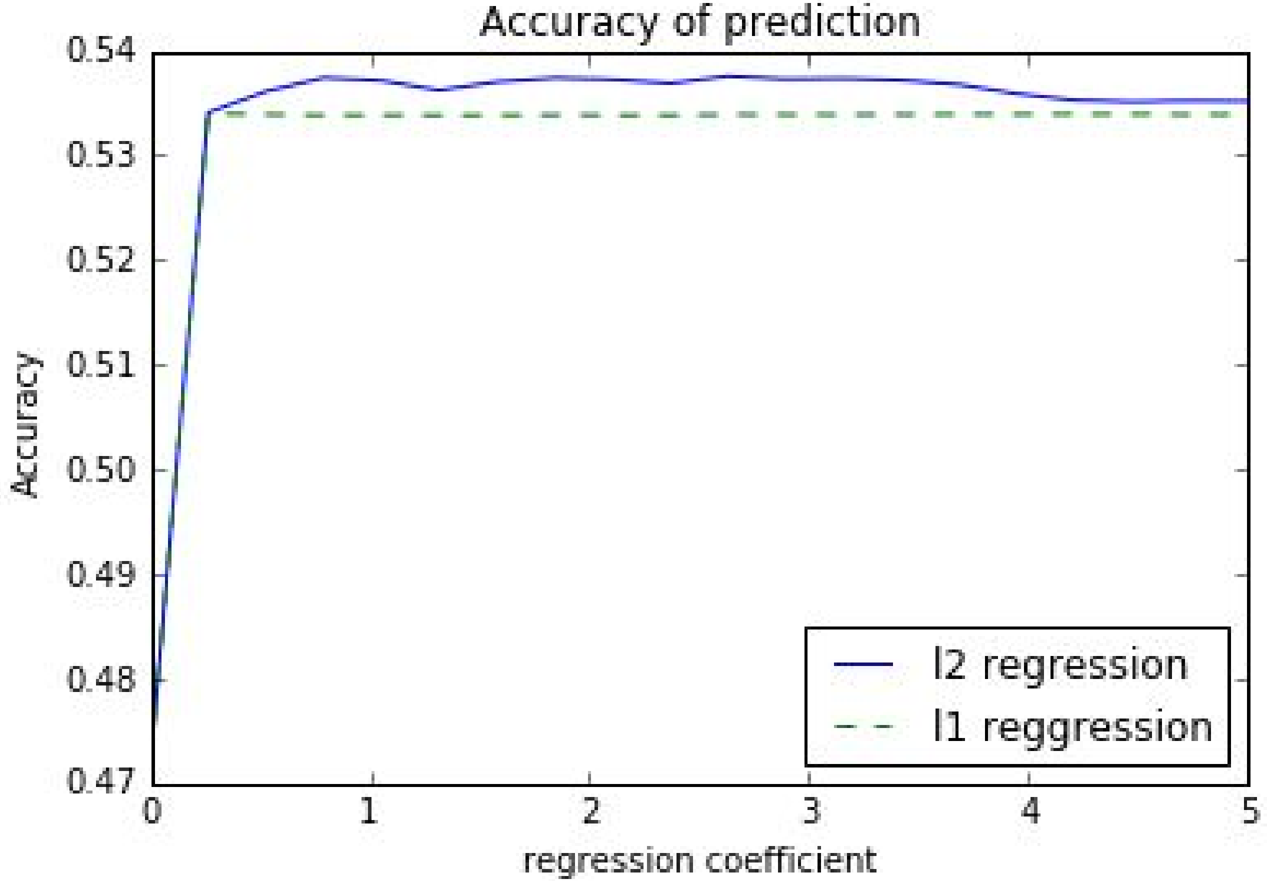
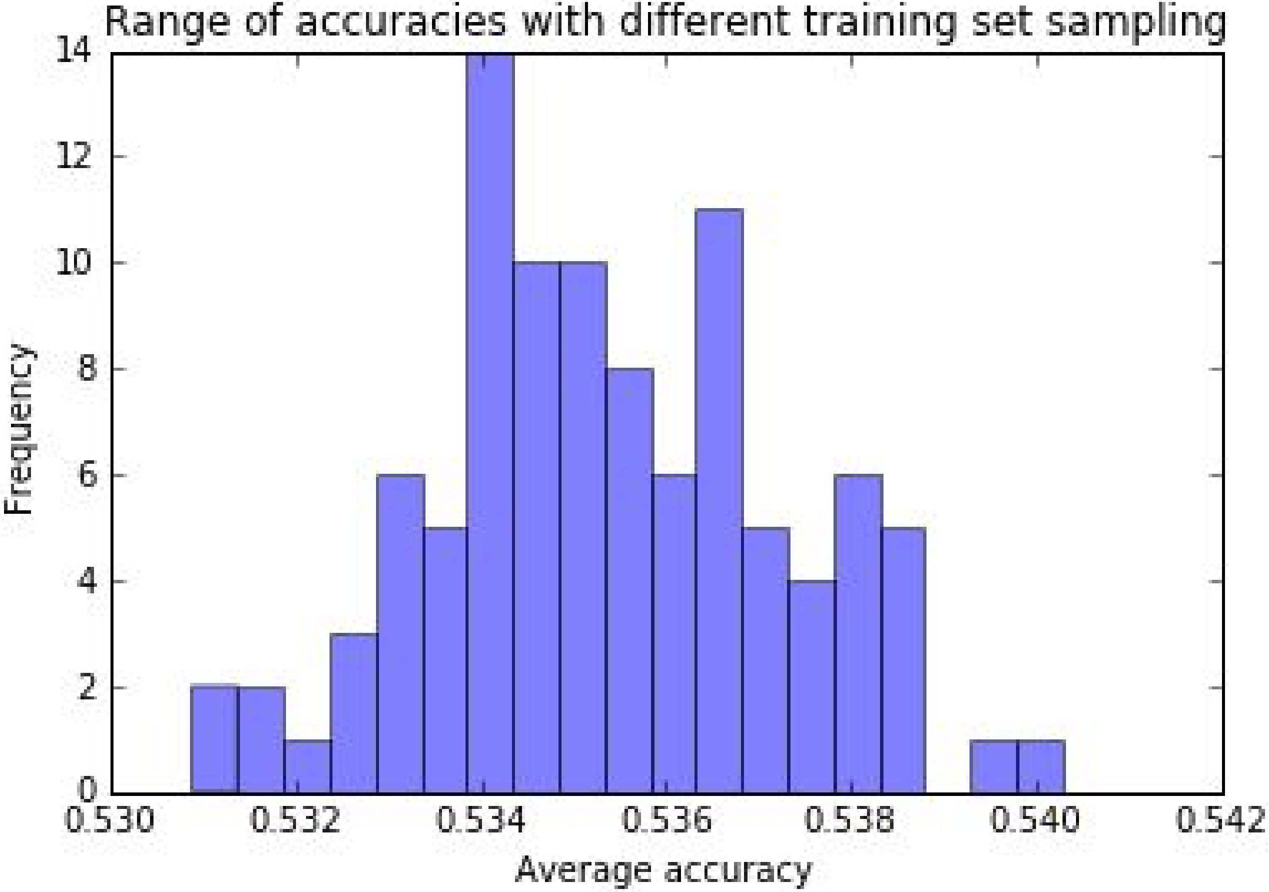
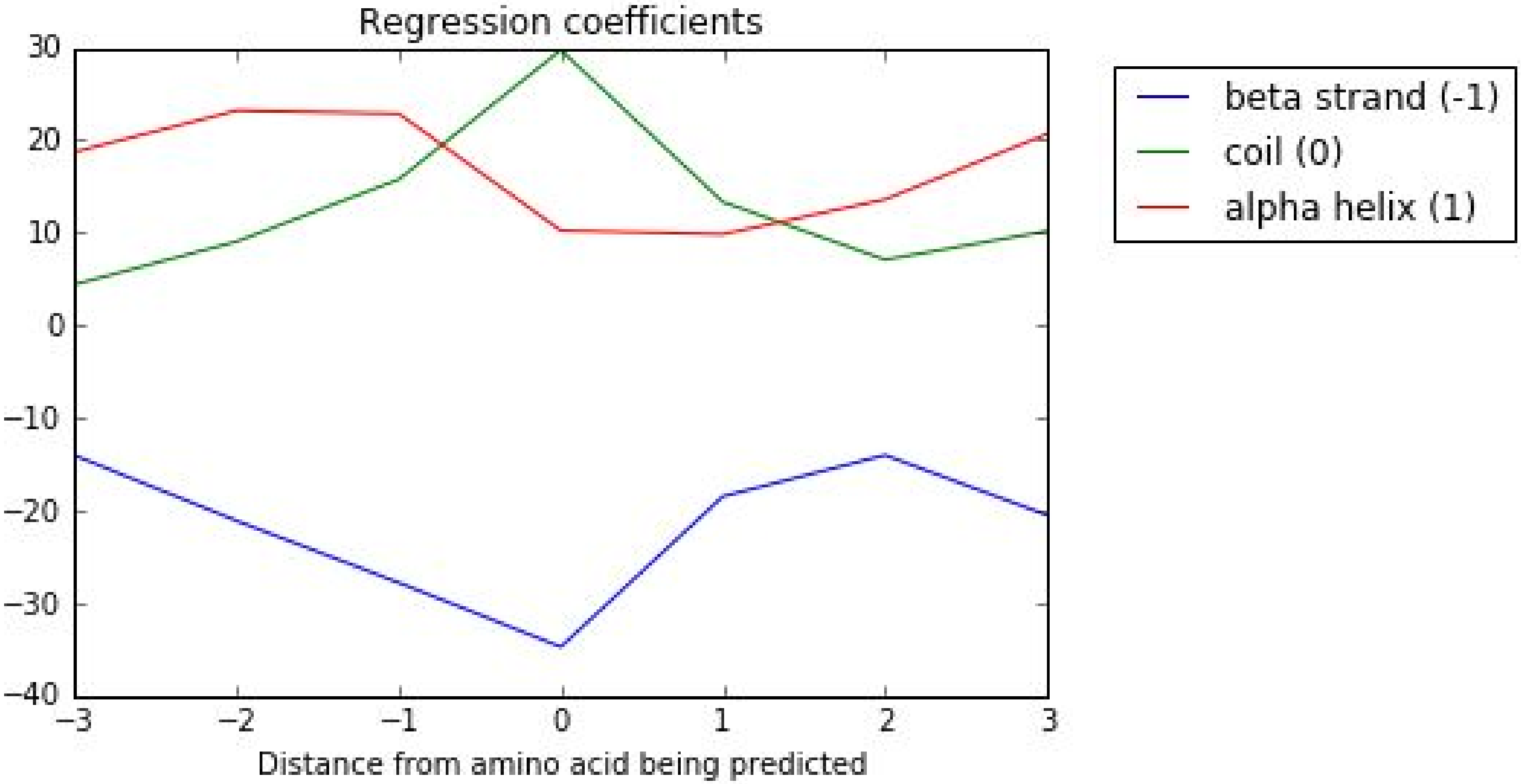
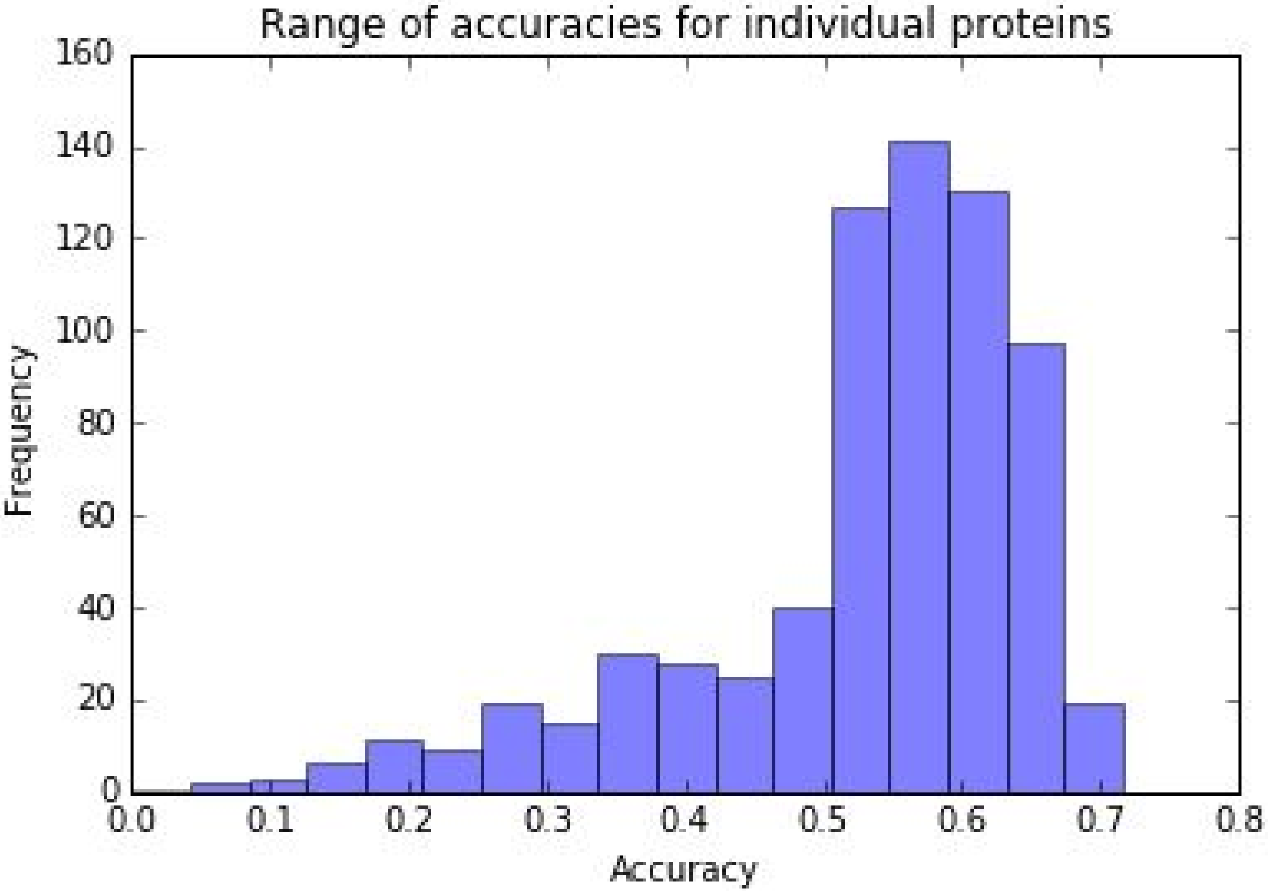


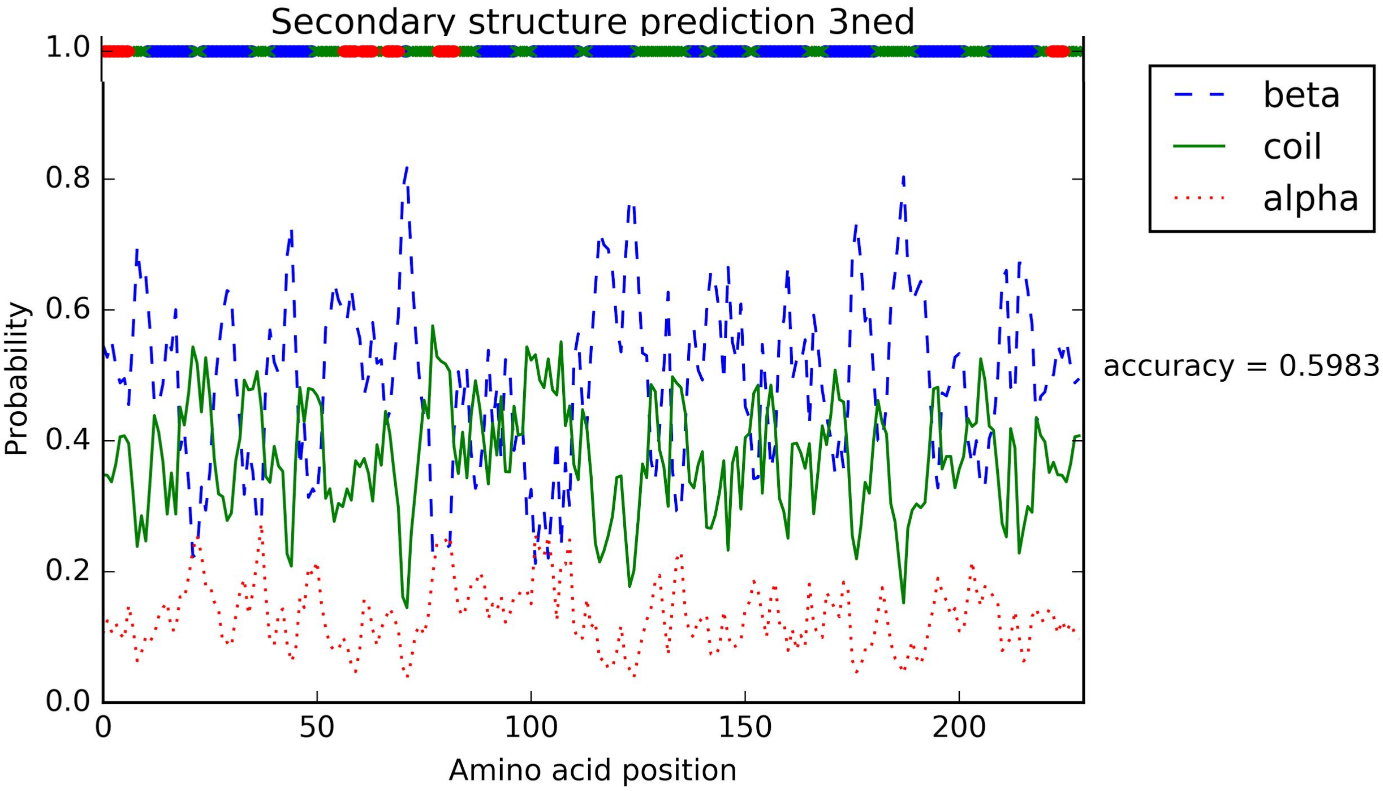
Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company


Accuracy of prediction
Accuracy
0.54
0.53
0.52
0.51
0.50
0.49
0.48
0.47
0
1
2
3
4
5
regression coefficient
l2 regression
l1 reggression

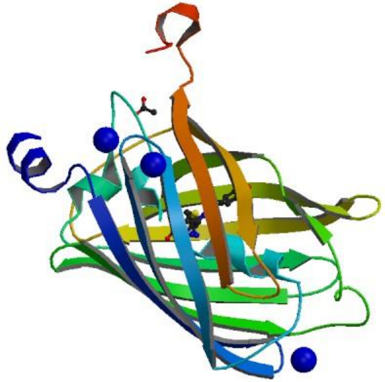

Range of accuracies with different training set sampling
Frequency
0
2
4
6
8
10
12
14
0.530
0.532
0.534
0.536
0.538
0.540
0.542
Average accuracy

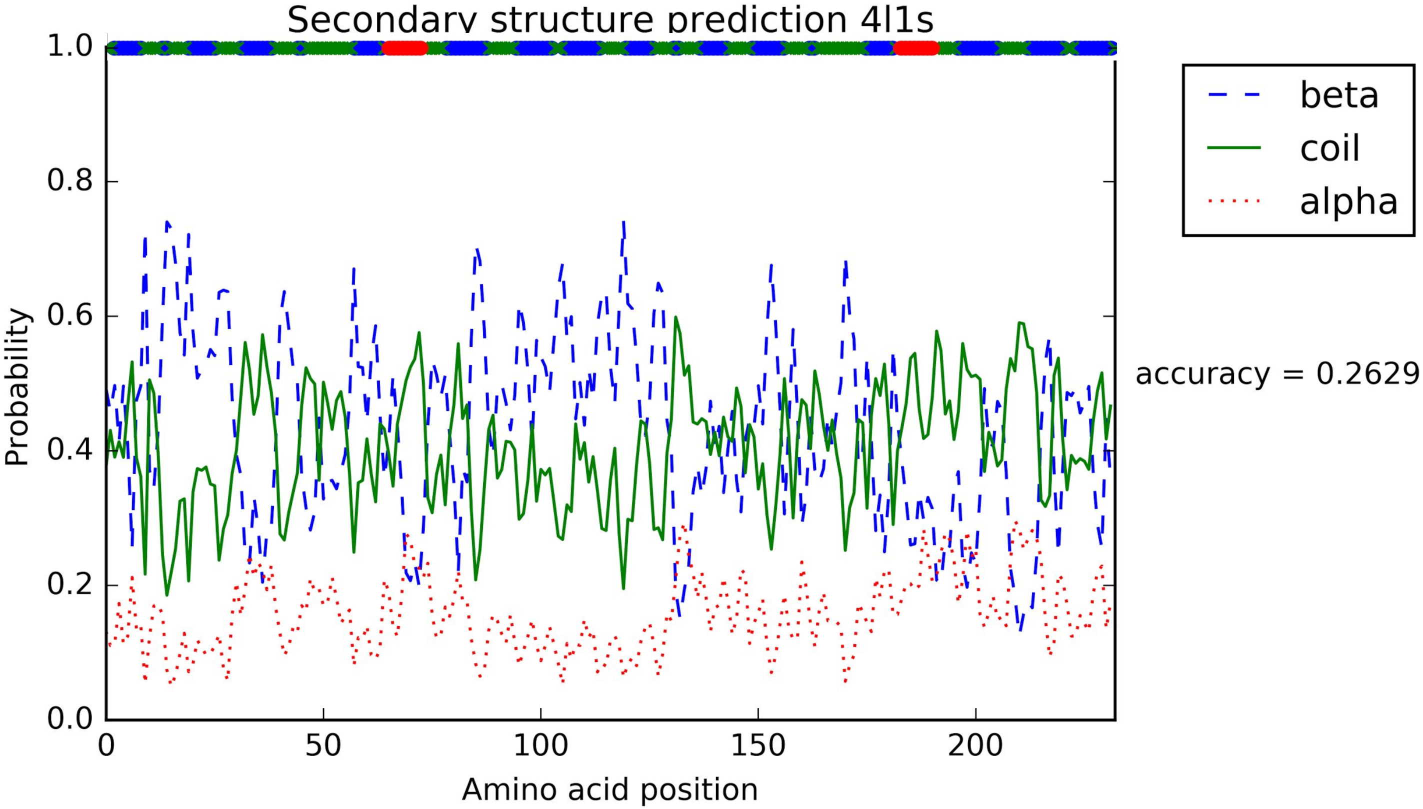

Regression coefficients
30
20
10
0
−10
−20
−30
−40
−3
−2
−1
0
1
2
3
Distance from amino acid being predicted
beta strand (-1)
coil (0)
alpha helix (1)

Range of accuracies for individual proteins
Frequency
0
20
40
60
80
100
120
140
160
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

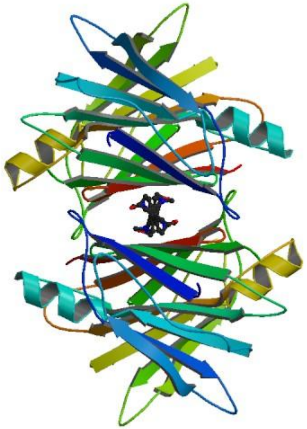


accuracy = 0.5983





accuracy = 0.2629



GOR IV

Information Theory
Bayesian Statistics

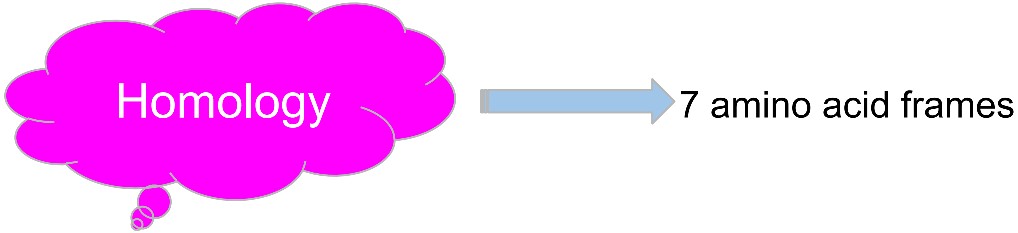
Scoring algorithm

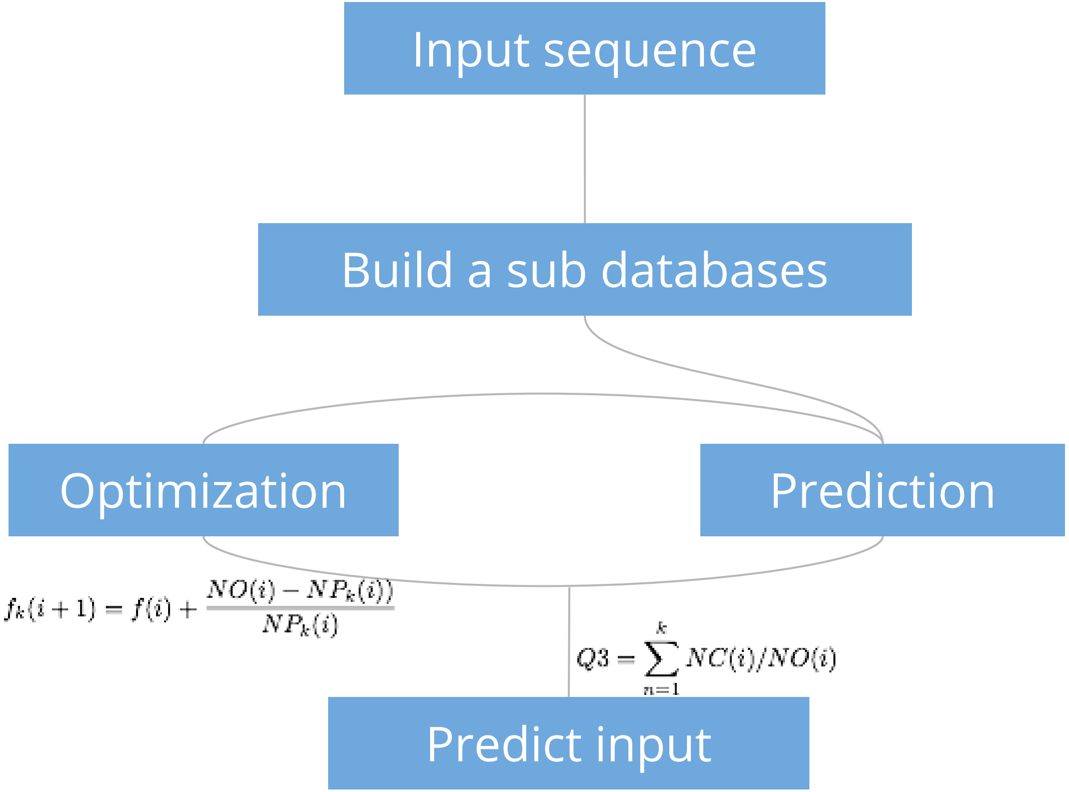
$$I(\Delta S_i; R_1, ..., R_n) = log[\frac{P(S_i, R_1, ..., R_n)}{P(n - S_i, R_1, ..., R_n)}] + log[\frac{P(n - S)}{P(S)}]$$

$$log[\frac{P(S_i, R_1, ..., R_17)}{P(n - S_i, R_1, ..., R_17)}] = \frac{2}{17} \sum_{m=-8, n>m}^{+8} log[\frac{P(S_i, R_i + m, R_i + n)}{P(n - S_i, R_i + m, R_i + m}]$$

$$-\frac{15}{17} \sum_{m=-8}^{+8} log[\frac{P(S_i, R_i + m)}{P(n - S_i, R_i + m}]$$

SOPM

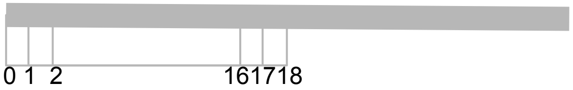




$$f_k(i+1) = f(i) + \frac{NO(i) - NP_k(i))}{NP_k(i)}$$

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

For all proteins in sub database:
Predict secondary structure by
Sequence homology with other proteins



s2D

BMRB dataset

s2D dataset

PSI-BLAST

SLFN (window=11)

SLFN (window=15)

Secondary-structure populations

N-to-1 network

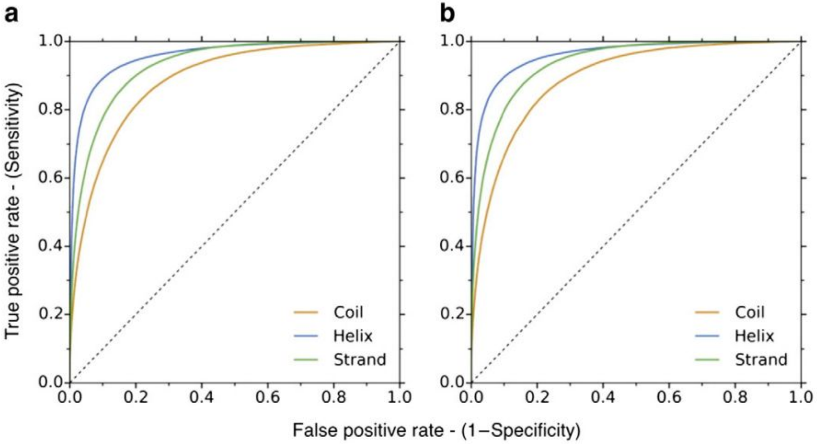
Mean secondary-structure pop

SLFN

Corrections (window=5)

Data sets

B-heavy data set



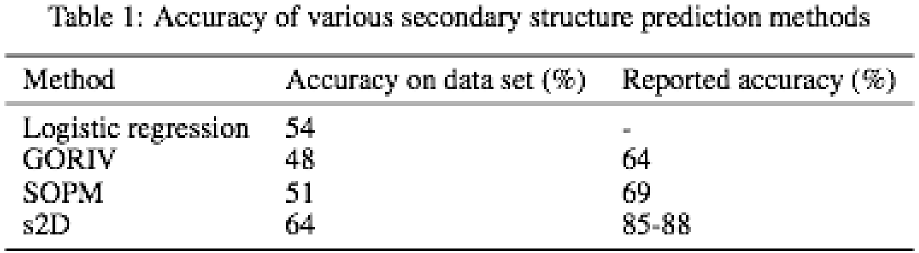
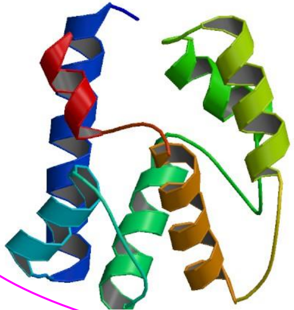
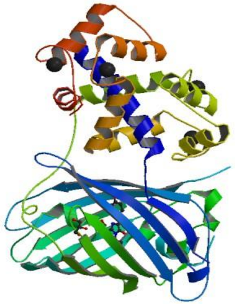


Table 1: Accuracy of various secondary structure prediction methods

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





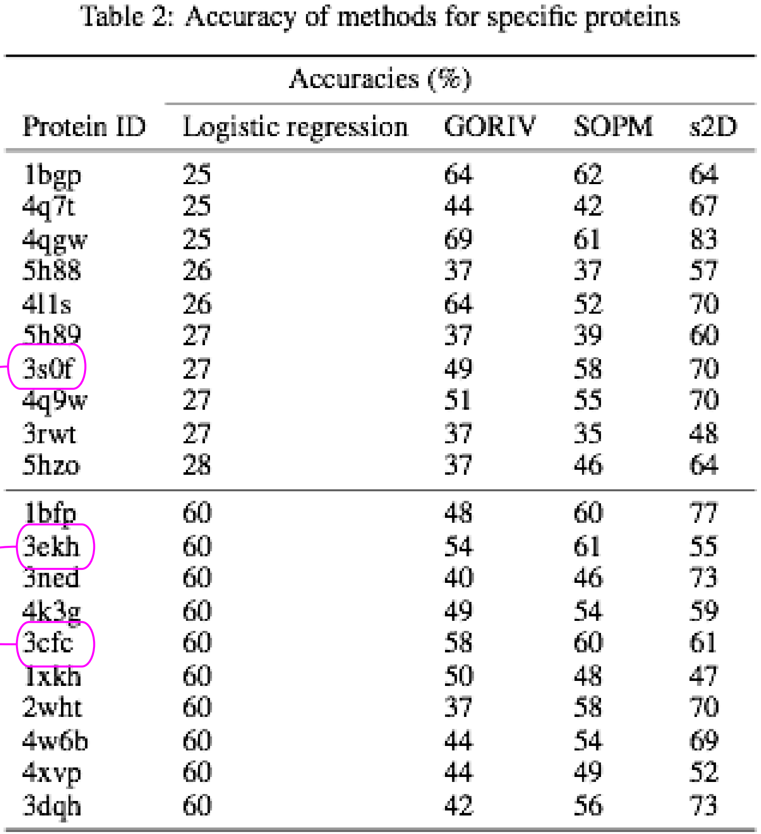
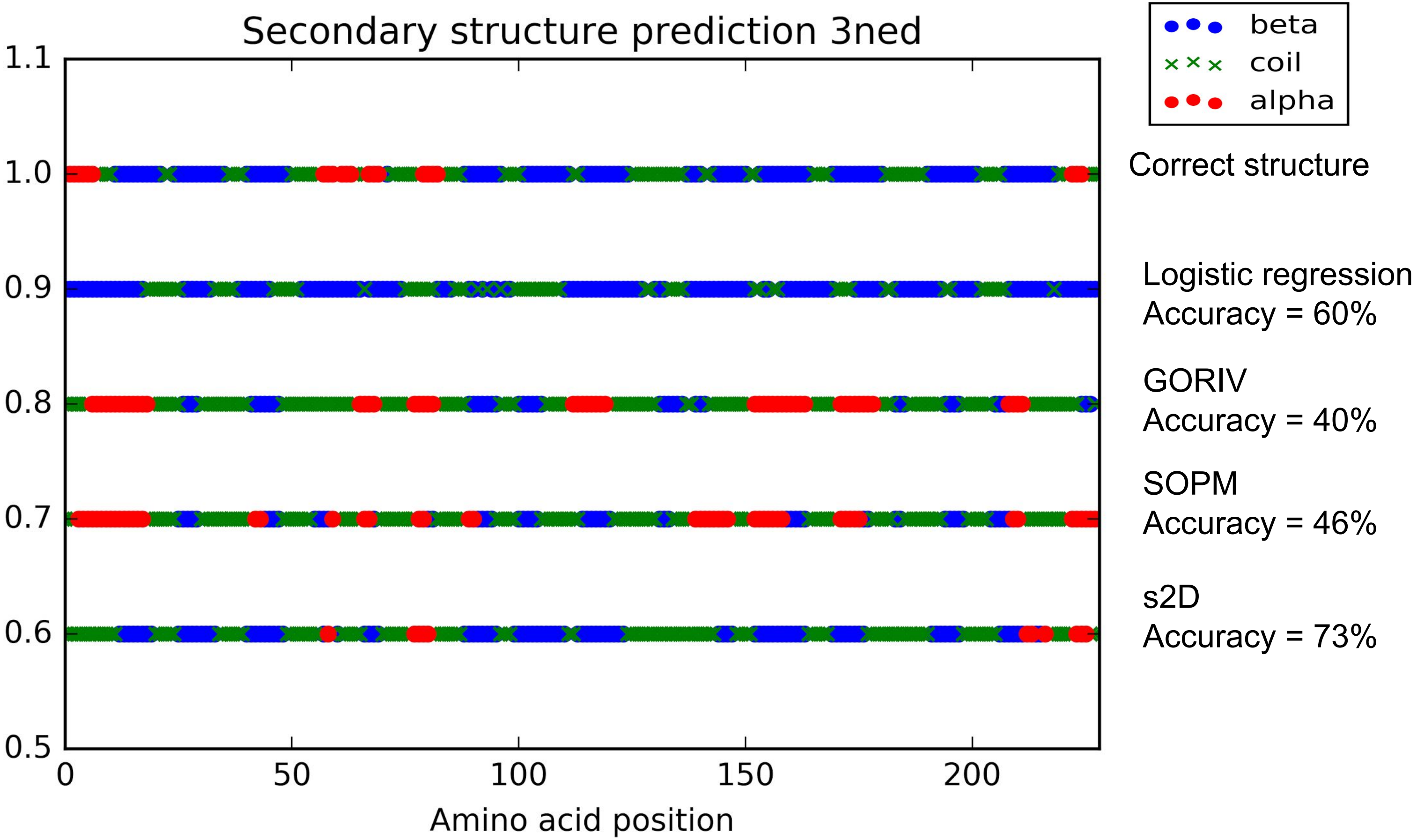


Table 2: Accuracy of methods for specific proteins

	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



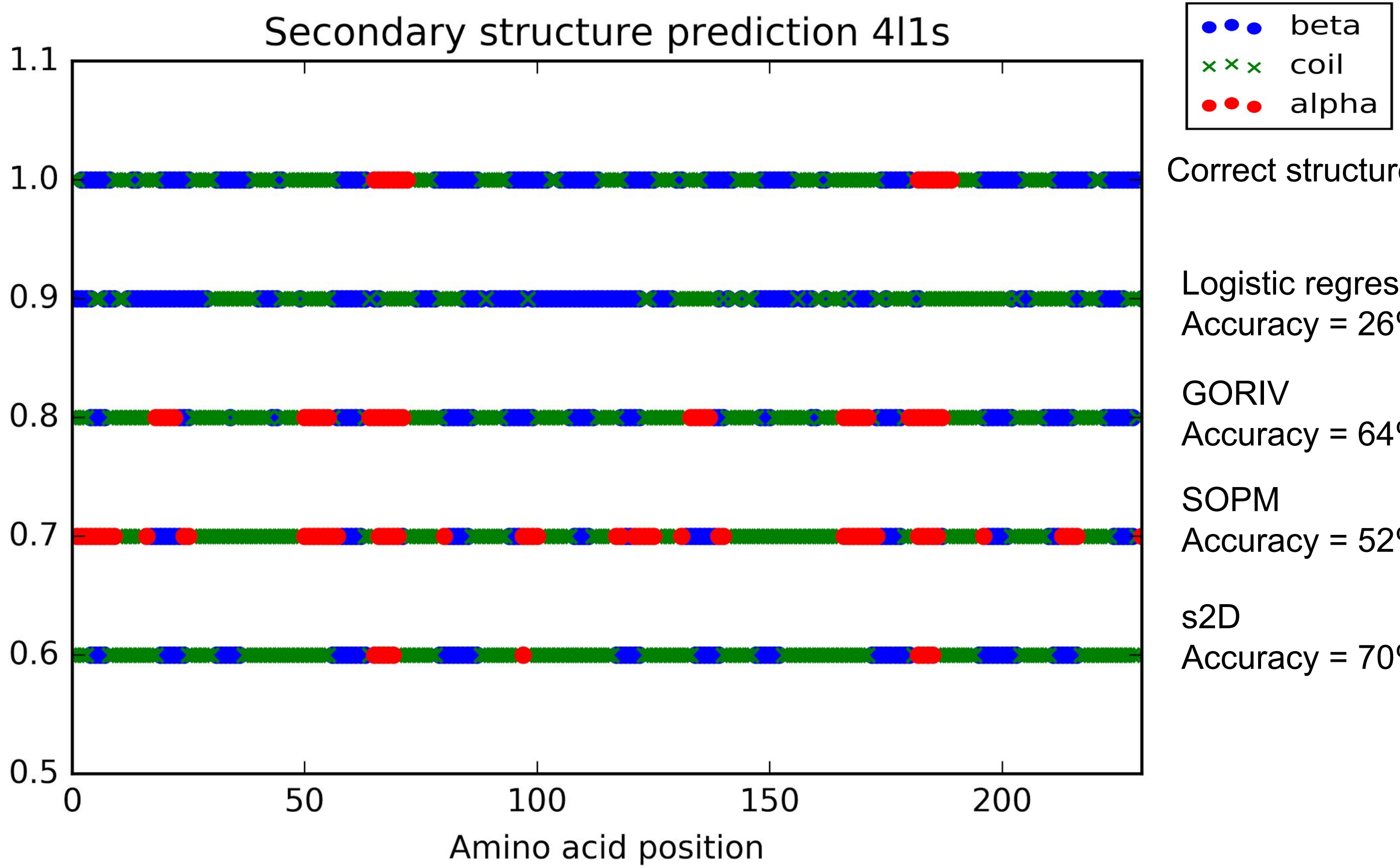
Correct structure

Logistic regression
Accuracy = 60%

GORIV
Accuracy = 40%

SOPM
Accuracy = 46%

s2D
Accuracy = 73%



Correct structure

Logistic regression
Accuracy = 26%

GORIV
Accuracy = 64%

SOPM
Accuracy = 52%

s2D
Accuracy = 70%

Conclusions

For a group of proteins with conserved structures, training on similar proteins is beneficial.

Bias of training set affects generalization.

Helices may require longer range interactions.

Logistic regression on a biased data set did outperform older prediction methods, but neural network methods are still more accurate.

Future Directions

Protein folding

Post transcriptional modifications

Promising techniques with bias training sets

Other methods: HCAM, YASSP

Directed mutations in the lab

Literature Cited

[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 method 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.

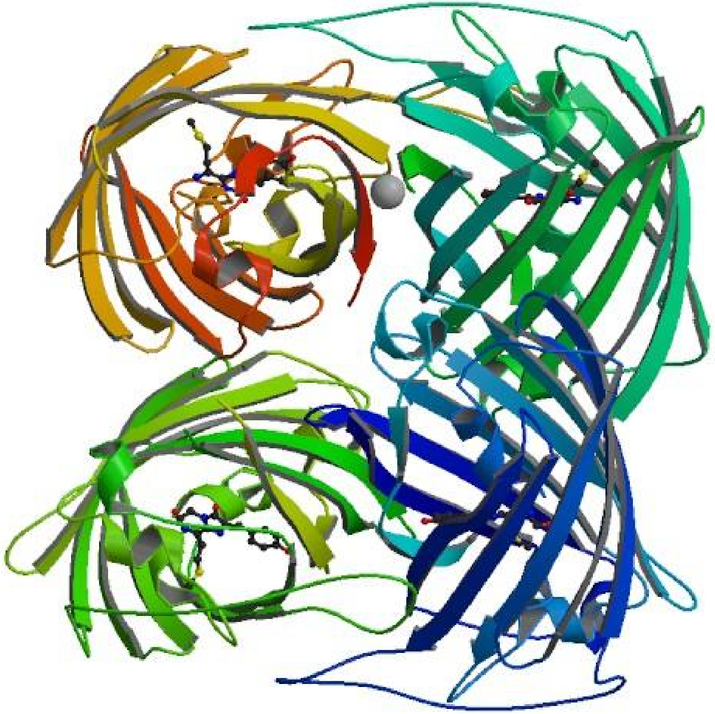
[10] RCSB Protein Data Bank. An Information Portal to 124928 Biological Structures. 739 proteins used, please see additional resources for these proteins and acknowledgments to all the scientists to whom these 739 proteins structures are acknowledged.

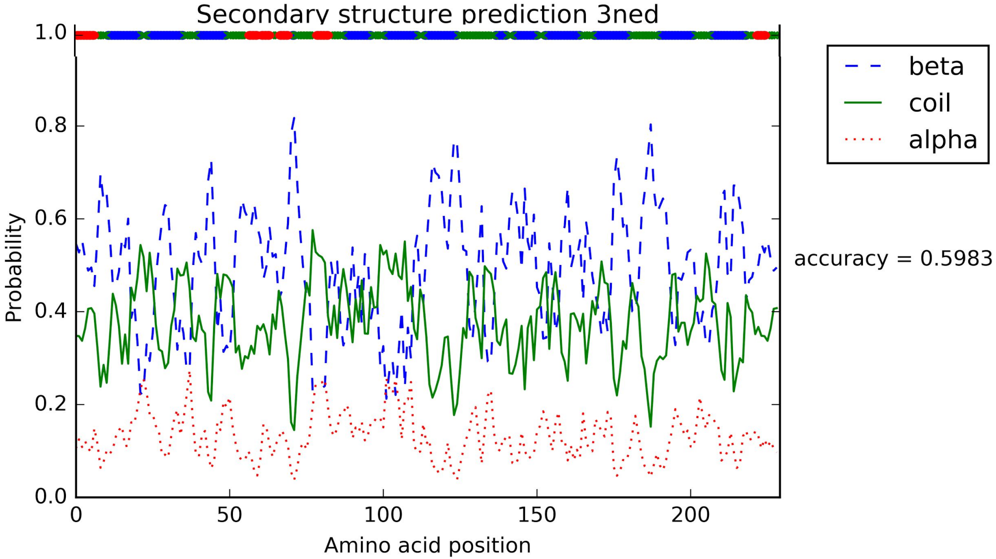
[11] Needleman, S.B. Wunsch, C.\ (1970) J. Mol. Biol. 48, 443-453.

[12] Ulrich, E.L., Akutsu, H. Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J. et al. (2007) BioMagResBank. Nucleic Acid Res 36: D402-8.

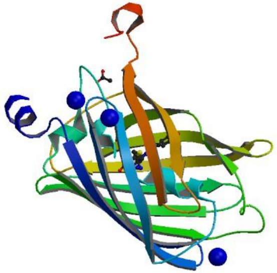
[13] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

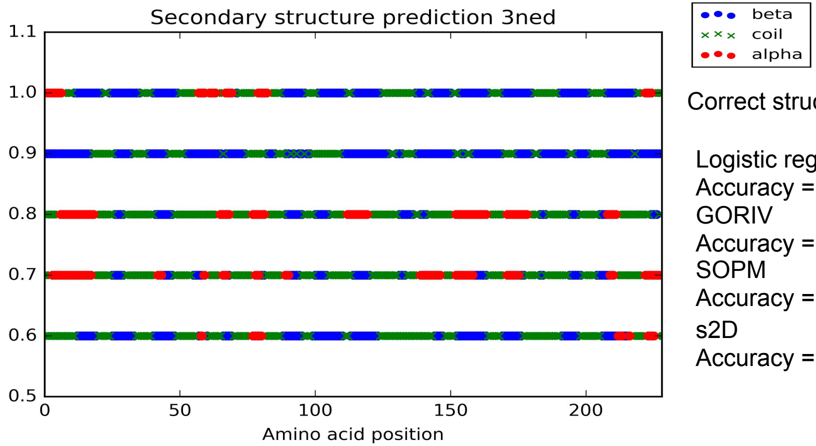
Questions?





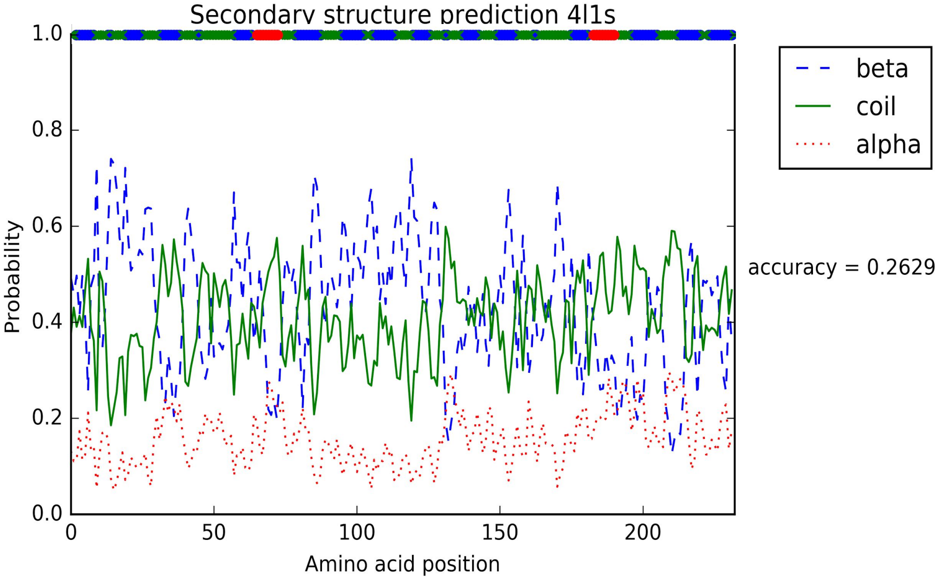
accuracy = 0.5983



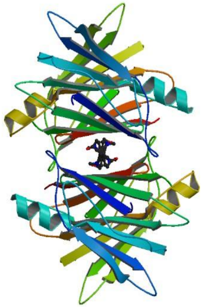


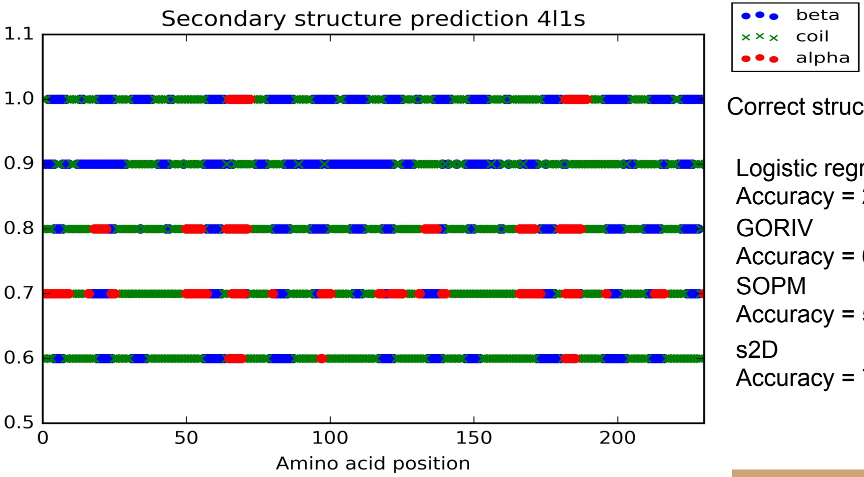
Correct structure

Logistic regression
Accuracy = 60%
GORIV
Accuracy = 40%
SOPM
Accuracy = 46%
s2D
Accuracy = 73%



accuracy = 0.2629





Correct structure

Logistic regression
Accuracy = 26%

GORIV
Accuracy = 64%

SOPM
Accuracy = 52%

s2D
Accuracy = 70%