β-strand segments prediction based on protein sequence and predicted neighboring structural information

Kedarisetti KD, Mizianty MJ, Dick S, and Kurgan LA

Department of Electrical and Computer Engineering, Univ. of Alberta, Edmonton, Canada

Introduction

Existing secondary structure predictors perform relatively poorly on β-strands when compared with the prediction of helices/coils [*1*]. Our analysis of 6 recently published/popular predictors (PROTEUS [2], PSI-PRED [3], SABLE [4], SPINE [5], SSpro [6] and YASPIN [1]) reveals that their SOVe ranges between 61 and 73% and that up to 18% of strand segments are never predicted. Recent works suggest that ensemble-based approaches may provide improvements [7] and show that correlations between neighboring secondary structures are stronger than between neighboring residues [8].

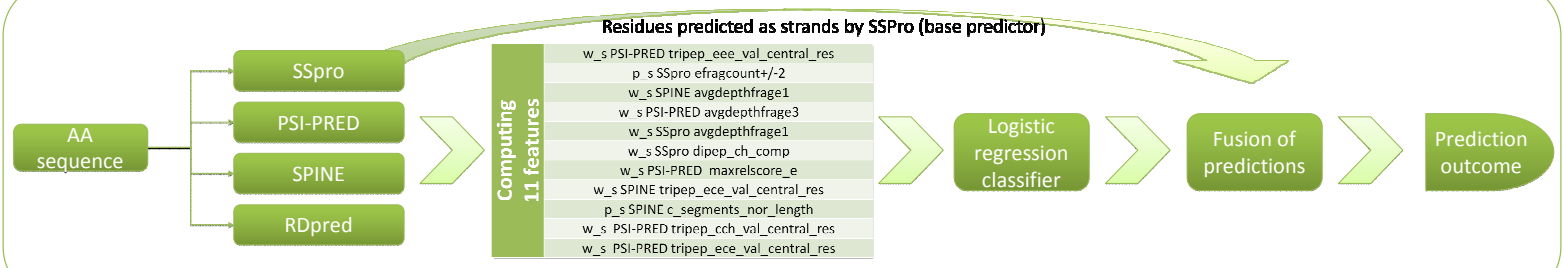

Diagram of the proposed prediction model

Materials & Methods

We propose a novel ensemble-based approach that exploits predicted local and global structural information to predict β-strand residues. Our method is intended to improve the coverage (by finding strands omitted by other methods) and quality (by improving SOVe) of strand predictions when compared with the current secondary structure predictors. We use the primary sequence, secondary structure predicted by SSpro, SPINE and PSI-PRED (three best-performing template-free predictors), and residue depth predicted with RDpred [9] to compute novel features that reveal local structures in the neighborhood of the predicted residue, and global information from the entire sequence. The method generates predictions by feeding a small set of 11 features, which were found by feature selection on a training dataset, as an input to a logistic regression classifier and the predictions are merged with the strand residues predicted by the best performing (on the training set) SSpro.

Experimental comparison between the proposed and competing predictors on the three independent test datasets, Test (432 proteins), CASP8 (118), and CASP8 (8 template free proteins). The strands were considered as found when at least 60 % of residues or one residue were correctly predicted, see columns 3 and 4, respectively.

Dataset	Method	#of strands found by matching at least		SOVe	Q_e observed	Q_e predicted	Accuracy	MCC
		60% of the residues	one residue					
Test	YASPIN	79.9%	82.6%	65.62	72.78	68.31	85.30	0.60
	PROTEUS	**86.5%**	**89.0%**	60.97	**79.96**	71.57	87.51	0.68
	SABLE	76.6%	82.3%	69.41	67.41	79.45	87.94	0.66
	SPINE	79.3%	85.0%	71.02	70.71	79.34	88.50	0.68
	PSI-PRED	80.1%	84.4%	70.88	72.94	77.96	88.51	0.68
	SSpro	79.7%	83.7%	73.12	71.07	**82.02**	**89.27**	0.70
	Proposed model	84.3%	88.0%	**74.60**	71.99	74.31	87.25	**0.71**
CASP8	YASPIN	**84.6%**	86.6%	66.73	**79.07**	73.41	88.12	0.68
	PROTEUS	81.1%	83.7%	61.86	74.89	72.56	87.20	0.66
	SABLE	75.0%	81.9%	67.74	66.56	79.87	87.97	0.66
	SPINE	79.7%	84.8%	68.48	70.81	80.18	88.81	0.68
	PSI-PRED	80.0%	84.5%	66.46	72.43	77.23	88.27	0.67
	SSpro	78.6%	83.2%	67.50	70.75	**82.22**	89.32	0.70
	Proposed model	83.7%	**88.3%**	**72.20**	76.05	79.60	**89.59**	**0.71**
Template free CASP8	YASPIN	**81.80%**	**87.0%**	62.52	**74.36**	76.48	**88.08**	**0.67**
	PROTEUS	68.80%	75.3%	67.22	61.32	78.42	86.35	0.61
	SABLE	71.40%	81.8%	64.80	59.62	82.30	86.93	0.63
	SPINE	75.30%	81.8%	63.94	60.90	**83.09**	87.35	0.63
	PSI-PRED	71.40%	75.3%	58.16	60.47	81.32	86.88	0.63
	SSpro	67.50%	77.9%	63.87	58.55	83.03	86.88	0.63
	Proposed model	80.5%	83.1%	**68.15**	64.32	82.02	87.77	0.65

Results

Tests show that the proposed method achieves SOVe of 74.6% and 72.2%, on 432 low-identity chains from the test dataset (at max pairwise identity of 40% within the test set and between test and training sets) and a set of 118 CASP8 targets, respectively. To compare, best performing secondary structure predictors based on 3-state accuracy, SSpro and SPINE, obtain SOVe of 73.1/71% and 67.5/68.5% on these two datasets, respectively. In addition, our approach misses only 12% and 11.7% of strand segments, while SSpro misses 16.3/16.8% and SPINE misses 15/15.2% of strand segments on the two datasets, respectively. Results for 8 template-free CASP8 proteins are slightly lower, as expected, and show that the proposed model outperforms other considered methods. PROTEUS and YASPIN over-predict strand residues on the test and template-free sets, respectively, see Qe values. When compared with SSpro (our base predictor) the proposed method improves SOVe between 1.5 and 4.7. Our study constitutes a step towards designing an accurate β-strand predictor that would, in the future, facilitate prediction of β-strand residue pairs and β-sheets.

[1] Lin K, Simossis VA, Taylor WR, Heringa J, **A simple and fast secondary structure prediction method using hidden neural networks**. *Bioinformatics* 2005; 21(2):152-159.
[2] Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS, **PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation.** *Nucleic Acids Res.* 2008; 36:W202-9.
[3] Jones DT, **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol.* 1999; 292(2):195-202.
[4] Adamczak R, Porollo A, Meller J, **Combining prediction of secondary structure and solvent accessibility in proteins.** *Proteins* 2005; 59:467-75.
[5] Dor O, Zhou Y, **Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training.** *Proteins* 2007; 66:838-845.
[6] Pollastri G, Przybylski D, Rost B, Baldi P, **Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles.** *Proteins* 2002; 47:228-235.
[7] Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS, **Improving the accuracy of protein secondary structure prediction using structural alignment**. *BMC Bioinformatics* 2006; 7:301.
[8] Crooks GE, Brenner SE, **Protein secondary structure: entropy, correlations and prediction**. *Bioinformatics* 2004; 20(10):1603-11.
[9] Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L, **Sequence based residue depth prediction using evolutionary information and predicted secondary structure**. *BMC Bioinformatics* 2008; 9:388.