Prediction of the Number of Helices for the Twilight Zone Proteins

Kanaka Durga Kedarisetti, Ke Chen, Aashima Kapoor and Lukasz Kurgan*
University of Alberta, Department of Electrical and Computer Engineering
Edmonton, CANADA, T6G 2V4
*lkurgan@ece.ualberta.ca (corresponding author)

***Abstract* – Protein structure prediction is one of the core research areas in bioinformatics. This paper addresses the protein secondary structure prediction problem for the twilight zone proteins, which are characterized by low, about 25% homology to the sets of known sequences. The commonly used sequence alignment based algorithms fail to provide accurate prediction for sequences of such low homology, and thus alternative solutions should be sought. We propose a novel method that aims at the prediction of the number of helical structures based on the twilight zone protein sequences. The method is based on a custom designed and compact feature based sequences representation and applies a decision tree prediction algorithm. The performed experimental study shows superiority of the proposed method over three other prediction algorithms and the results provided by YASPIN algorithm, which is a state-of-the-art alignment based secondary structure prediction method designed using low homology sequences.**

***Keywords:* Protein Secondary Structure, Protein Sequence, Twilight Zone, Helix Prediction.**

I. INTRODUCTION

Knowledge of protein structure is crucial for understanding its functions and the related biological processes. Proteins are composed of amino acid (AA) chains that fold into three dimensional molecules. The experimental methods to recover the tertiary (three dimensional) structure, which include X-ray crystallography and NMR, are relatively expensive, labor intensive, and time consuming. These facts and the growing gap between the number of known proteins vs. the number of proteins for which the structure is known, motivates development of computational methods for the protein structure prediction.

Due to the significant complexity and growing quantity of the protein data, the predictions are performed at various levels of protein structure, including tertiary structure (Bujnicki, 2006), secondary structure (Pollastri et al., 2002; McGuffin and Jones, 2003; Lin et al., 2005) structural class (Wang and Yuan, 2000; Cai et al., 2003; Kurgan and Homaeian, 2006), and secondary structure content (Zhang et al., 2001; Lin and Pan, 2001). The computational approaches can be categorized into:

1. Multiple-sequence alignment based, in which, for a given query sequence, homologous sequences are found and the query sequence's structure is deduced based on the known structure of the homologous sequences.
2. Threading based, which compare a query sequence with a library of known folds. The comparison results in 'similarity' scores, which are ranked, and the structural template with the best score becomes the predicted structure of the query sequence. The main shortcoming of threading methods is that they are unable to recognize previously unencountered structures.
3. Fragment assembly based, which are based on an observation that the protein backbone structure can be accurately represented using short fragments taken from other proteins (Rohl et al, 2004; Kim et al., 2004).

Recent research results in hybrid methods, which combine fragment assembly, lattice based folding simulations and threading (Skolnick et al., 2001; Zhang and Skolnick, 2004), and sequence alignment and threading (Shan et al., 2001; Skolnick et al., 2004) to improve accuracy.

The protein secondary structure prediction is currently dominated by the sequence alignment methods (Cuff and Barton, 2000; Rost and Sander, 2000; Pollastri et al., 2002; Pollastri and McLysaght, 2005, Lin et al., 2005), which are shown to provide superior accuracy (McGuffin and Jones, 2003). CASP studies show that the best prediction results are achieved by methods that are based on sequence alignment and utilize a committee of several prediction methods (Moult et al., 2003). At the same time, the sequence alignment requires at least ~30% homology between the query protein and protein used to predict its structure (Sander and Schneider, 1991). The proteins characterized by lower, 20-30% homology with sequences that are used to predict their structure are called twilight zone proteins (Rost, 1999). More than 95% of all sequence pairs detected in the twilight zone have different structures (Rost, 1999), which significantly impacts quality of the structure prediction. The prediction of the secondary structure for homologous sequences by the state-of-the-art alignment secondary structure prediction methods yields about 80% accuracy (Petersen et al., 2000; Pollastri and McLysaght, 2005). In case of the twilight zone sequences, the accuracy substantially drops to about 65% to 68% (Lin et al., 2005). To compare, early secondary structure prediction methods that date back to late 1970's achieved similar, about 66%, accuracy (Chou and Fasman, 1978).

To this end, this paper addresses an aspect of the secondary structure prediction for the low homology proteins. We propose a novel method for prediction of the number of helical structures based on the sequences of the twilight zone proteins.

Since to the best of our knowledge, there are no existing methods for prediction of the number of helices, we compared the proposed method with a more general, recent method for the secondary structure prediction called YASPIN (Lin et al., 2005). The latter method was designed and tested on low, ≤ 25% sequence homology dataset, which is a superset of the data used in this paper. The proposed prediction method is shown to give prediction results better than those given by the multiple-sequence alignment based YASPIN method and thus can be used to improve quality of secondary structure prediction of the alignment based methods.

Next, the related work is overviewed and the proposed prediction method is contrasted with existing prediction methods.

A. Related Work

The idea of protein secondary structure prediction stems from an observation that short sequence fragments prefer certain local structural arrangements. In general, the tertiary protein structure is defined by the packing of these arrangements (α-helices denoted by H, β-strands denoted by E and coils denoted by C), which constitute the secondary structure. A number of prediction methods related to the secondary structure were proposed:

- methods that directly predict the secondary structure,
- content prediction methods, which predict the percentage amount of the residues that constitute the individual secondary structures, i.e. helices and strands,
- structural class prediction methods, which classify protein chains into corresponding structural classes (α,β,α+β, and α/β) that are defined based on the inclusions of the helices and strands in the structure.

The last two methods provide important information to understand the protein confirmation, and to analyze and define structural and functional similarities between different proteins (Murzin et al., 1995).

The content and structural class predictions are performed based on a substantially different approach when compared with the alignment based secondary structure prediction. Instead of using alignment profiles, these methods convert the protein sequence into a feature based representation and use these vectors in combination with a variety of Machine Learning algorithms to perform the prediction.

The first content prediction effort was undertaken using Multiple Linear Regression (MLR) method and the composition vector based representation of protein sequence (Krigbaum and Knutton, 1973). Later, a number of approaches, which used different combinations of the composition vector, molecular weight, and hydrophobicity based auto-correlation functions to represent sequences and neural networks (Muskal and Kim, 1992; Ruan et al., 2005), analytic vector decomposition technique (Eisenhaber et al., 1996a), and MLR (Zhang et al., 1996; Zhang et al., 1998; Zhang et al., 2001; Lin and Pan, 2001; Kurgan and Homaeian, 2005) were developed.

Early structural class prediction methods again used relatively simple composition vector based sequence representation and applied discriminant analysis with Euclidean distance (Nakashima et al., 1986), Hamming distance (Klein and Delisi, 1986), and Mahalanobis distance (Chou and Zhang, 1994). Next generation prediction methods used more complex classification algorithms based on the maximum component coefficient principle (Zhang and Chou, 1992), least correlation angle algorithm (Chou and Zhang, 1993), fuzzy clustering (Zhang et al., 1995), neural network (Dubchak et al., 1999), vector decomposition (Eisenhaber et al., 1996b), component coupled geometric classification algorithm (Chou and Maggiora, 1998), Bayesian classification (Wang and Yuan, 2000), and support vector machines (Cai et al., 2003). Recent works improve structural class prediction by using alternative sequence representation, which includes auto-correlation functions based on non-bonded residue energy (Bu et al., 1999), polypeptide composition (Luo et al., 2002), functional domain composition (Chou and Cai, 2003), and physiochemical properties (Kurgan and Homaeian, 2006).

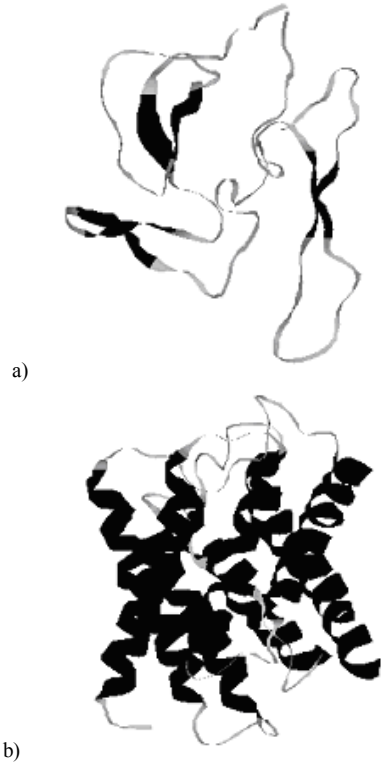
B. Problem Definition and Goals

We propose a novel prediction method that aims at prediction of the number of helical structures using the protein sequence as the input. We chose this particular prediction since helical structures are regular and local, which result in higher accuracy of the corresponding secondary structure prediction, i.e. a recent study that concerns the twilight zone sequences reported about 70% accuracy for helix and 55% accuracy for strand (Lin et al., 2005). In contrast, coils are irregular and strands form long range interactions to form sheets. Knowledge of the number of helices for a given sequence may be useful in assigning structural class, learning topology of the corresponding protein and performing structural alignment studies. Since the proposed method aims at the prediction for the twilight zone sequences, for which alignment does not provide reliable information, we decided to develop our solution based on ideas proposed by authors of the structural class and content prediction methods. Therefore, the method converts the sequence into custom designed feature representation, and uses the feature values as an input to a selected prediction algorithm. For example, for the following primary sequence and secondary structure:

- primary sequence: `GTMLLGMLMICSATEK`
- secondary structure: `CCHHHHHCCHHHHCCC`

the method should predict the number of helical structures as 2. Note that since one helical turn requires at least three AAs, hence the helical structures that include at least three residues are counted.

We illustrate the proposed prediction method using two example proteins: merozoite surface protein 1 from malaria parasite (Protein Data Bank (PDB) (Berman et al., 2000) ID: 1CEJ) and human aquaporin-1 protein (PDB ID: 1FQY). The former protein contains no helices, while the latter contains 8 helices, see Figure 1.



a)

b)

Fig. 1. Ribbon picture of the secondary structure of a) merozoite surface protein 1 (PDB ID: 1CEJ), and b) aquaporin-1 protein (PDB ID: 1FQY). The helices and strands are shown in black, while coils are shown in grey.

The proposed method correctly predicts no helices based on the primary sequence of the merozoite surface protein 1 and 8 helices based on the sequence of the aquaporin-1 protein.

Figure 2 contrasts and integrates the proposed method with other prediction approaches.

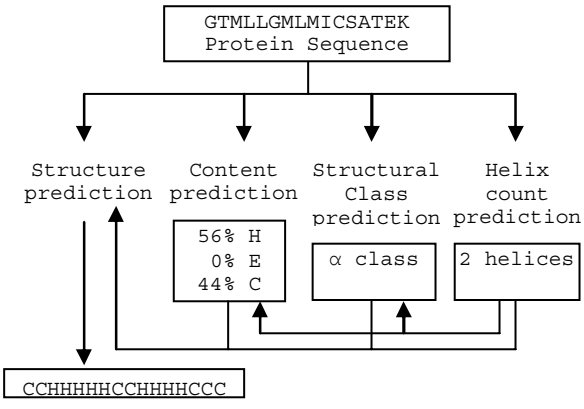


Fig. 2. Integration of the existing and the proposed method for secondary protein structure prediction.

We aim at minimizing the prediction error through optimization of the feature representation, i.e., selection of a small set of well performing features, and careful selection of the best performing prediction algorithm. The main goal is to obtain the method that surpasses the quality of the alignment based secondary structure prediction that can also be used to infer the number of helices. Therefore, the developed method was compared with YASPIN methods, which is the most recent alignment based secondary structure prediction methods that was originally developed and tested on the twilight zone proteins (Lin et al., 2005).

II. Proposed Method

The design and test procedures performed with the proposed method are summarized in Figure 3.

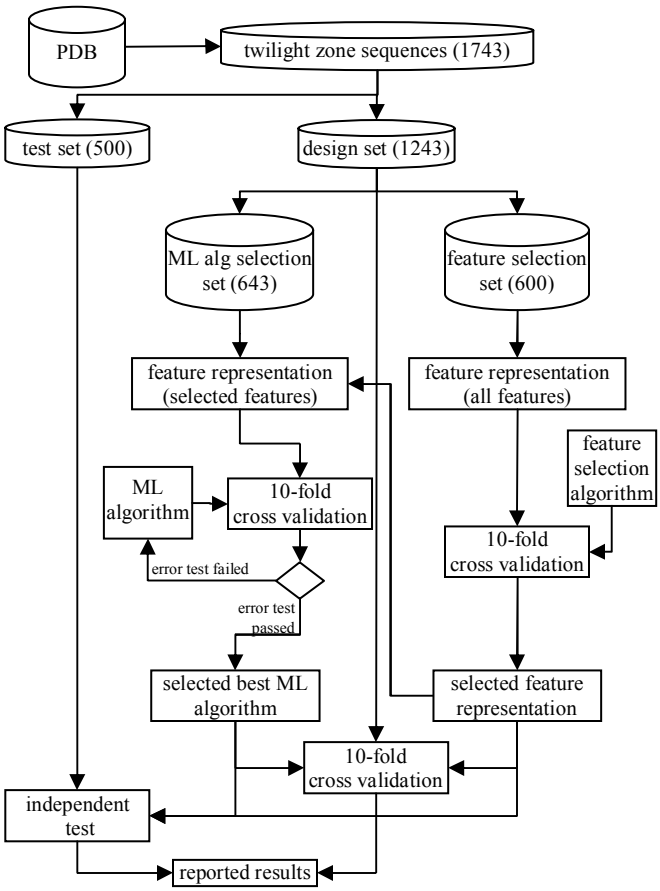


Fig. 3. Design and test procedures for the proposed prediction method.

First, a suitable set of 1743 twilight zone sequences was selected among the protein stored in the PDB. For these sequences, secondary structure was assigned based on DSSP (Dictionary of Secondary Structures of Proteins) method (Kabsch and Sander, 1983), and this information was used to compute the number of helical structures. Next, a randomly selected set of 500 sequences was set aside as an independent test set. The remaining sequences were randomly split into two

sets: 600 were used for design of feature representation, and the other 643 were used for the selection of the optimal learning parameters for the considered prediction algorithms when using the designed representation. The design of sequence representation and optimization of prediction algorithms tasks were performed using 10-fold cross validation tests on disjoint sets of proteins to minimize the negative impact of the learning bias and the overfitting. The resulting prediction approach was tested both on the set of 1243 proteins (using 10-fold cross validation) and on the independent test set (using the optimized algorithm's setting and 1243 protein as the training set). The latter results, which are not biased by the performed design, were compared with the prediction on the independent test set performed by the YASPIN method.

Next, we explain the design procedure.

A. Dataset Preparation

The main goal for the data preparation was to select well defined set of the twilight zone sequences. The dataset was generated using PISCES (Wang and Dunbrack, 2003), a protein sequence culling server that applies combination of structural and sequence (using PSI-BLAST) alignment to limit the sequence homology. We used the cullpdb25 data set as of May 2006, which includes sequences with the maximum of 25% sequence identity measured at 3.0 resolution and with R-factor equal 1. The original data set includes 4127 protein sequences, which were further filtered according to a set of rules defined in Table I to eliminate errors and inconsistencies. The final, filtered dataset includes 1743 sequences.

TABLE I
FILTERS USED TO SELECT HIGH QUALITY SEQUENCES

Filter	# removed sequences
sequences with structure that is incomplete in PDB	2213
sequences that include non-standard AAs	192
sequences that do not have the side chain coordinates, which may cause incorrect DSSP assignment of the secondary structure	76
sequences with less than 20 AAs	72

B. Sequence Representation Design

The transformation of the protein sequence into a feature space is performed by utilizing various physiochemical properties of AA, which were drawn from the past research studies. The related features are grouped into sets. Table II presents the considered features sets and points to references that motivated our choice and that give further details. The resulting 66 features, which are grouped into 8 feature sets, were fed into a feature selection method to design the sequence representation for the proposed prediction method. The Feature Subset Consistency (FSC) (Liu and Setiono, 1996) method, which aims at maximizing consistency of the target class values (helix counts) when the training instances are projected onto the feature subset, was applied on the 600 sequences using 10-fold cross validation. The final feature representation consists of the features sets that include at least 25% of features that were selected by the FSC method in at least 5 out of 10 cross validation folds, and thus provide significant information for the prediction. The results are summarized in Table III (the selected feature sets are underlined and bolded). Four feature sets that include total of 28 features were selected.

TABLE II
FEATURE SETS CONSIDERED FOR THE SEQUENCE REPRESENTATION

Feature set with description	Abbr.	# features	reference
Sequence length - # of AA residues in protein sequence (related to content prediction)	L	1	(Muskal and Kim, 1992; Syed and Yona, 2003)
Composition vector-percentage of each AA in the protein sequence (most common attribute considered in content and structural class predictions)	CV	20	(Krigbaum and Knutton, 1973; Muskal and Kim, 1992; Eisenhaber et al., 1996a; Zhang et al., 1996; Zhang et al., 1998; Zhang et al., 2001; Ruan et al., 2005; Kurgan and Homaeian, 2005)
1st order composition moment vector -composition vector that incorporates position of AA in the sequence (improves content prediction)	CMV	20	(Ruan et al., 2005; Kurgan and Homaeian, 2005; Kurgan and Homaeian, 2006)
R group-divides AAs into nonpolar aliphatic, polar uncharged, aromatic, positively and negatively charged (used to measure protein concentration)	RG	5	(Nelson and Cox, 2000)
Exchange group - divided AAs based on their structure-conserved mutations (describe replacements through evolution)	EXG	3	(Wang et al., 2000; Yang and Wang, 2003)
Electronic group- divides AAs into neutrals, electron donors or acceptors (describe electrostatic forces that stabilize the structure)	EG	5	(Ganapathiraju et al., 2004)
hydrophobic group -divides AAs into hydrophobic and hydrophilic (hydrophobic force is one of the strongest determinant factor of a protein structure)	HG	2	(Hobohm and Sander, 1995; Syed and Yona, 2003)
Auto-correlations-auto-correlations based on hydrophobic indices(describes correlation of hydrophobic profile along a protein sequence)	AC	10	(Zhang et al., 2001; Lin and Pan, 2001)

TABLE III
FEATURE SELECTION RESULTS

Feature Set	Total # of features	# features selected in at least 5-folds
Sequence length	**1**	**1**
Composition vector	**20**	**6**
1st order composition moment vector	20	5
R group	**5**	**3**
Exchange group	3	0
Electronic group	5	1
Hydrophobic group	**2**	**2**
Auto-Correlations	10	2

The selected features describe the length of the sequence, its composition, and several physiochemical properties, such as polarity, charge, aromaticity, and hydrophobicity, of AAs in the sequence. These features extend the most commonly used representation that considers only the sequence composition (see section I.A).

C. Optimization of the Prediction Algorithms

The 643 sequences transformed into the 28-features based representation were used to optimize prediction algorithms. Four major prediction algorithms types were considered:

- IBk, which is a k-nearest neighbor algorithm (Aha and Kibler, 1991)
- Decision table algorithm that generates production rule sets (Kohavi, 1995)
- M5, which is a decision tree based algorithm that can handle continuous class features (Wang and Witten, 1997)
- Support vector regression algorithm (Shevade et al., 2000)

WEKA platform was used to perform prediction experiments (Witten and Frank, 2005).

Similarly to (Zhang et al., 2001; Lin and Pan, 2001; Kurgan and Homaeian, 2005) the quality of the prediction was measured based on the absolute average error (*error*) and the corresponding standard deviation (*stdev*):

$$error = \sum_{i=1}^{N} \frac{|\, aHc_i - pHc_i \,|}{N}$$

$$stdev = \sqrt{\sum_{i=1}^{N} \frac{(|\, aHc_i - pHc_i \,| - error)^2}{N}}$$

where N is the number of test sequences, and aHc_i and pHc_i are the actual and predicted number of helices for the i^{th} test sequence, respectively. We note that the root mean squared error could be used, but we decided to use the same criteria as in the content prediction field. We caution the reader that the *error* values are much smaller than the root mean squared error values.

The 10-fold cross validation test was used, and the error was minimized by tweaking the learning parameters for each of the algorithms. Negative predictions were rounded up to zero and all positive predictions were rounded to the nearest integer.

III. Experimental Evaluation

Two sets of experimental results were performed

- 10-fold cross validation test on the set of 1243 sequences that share less that 25% homology with each other. This test was performed for each of the four optimized prediction methods and both 66 and 28 feature based representations.
- Test on the independent test set that includes 500 sequences that share less than 25% homology with each other and with sequences in the training set. The four optimized prediction algorithms were trained with the 1243 sequences that were represented using the selected 28 features and the entire set of 66 features, and tested on the 500 sequences.

These tests allow evaluating the impact of the feature selection and the selection of the best prediction algorithm. They were also compared with results of YASPIN method to verify if the proposed method can improve the results provided by alignment based secondary structure prediction algorithms. In the latter case, the independent test set with 500 sequences was fed into the PSI-BLAST to generate PSSM profiles, and these profiles were utilized by the YASPIN method to predict the secondary structures. Next, the predicted structures were used to compute the number of helices and the corresponding error and standard deviation were reported. We note that the original YASPIN implementation that was trained with a large set of 3553 sequences was used (Lin et al., 2005). Although these sequences also shared low, 25% homology, this training set is almost three times larger than the training set used by the proposed method, and thus we note that the YASPIN method's results may be overestimated when compared to our results.

Prediction results for the 10-fold cross validation tests on the training data with 1243 sequences are summarized in Table IV.

TABLE IV
10-FOLD CROSS VALIDATION RESULTS ON THE SET OF 1243 PROTEINS

Prediction algorithm	*error* (*stdev*)	
	using all 66 features	using selected 28 features
M5	**1.62 ±1.66**	**1.62 ±1.64**
Support vector regression	**1.61 ±1.69**	**1.59 ±1.64**
Decision Table	1.87 ±1.84	1.79 ±1.83
IBk	1.75 ±1.85	1.74 ±1.86

The results show that M5 and support vector regression methods provide superior predictions (see bolded values in Table IV). The results also show that the sequence representation based on the selected 28 features provides the same or slightly better predictions when compared with the full set of 66 features.

Prediction results for the independent set of 500 sequences are summarized in Table V.

TABLE V
PREDICTION TEST RESULTS ON THE INDEPENDENT SET OF 500 PROTEINS

Prediction algorithm	*error* (*stdev*)	
	using all 66 features	using selected 28 features
M5	**1.57 ±1.54**	**1.53 ±1.53**
Support vector regression	1.66 ±1.68	1.59 ±1.60
Decision Table	1.98 ±1.90	1.94 ±1.89
IBk	1.78 ±1.80	1.74 ±1.79
YASPIN	1.74 ±1.73	

The results again confirm superiority of both M5 and support vector regression methods, although in this test the former one gives smaller errors (see bolded values in Table V). The best, M5, algorithm is also characterized by the lowest standard deviation. The selected sequence representation results in slight improvements in the error values, while it is substantially more compact than the original representation that includes 66 features.

Most importantly, the proposed method performs substantially better than the alignment based YASPIN algorithm. The difference is the performance between the M5 algorithm and the YASPIN method is visualized in Figure 4.

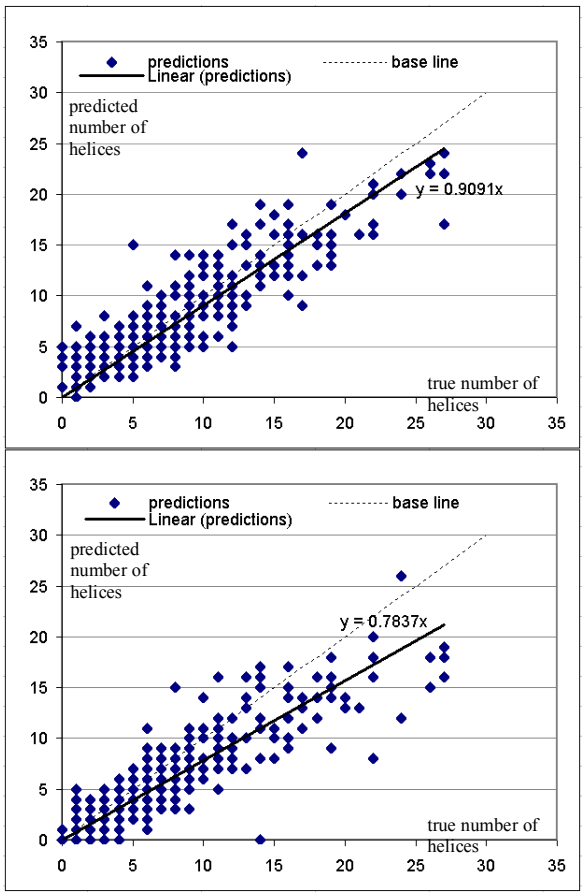


Fig. 4. Scatter plots that show predicted vs. true number of helices for the M5 algorithm (upper plot) and the YASPIN method (lower plot).

This Figure shows a scatter plot of the actual vs. predicted helix counts for both methods. The dotted line represents the perfect results, while the diamond shaped markers show the predictions. The solid line shows linear regression of the predictions. The plots shows that YASPIN method more significantly underestimates the number of helical structures when compared to the results of the proposed method, i.e., the corresponding linear regression coefficient equal 0.78 and 0.91. The results from the proposed method are clearly clustered around the diagonal, while the YASPIN's predictions for sequences with the large number of helices suffer more substantial errors.

In short, the scatter plots reveal that the proposed method provides a high quality alternative for prediction of the number of helical structures in the twilight zone protein sequences.

IV. Summary and Conclusions

This paper presents a novel system for the computational prediction of the number of helical structures in the twilight zone protein sequences, which are characterized by low, below 25%, homology.

The system performs prediction based on a novel and compact feature based sequence representation and uses a decision tree based prediction algorithm. Experimental comparison of the proposed method with three other prediction algorithms and prediction results based on a state-of-the-art alignment based secondary structure prediction algorithm show superiority of the former method. We note that the presented method can be further improved, e.g. by trying other alternative feature sets to represent the sequences and by using boosting and classifier fusion techniques. At the same time, the main strength of the proposed method is its simplicity and user-friendliness, in terms of a compact and easy to compute sequence representation, fast execution time and application of a popular, and thus easy to obtain and use, prediction algorithm.

The results from the proposed prediction methods provide useful information that can be utilized to improve accuracy of the alignment based secondary structure prediction methods for low homology sequences, including the YASPIN method. They can be also used as input to the protein structural class and content prediction methods, and in investigations related to finding evolutionary relationships between protein structures and sequences.

The proposed method can be extended to predict number of strands, which will constitute our future work. We anticipate that the feature selection and optimization of the prediction algorithms should be repeated to assure acceptable accuracy.

Acknowledgment

We would like to thank the authors of the YASPIN method for providing their software. KDK, KC and LK gratefully acknowledge support provided by NSERC Canada. KC was also partially supported by MITACS Canada under the Internship program.

References

1. Aha, D., and D. Kibler, Instance-based learning algorithms", *Machine Learning*, 6, 37-66, 1991
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P., Protein Data Bank, *Nucleic Acids Research*, 28, 235-242, 2000
3. Bu WS, Feng ZP, Zhang Z, and Zhang CT, Prediction of Protein (Domain) Structural Classes Based on Amino Acid Index, *European Journal of Biochemistry*, 266, 1043-1049, 1999
4. Bujnicki J., Protein-structure Prediction by Recombination of Fragments, *Chembiochem*, 7:1, 19-27, 2006
5. Cai YD, Liu XJ, Xu XB, and Chou KC, Support Vector Machines for Prediction of Protein Domain Structural Class, *Journal of Theoretical Biology*, 221, 115-120, 2003
6. Chou P.Y., and Fasman G.D., Prediction of the Secondary Structure of Proteins from Their Amino Acid Sequences, *Advances in Enzymology*, 47, 45-148, 1978
7. Chou KC and Zhang CT, A New Approach to Predicting Protein Folding Types, *Journal of Protein Chemistry*, 12, 169–178, 1993
8. Chou KC and Zhang CT, Predicting Protein-folding Types by Distance Functions that Make Allowances for Amino-acid Interactions, *Journal of Biological Chemistry*, 269, 22014-22020, 1994
9. Chou, KC, and Maggiora, GM, Domain structural class prediction, *Protein Engineering*, 11, 523–538, 1998
10. Chou KC, and Cai YD, Prediction Protein Structural Class by Functional Domain Composition, *Biochemical and Biophysical Research*

Communications, 321, 1007-1009, 2004
11. Cuff J. and Barton G., Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction, *Proteins*, 40, 502–511, 2000
12. Dubchak I, Muchnik I, Mayor C, Dralyuk I and Kim SH, Recognition of a Protein Fold in the Context of the SCOP Classification, *Proteins*, 35, 401-407, 1999
13. Eisenhaber, F., Imperiale F, Argos P, and Frommel C., Prediction of Secondary Structural Contents of Proteins from Their Amino Acid Composition Alone, I. New Analytic Vector Decomposition Methods, *Proteins*, 25:2, 157-168, 1996a
14. Eisenhaber F, Frömmel C, and Argos P., Prediction of secondary structural content of proteins from their amino acid composition alone, II The paradox with secondary structural class, Proteins, 25, 169–179, 1996b
15. Ganapathiraju, M, Klein-Seetharaman, J, Balakrishnan, N and Reddy, R, Characterization of Protein Secondary Structure Using Latent Semantic Analysis, *IEEE Signal Processing Magazine*, 15, 78-87, 2004
16. Hobohm, U., and Sander, C., A Sequence Property Approach to Searching Protein Databases, *Journal of Molecular Biology*, 251, 390-399, 1995
17. Kabsch, W. and Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22, 2577-2637, 1983
18. Kim, D. E., Chivian, D., and Baker, D., Protein Structure Prediction and Analysis Using the Robetta Server, *Nucleic Acids Research*, 32, W526-531, 2004
19. Klein P and Delisi C, Prediction of Protein Structural Class from the Amino-acid Sequence, *Biopolymers*, 25, 1659-1672, 1986
20. Krigbaum, W. R., and Knutton, S. P., Prediction of the Amount of Secondary Structure in a Globular Protein from its Amino Acid Composition, *Proceedings of the National Academy of Science*, 70, 2809-2813, 1973
21. Kurgan, L. and Homaeian, L., Prediction of Secondary Protein Structure Content from Primary Sequence Alone - a Feature Selection Based Approach, *Proceedings of the 2005 International Conference on Machine Learning and Data Mining in Pattern Recognition*, 334-345, 2005
22. Kurgan, L. and Homaeian, L., Prediction of Structural Classes for Protein Sequences and Domains - Impact of Prediction Algorithms, Sequence Representation and Homology, and Test Procedures on Accuracy, *Pattern Recognition*, special issue on Bioinformatics, available online 11 April 2006
23. Kohavi. R , The power of decision tables*, Proceedings of the 8th European Conference on Machine Learning* (ECML'95), 174-189, 1995
24. Lin Z and Pan X-M, Accurate Prediction of Protein Secondary Structural Content, *Journal of Protein Chemistry*, 20:3, 217-220, 2001
25. Lin K, Simossis VA, Taylor WR, and Heringa J., A Simple and Fast Secondary Structure Prediction Method using Hidden Neural Networks, *Bioinformatics*, 21:2, 152-9, 2005
26. Liu H and Setiono R, A Probabilistic Approach to Feature Selection - A Filter Solution, *Proceedings of the 13th International Conference on Machine Learning*, 319-327, Italy, 1996
27. Luo R, Feng Z, and Liu J, Prediction of Protein Structural Class by Amino Acid and Polypeptide Composition, *European Journal of Biochemistry*, 269, 4219-4225, 2002
28. McGuffin L and Jones D., Benchmarking Secondary Structure Prediction for Fold Recognition, *Proteins*, 52:2, 166-75, 2003
29. Moult J. et al., Critical assessment of methods in protein structure prediction (CASP) – Round V, Proteins, 53, 334-339, 2003.
30. Murzin A. G., Brenner S. E., Hubbard T., and Chothia C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, 247, 536-540, 1995
31. Muskal, S.M., and Kim, S-H., Predicting Protein Secondary Structure Content: a Tandem Neural Network Approach, *Journal of Molecular Biology*, 225, 713-727, 1992
32. Nakashima H, Nishikawa K, and Ooi T, The Folding Type of a Protein is Relevant to the Amino Acid Composition, *Journal of Biochemistry*, 99, 153–162, 1986
33. Nelson, D., and Cox, M., *Lehninger Principles of Biochemistry*, Worth Publishers, 2000
34. Petersen T. et al., Prediction of Protein Secondary Structure at 80% Accuracy, *Proteins*, 41, 17-20, 2000
35. Pollastri,G., Przybylski,D., Rost,B. and Baldi,P., Improving the Prediction of Protein Secondary Structure in Three and Eight Classes using Recurrent Neural Networks and Profiles. *Proteins*, 47, 228–235, 2002
36. Pollastri G and McLysaght A., Porter: A New, Accurate Server for Protein Secondary Structure Prediction, Bioinformatics, 21:8, 1719-1720, 2005
37. Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D., Protein Structure Prediction using Rosetta, *Methods Enzymol*, 383, 66-93, 2004
38. Rost B., Twilight Zone of Protein Sequence Alignments, *Protein Engineering*, 12, 85-94, 1999
39. Ruan, J., Wang, K., Yang, J., Kurgan, L.A., Cios, K.J., Highly Accurate and Consistent Method for Prediction of Helix and Strand Content from Primary Protein Sequences, Artificial Intelligence in Medicine, special issue on Computational Intelligence Techniques in Bioinformatics, 35:1-2, 19-35, 2005
40. Sander, C. and Schneider, R., Database of homology-derived structures and the structural meaning of sequence alignment, *Proteins*, 9, 56-68, 1991
41. Shan Y.B. Wang G.L. and Zhou H.X., Fold Recognition and Accurate Query-template Alignment by a Combination of PSI-BLAST and Threading, *Proteins*, 42, 23–37, 2001
42. Shevade, SK., Keerthi, SS., Bhattacharyya, C., and Murthy, KRK, Improvements to the SMO algorithm for SVM regression, *IEEE Transaction on Neural Networks*, 11:5, 1188-1183, 2000
43. Skolnick J., Kolinski A., Kihara D., Betancourt M.R., Rotkiewicz P. and Boniecki M., Ab initio Protein Structure Prediction via a Combination of Threading, Lattice Folding, Clustering, and Structure Refinement, *Proteins*, 5, 149-156, 2001
44. Skolnick J. Kihara D. and Zhang Y., Development and Large Scale Benchmark Testing of the PROSPECTOR 3_0 Threading Algorithm, *Proteins*, 56, 502–518, 2004
45. Syed, U., and Yona, G., Using a Mixture of Probabilistic Decision Trees for Direct Prediction of Protein Function, *Proceedings of RECOMB 2003*, 224-234, 2003
46. Wang, J., et al., Application of Neural Networks to Biological Data Mining: a Case Study in Protein Sequence Classification, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 305-309, 2000
47. Wang G. and Dunbrack, R.L. Jr., PISCES: a protein sequence culling server, *Bioinformatics*, 19:1589-1591, 2003
48. Wang Z-X, and Yuan Z, How Good is the Prediction of Protein Structural Class by the Component-Coupled Method?, *Proteins*, 38, 165-175, 2000
49. Wang Y., and Witten, I.H., Inducing Model Trees for Continuous Classes, *Proceedings of the 9th European Conference on Machine Learning* (ECML'97), 128-137, 1997
50. Witten I.H. and Frank E., *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005
51. Yang, X., and Wang, B., Weave Amino Acid Sequences for Protein Secondary Structure Prediction, *Proceedings of the 8th ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery*, 80-87, 2003
52. Zhang CT and Chou KC, An Optimization Approach to Predicting Protein Structural Class from Amino-acid Composition, *Protein Science*, 1, 401-408, 1992
53. Zhang CT, Chou KC, and Maggiora GM, Predicting Protein Structural Classes from Amino Acid Composition: Application of Fuzzy Clustering, *Protein Engineering*, 8, 425-435, 1995
54. Zhang, C.T., Zhang, Z., and He. Z., Prediction of the Secondary Structure of Globular Proteins Based on Structural Classes, *Journal of Protein Chemistry*, 15, 775-786, 1996
55. Zhang, C.T., et al., Prediction of Helix/Strand Content of Globular Proteins Based on Their Primary Sequences, *Protein Engineering*, 11:11, 971-979, 1998
56. Zhang, Z.D., Sun, Z.R., and Zhang, C.T., A New Approach to Predict the Helix/Strand Content of Globular Proteins, *Journal of Theoretical Biology*, 208, 65-78, 2001
57. Zhang Y. and Skolnick J., Automated structure prediction of weakly homologous proteins on a genomic scale, *Proceedings of the National Academy of Science*, 101, 7594-7599, 2004