Secondary Protein Structure Fragments – Feasibility Study in Prediction and Analysis

Lukasz A. Kurgan and Kanaka D. Kedarisetti
University of Alberta, department of Electrical and Computer Engineering, Edmonton, CANADA
{lkurgan, kanaka}@ece.ulberta.ca

Abstract

The prediction of secondary protein structure is one of very active research subjects in computational biology. It aims to provide computational for prediction of secondary protein structure from the primary sequence. This paper performs feasibility study for development of a novel method for the prediction. The main aim is to provide a method that breaks current 88% accuracy limit of the third generation prediction methods.

The study uses novel and comprehensive feature representation for primary sequences, and performs prediction of secondary structure for uniform, in terms of the secondary structure, protein fragments. It considers a wide range of state of the art classification systems on a large and high quality protein dataset extracted from the Protein Data Bank.

The experimental results indicate that the multiple layer perceptron neural networks and boosted decision trees achieve best and significantly better results, when compared to 6 other classification systems. The accuracy limit of the develop solution is 72%, which prevent from applying these results for a system that will break the 88% accuracy limit. At the same time the results show high specificity of about 85%, which indicates that the generated models are very selective, and further improvements are possible and will be pursued. We also discovered an interesting finding that shows that higher accuracy is achieved for the protein fragments closer to the protein head, indicating possible importance of the subsequence position when predicting the secondary sequence. We also note that more evidence should be collected to further substantiate the claim.

Keywords: *Computational Biology, Bioinformatics, Protein Secondary Structure, Protein Primary Structure, Protein Structural Class, Three-State Protein Structure, Classification*

1 Introduction

One of important and heavily explored problems in computational biology is computational prediction and analysis of protein structure. In general proteins have a complex three-dimensional (tertiary) structure. Ability to know and analyze it is the key to understanding biological functions of proteins. The tertiary structure can be learned by experimental approaches, such as X-ray crystallography and NMR (Engel, 1982). These methods are expensive, tedious, and impossible to perform for some proteins, and therefore computational approaches gain their momentum (Ganapathiraju et al., 2004). The computational methods use primary amino acid sequence and predict two dimensional (secondary) structures as an intermediate step to tertiary structure prediction. Their development is further motivated by large number, counted in millions, of currently publicly known primary proteins structures, compared to only about 30 thousands of known tertiary structures, which are stored in Protein Data Bank (PDB) (Berman et al., 2000).

This paper focuses on performing feasibility study to design a novel approach to predict two dimensional protein structures. The main difference between numerous existing secondary structure prediction methods and the new method lays in fundamental methodology and goals. The existing methods aim to predict secondary structure for entire protein and currently achieve up to 80% accuracy. The new method focused on predicting secondary structure for some parts of the protein called fragments, aiming to obtain very high accuracy of possibly over 90%, and leaving the remaining parts of the proteins unrecognized. The highly reliable information about secondary structure of protein fragments will provide invaluable help to predict the remaining parts of the protein with high accuracy using existing methods. Both approaches are compared in Figure 1.

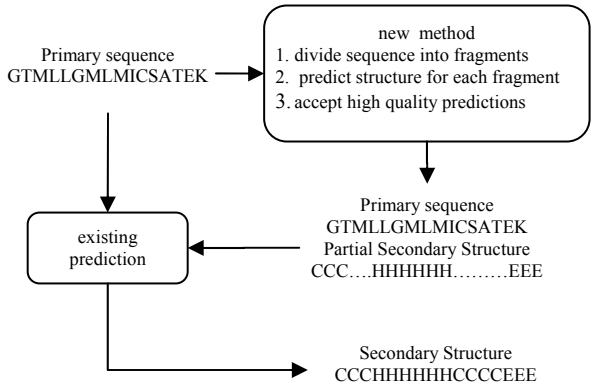

Figure 1. The comparison between the existing and the proposed method for secondary protein structure prediction.

The new method applies divide and conquer approach to increase the secondary prediction accuracy. The prediction of secondary structure for fragments is performed in three steps. First, a given primary sequences is cut into fragments. Second, for each fragment the secondary structure is predicted. Third, the predicted structure is kept only for predictions characterized by high confidence.

The first step is performed by utilizing hydrophobic information about the primary protein structure. It was originally proposed by Dr. Ruan from Nankai University, and currently is in preparation for publication. The method allows dividing the protein sequences into a set of disjoint fragments. The characteristic feature of each fragment is that it has uniform secondary structure, and therefore it is called *uniform fragment*. Some other alternative methods to divide the protein into a set of uniform fragments are also under development, but are outside of the scope of this paper. This paper assumes that the protein primary sequence is divided into set of uniform fragment, and performs feasibility study of the second step to verify if high accuracy classification of the fragments into their corresponding secondary structures can be performed. The paper also performs additional experiments that aim to shed more light into the relation between primary sequence fragments for given secondary structures and the quality of their secondary structure prediction. The results of this study will be used in the design of the third step.

1.1 Related Work

The secondary protein structure prediction is used as an important intermediate step for predicting tertiary structure, protein function, and protein structural change, as well as for computer-assisted molecular design (Truhlar et al., 1999). The molecular design is a basis for rational drug design, and development of novel treatments for diseases such as cancer, cystic fibrosis, and autoimmune disorders. The Dictionary of Secondary Structures of Proteins annotates each amino acid that constitutes the primary structure as belonging to one of seven secondary structure types, which are typically reduced to three states: helix (H), strand (E), and coil (C) (Engle, 1982). Secondary structure of a protein refers to the folding of the chain of amino acids in the three states.

Predicting protein secondary structure using computational approaches has over 30 years of history. Different prediction methods have been developed and continue to improve prediction accuracy, from early results of about 60% accuracy to state of the art algorithms that achieve about 80% accuracy (Rost, 2001). The first generation prediction methods were based on single amino acid propensities (Chou and Fasman, 1978; Garnier et al., 1978). Second-generation prediction methods are based on 3-51 adjacent residues propensities (Gibrat et al., 1987; Rost and Sander, 1994a; Rost et al., 1994b). The third generation prediction methods use evolutionary information and large protein databases to consider global properties associated with protein families. They use position specific profiles, and facilitate structure discovery based on sequence alignment between the query protein and other known proteins using PSI-BLAST and hidden Markov models (Altschul et al., 1997; Hargbo and Elofsson, 1999; Rost and Sander, 2000). The third generation methods are based mostly on considering global protein properties based on advanced multiple alignment procedures, and aim to predict secondary structure for the entire proteins.

1.2 Motivation and Goals

Existing secondary structure prediction methods are statistical in nature. The three-state prediction accuracy of third generation methods that are based solely on multiple sequence alignment is limited to the level of 88%. They cannot be expected to overcome this accuracy limit due to natural variations observed in structural families (Rost et al., 1994b; MacCallum, 1997).

The investigated new method concentrates on the prediction of secondary structure for protein fragments for which the highest confidence was achieved, instead of finding the secondary structure of the entire protein. We aim to develop a method that will be able to break the 88% accuracy limit for the selected protein

fragments providing reliable data for standard third generation methods, which will complete the prediction process. The new method applies divide and conquer principle by dividing the entire protein into uniform, in terms of the secondary structure, fragments and performing prediction for each of the fragments individually. Our investigation focuses on designing a classifier that will predict secondary structure for the protein fragments using their primary sequence as the input.

The rest of the paper is organized as follows. Section 2 gives background concepts and presents the proposed method and goals for the feasibility study, while Section 3 describes experimentation, results, and conclusions. The paper ends with summary and future work.

2 Background and Goals

The method for prediction of secondary structure for uniform protein fragments is shown in Figure 2.

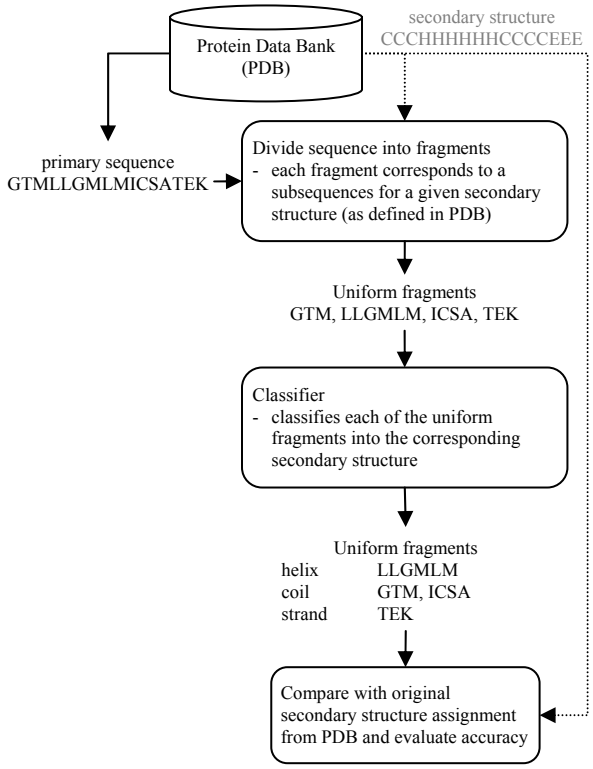

Figure 2. The procedure used to assess quality of prediction of secondary structure for protein fragments.

2.1 Selection of Uniform Protein Fragments

The assessment of the quality of the prediction for uniform fragments is based on two assumptions. First, the PDB is used as a source of data that includes primary and secondary protein structure. A custom set of filters, which are described in the next section, is used to guarantee high quality of the input data. Second, this study divides the primary sequences into subsequences that correspond to secondary structures from PDB. This is an idealized situation where the fragments are fully uniform and span the entire corresponding primary subsequence. The actual known method for dividing the sequences into fragments guarantees uniformity, but at the same time the resulting fragment usually are only subsequences of the entire corresponding primary subsequences. For instance, based on the following primary and secondary protein sequences

- primary sequence: `GTMLLGMLMICSATEK`
- secondary sequences: `CCCHHHHHHCCCCEEE`

the idealized scenario realized in this paper divides the primary sequence into GTM, LLGMLM, ICSA, and TEK, while the actual division might be GTM, LLG, MLM, IC, SA, and TEK.

While application of the idealized scenario may results in overestimating the accuracy of the resulting classification, it provides certain benefits. It allows for direct evaluation of the accuracy of classification of the fragments into the secondary structures. It also allows performing additional studies related to investigation of the relation between uniform fragments and their secondary structure.

2.2 Protein Representation

The prediction of secondary structure is performed with an intermediate step that transforms the primary sequences into their feature space representation. This is due to the differences in length of the primary sequences for different proteins, i.e. the primary sequences length can vary between several to several hundred residues (amino acids), while prediction algorithms usually assume input data of fixed length. The usual features describe amount and position of amino acids that compose a given proteins, and can vary between about 20 features to several hundred features (Kurgan and Homaeian, 2005). This investigation assumes comprehensive feature representation that is based on protein and amino acid properties listed in Table 1. The table also provides motivation for introduction of each of the features. Each uniform protein fragment is described using set of 83 features and is classified into one of the three secondary protein structures.

Table 1. Feature representation of a primary sequence for a uniform protein fragment

Property Name	Feature index	Description	Motivation
Sequence Length	1	Length (# amino acids) of a protein fragment	The length relates to quantity of the three-state structures
# of duplicates	2	Number of times a given uniform fragment occurs in a protein dataset	Higher occurrence of a uniform fragment gives higher confidence of quality of its primary structure
Relative position	3	Position of a given uniform fragment in the protein. The position describes in which quarter of the protein the majority of the fragment resides.	Fragment's position relates to the secondary structure
Average hydrophobicity	4	Average hydrophobicity value of the uniform protein fragment using Esienberg's hydrophobic index table (Cornette et al., 1997)	Hydrophobic force is one of the strongest determinant factors of a protein structure.
Accumulated average hydrophobicity	5	Accumulated (summed starting from the protein head) average hydrophobicity value of the uniform protein fragment using Esienberg's hydrophobic index table (Cornette et al., 1997)	Hydrophobic force is one of the strongest determinant factors of a protein structure. Sequences at the protein head determine the rest of the protein.
Average log Hydrophobicity	6	Average hydrophobicity value of the uniform protein fragment using hydrophobicity values of the Black and Mould hydrophobic index table (Black and Mould, 1991)	Hydrophobic force is one of the strongest determinant factors of a protein structure. Log value allows better scaling of the value.
Accumulated average log hydrophobicity	7	Accumulated (summed starting from the protein head) average hydrophobicity value of the uniform protein fragment using hydrophobicity values of the Black and Mould hydrophobic index table (Black and Mould, 1991)	Hydrophobic force is one of the strongest determinant factors of a protein structure. Sequences at the protein head determine the rest of the protein. Log value allows better scaling of the value.
Molecular weight	8	The molecular weight of the uniform protein fragment. It is computed as a sum of molecular weights of the neutral, free amino acids.	The three-state structures are related to their weight. Amino acids are very small biomolecules with an average molecular weight of about 135 Daltons.
Composition vector (1-20)	9-28	Normalized, by the protein length, composition percentage of each amino acid in the primary sequence of the uniform protein fragment	Most structure prediction methods use this property (Eisenhaber, 1996; Zhang et al., 2001; Ruan et al., 2005)
1st order compo-sition moment vector (1-20)	29-48	Normalized, by the protein length, composition percentage of each amino acid that additionally takes into consideration position of the amino acids in the primary sequence of the uniform protein fragment	Measure used for protein content prediction (Ruan et al., 2005)
Auto correlation function (1-10)	49-58	Reflects the profile of the hydrophobicity indices of residues along the amino acid sequence of the uniform protein fragment	Measure used for protein structure prediction (Zhang et al., 2001)
Electronic group (1-6)	59-64	Divides amino acids based on the electronic property, i.e. if they are neutral, electron donor or electron acceptor	Electrostatic forces are strong, and stabilize secondary and tertiary structure (Ganapathiraju et al., 2004)
Chemical group (1-19)	65-83	Divides amino acids based on chemical groups	There are 19 chemical groups of which AAs are composed (Ganapathiraju et al., 2004; Kurgan and Homaeian, 2005)
Class	84	Three state secondary structures: helix, strand, and coil	Target predicted attribute

2.3 Feasibility Study Goals

The goal of this paper is to verify the following two hypotheses:

1. Is it possible to achieve prediction accuracy significantly higher than 88% for the task of classification of uniform protein fragments?
2. What is the relation between the position of the uniform proteins fragments, with the respect to the beginning (head) of the protein sequences and the quality of secondary structure prediction?

Satisfying the first goal would provide a valuable solution to the overall new prediction method presented in Figure 1. The 88% accuracy limit is related to the current accuracy limit of the third generation secondary structure prediction methods.

The answer to the second hypothesis would provide a valuable insight into the analysis of secondary protein structures. A recent paper shows that long primary protein sequences stored in PDB are a covering set of all smaller peptides in three dimensional structures (Kichara and Skolnik, 2003). This means that short protein sequences can be found as subsequences of longer proteins. At the same time, biologist argue that for a set of primary sequences that have the same amino acids at first t sites, amino acids situated after site t, which constitute so called tail, converge to similar sequences with increasing value of t. Based on these two observations, it would be interesting to investigate if certain, say located near the

protein head, uniform proteins fragments are characterized by being "more characteristic" for given secondary sequence structures since they dictate the remaining, tail, portion of the primary sequence.

3 Experiments and Results

The two hypotheses defined in Section 2.3 were verified experimentally. A comprehensive set of experiments, which included careful preparation of input data based on proteins published in PDB, strict evaluation of prediction accuracy of several state of the art classification systems for the secondary structure prediction for the uniform fragments, and finally careful analysis of the generated results, which led to verification of the defined hypotheses, were performed. We first describe the data preparation process, which is followed by description of the selected classification systems, experimental results, and summary of the results.

3.1 Dataset Preparation

The main goal for data preparation procedure was to assure high quality of the used proteins. The primary and secondary structures of the considered in the experimentation proteins were extracted from PDB, release as of August 12th 2004. Analogically to the procedure performed in (Kurgan and Homaeian, 2005) for the proteins that have isotopes, the last one was selected. The proteins were filtered according to a set of rules defined in Table 2 to eliminate errors and inconsistencies.

Table 2. Filters used to select high quality protein sequences

Filter	# removed sequences
The length of the sequence was less than 4.	455
Number of residues did not match the sequence length.	9
The sequence had the residue called *UKN* (unknown)	25
There were some residue(s) other that the legal twenty two amino acids.	11540
There was some helix of length less that 3.	1291
There was some strand of length less that 2.	19022
There was some helix indexed out of the sequence.	10038
There was some strand indexed out of the sequence	8023
There was some coil indexed out of the sequence	219
Overlap between helix and strand	782
Overlap between helix and coil	1342
No secondary structure	9972
No primary structure	13

Additionally, all sequences with identical primary sequences and different secondary sequences were eliminated. Lastly, sequences with ambiguous amino acids in the primer, i.e. B or Z, were removed resulting in a dataset that included 5834 sequences. The dataset was further filtered using the 25% PDB SELECT list (Hobohm and Scharf, 1992; Hobohm and Sander, 1994). The list is a subset of PDB proteins that excludes proteins of low quality and homologous proteins of lower quality (the list of proteins can be obtained from http://homepages.fh-giessen.de/~hg12640/pdbselect/). The filtering resulted in the final set of 539 proteins.

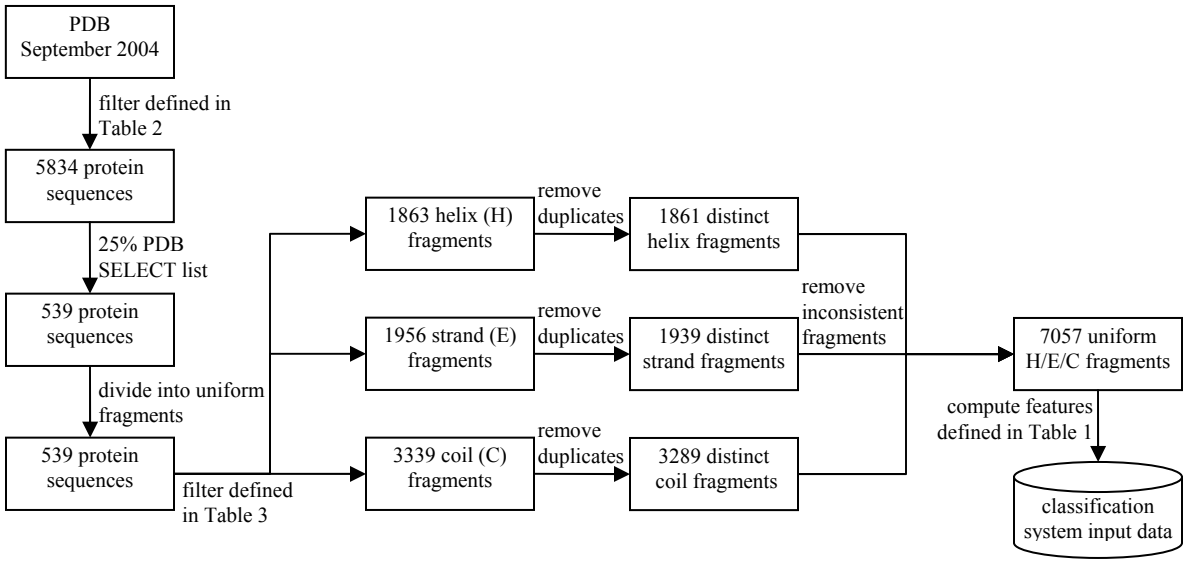

Figure 3. The dataset preparation procedure.

The resulting set of proteins was divided into three sets corresponding to uniform fragments for helix, strand, and coil structures. The uniform fragments were additionally filtered to eliminate errors similar to the ones defined for entire primary sequences, which are summarized in Table 3.

Table 3. Filters used to select high quality uniform protein fragments

State/class	Filter
H (helix fragments)	Helix fragments of length ≤ 3 and those adjacent to strand fragment were removed
E (strand fragments)	Strand fragments of length ≤ 2 and those adjacent to helix fragment were removed
C (coil fragments)	Coil fragments of length ≤ 2 were removed

Within each of the uniform fragment sets duplicate structures were removed and value of the second feature from Table 1 was computed. Next, inconsistent uniform fragments, i.e. fragments of identical primary structure and different secondary structure, between the three sets were removed. Schematic diagram for the entire dataset preparation procedure is shown in Figure 3.

Finally, the following 8 datasets were created:

- *DA*, which includes all 7057 uniforms fragments
- *D1*, *D2*, *D3*, and *D4*, which include only first, first two, first three, and first four uniform fragments respectively, with respect to the protein head, for each protein from the set of 539 proteins.
- *d1*, *d2*, *d3*, and *d4*, which include only first, second, third, and fourth uniform fragments respectively, with respect to the protein head, for each protein from the set of 539 proteins. We note that *d1* and *D1* are the same dataset.

The *DA* dataset was used to investigate the first hypothesis, while the remaining datasets were used to investigate the second hypothesis. The summary information for the datasets is shown in Table 4. The resulting sets uniform fragments were converted into the feature representation, as defined in Table 1.

3.2 Experimental Setup

The prediction of secondary structures for all considered datasets of uniform fragments was performed using a comprehensive set of classification systems. They were selected to include all state of the art systems, as well as to cover all major families of systems. The classification systems can be divided based on the generated model into the following families:

- black-box systems, which generate a model that cannot be interpreted by the user
- white-box systems, generate an interpretable model

The white-box systems can be further divided into:

- rule-based systems, which generate models that consists of production rule sets
- decision tree systems, which generate models that consists of decision trees
- probabilistic systems, which generate probabilistic models

Our paper uses state of the art methods in each of the classification systems families. A detailed list of the considered classification systems, which includes 8 methods, is shown in Table 5.

Table 4. Summary information for the considered datasets of uniform protein fragments

Data set	State	# fragments	length min	length max	Data set	State	# fragments	length min	length max
DA	H	1863	4	68	d1	H	15	4	32
	E	1956	3	26		E	2	12	13
	C	3339	3	74		C	406	3	74
D1	H	15	4	32	d2	H	267	4	68
	E	2	12	13		E	223	3	19
	C	406	3	74		C	11	3	15
D2	H	282	4	68	d3	H	13	5	24
	E	225	3	19		E	3	5	10
	C	417	3	74		C	398	3	44
D3	H	295	4	68	d4	H	258	4	59
	E	228	3	19		E	184	3	17
	C	815	3	74		C	20	3	18
D4	H	553	4	68					
	E	412	3	19					
	C	835	3	74					

Table 5. List of considered classification systems

system type		classification system name	reference
black-box		Multiple Layer Perceptron (MLP) Neural Network	(Hornik et al., 1989)
white-box	rule-base	RIPPER	(Cohen, 1995; Cohen, 1996)
		SLIPPER	(Cohen and Singer, 1999)
	decision trees	ID3	(Quinlan, 1986)
		CART	(Breaiman et al., 1984)
		C5.0	(RuleQuest, 2003)
		boosted C5.0	(RuleQuest, 2003)
	probabilistic	Naïve Bayes (NB)	(Duda and Hart, 1973)

The selected methods were used for the three-state prediction of the uniform fragments based on the feature representation. The implementation of the methods was obtained from the authors, and in case of the ID3, CART, MLP, and NB systems, the TANAGRA ver. 1.1.3 software, available at eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html, was used. The C5.0 system was combined with boosting option, which generates and combines several models to increase the prediction accuracy (Schapire and Singer, 1998). Boosting potentially results in increasing the accuracy of prediction in expense for more complex model.

We also note that the features generated according to Table 1 are continuous. All above classification systems can use continuous data, except the NB, for which Equal-Frequency discretization was performed.

3.3 Experiments, Results, and Conclusions

The selected 8 classification systems were applied on the 8 datasets. Each prediction was performed using ten-fold cross-validation procedure. The results report average accuracy together with standard deviation. Also, for all classification systems, except RIPPER and SLIPPER, average, over the three classes, sensitivity and specificity values were computed to give further insights. The results are summarized in Table 6.

To easy the analysis, the experimental results are represented using a series of figures. The analysis of the results needs to be proceeded by closer analysis of the input datasets. The Figure 4 shows distribution of secondary structures for the 8 datasets.

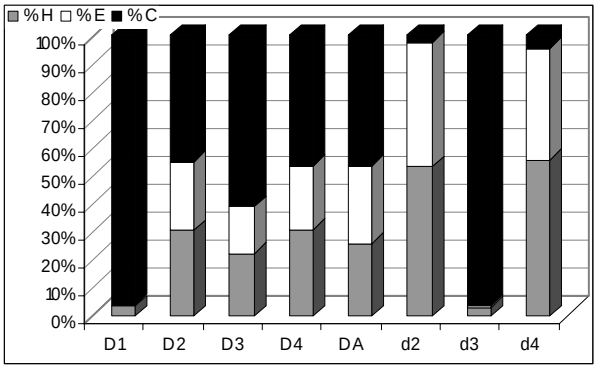

Figure 4. Distribution of the secondary structures for the considered 8 datasets

Table 6. Summary of experimental results

Classification System	dataset DA					dataset D1					dataset D2					dataset D3				
	accuracy			sensitivity	specificity	accuracy			sensitivity	specificity	accuracy			sensitivity	specificity	accuracy			sensitivity	specificity
	min	max	avg (stdev)			min	max	avg (stdev)			min	max	avg (stdev)			min	max	avg (stdev)		
RIPPER	67.52	72.34	**69.11** (1.8)	---	---	92.68	97.56	**95.21** (2.0)	---	---	62.64	74.73	**69.36** (3.8)	---	---	66.92	78.20	**72.22** (3.2)	---	---
SLIPPER	65.11	70.08	**67.75** (1.7)	---	---	92.68	97.56	**96.18** (1.7)	---	---	63.00	74.73	**70.81** (3.7)	---	---	66.92	80.45	**72.29** (4.4)	---	---
ID3	65.36	67.21	**66.59** (0.5)	64.72	82.30	95.85	96.34	**96.00** (0.2)	33.33	66.67	67.03	70.44	**68.36** (1.1)	66.78	83.26	70.45	73.68	**72.30** (1.0)	62.94	81.39
MLP	71.04	72.96	**72.43** (0.5)	70.63	85.33	94.88	96.10	**95.68** (0.5)	34.12	67.34	72.64	75.60	**74.29** (0.9)	72.29	86.66	74.89	77.37	**75.92** (0.7)	67.54	84.58
CART	66.61	68.14	**67.58** (0.5)	64.35	82.23	95.61	96.10	**95.85** (0.1)	33.54	66.85	67.69	70.99	**69.26** (1.1)	71.05	85.09	71.65	75.26	**73.40** (1.1)	64.18	82.15
NB	66.40	66.70	**66.54** (0.1)	67.15	83.29	95.61	96.34	**96.05** (0.3)	41.72	74.01	69.34	70.44	**69.91** (0.3)	71.05	85.09	68.05	68.72	**68.36** (0.2)	69.95	84.05
C5.0	64.40	68.80	**67.50** (1.4)	65.19	82.70	92.60	97.60	**95.20** (2.0)	39.50	72.05	59.80	76.10	**69.32** (4.9)	68.23	84.30	60.40	78.20	**70.64** (4.9)	62.71	81.80
boosted C5.0	70.10	74.60	**72.37** (1.4)	69.96	85.09	95.20	97.60	**96.16** (1.2)	35.56	68.63	67.00	81.50	**75.71** (4.4)	74.01	87.35	67.70	80.50	**75.22** (3.8)	66.01	83.35
AVERAGE	**67.1**	**70.1**	**68.7**	**67.0**	**83.5**	**94.4**	**96.9**	**95.8**	**36.3**	**69.3**	**66.1**	**74.3**	**70.9**	**70.6**	**85.3**	**68.4**	**76.5**	**72.5**	**65.6**	**82.9**

Classification System	dataset D4					dataset d2					dataset d3					dataset d4				
	accuracy			sensitivity	specificity	accuracy			sensitivity	specificity	accuracy			sensitivity	specificity	accuracy			sensitivity	specificity
	min	max	avg (stdev)			min	max	avg (stdev)			min	max	avg (stdev)			min	max	avg (stdev)		
RIPPER	65.92	75.14	**69.75** (2.8)	---	---	74.00	88.00	**79.20** (4.0)	---	---	2.44	100.0	**86.95** (29)	---	---	65.92	96.70	**69.75** (8.8)	---	---
SLIPPER	62.01	73.48	**66.71** (3.4)	---	---	66.00	82.00	**76.00** (5.1)	---	---	90.24	100.0	**96.14** (3.1)	---	---	62.01	73.48	**66.85** (3.4)	---	---
ID3	64.58	67.09	**65.82** (0.8)	63.62	81.63	79.00	80.80	**80.26** (0.6)	55.35	87.78	79.00	80.80	**80.26** (0.6)	33.33	66.67	77.61	78.70	**78.41** (0.4)	54.81	86.49
MLP	71.04	72.96	**72.43** (0.5)	71.19	85.91	80.60	84.20	**82.56** (1.4)	56.13	88.51	96.10	96.10	**96.10** (0.0)	33.33	66.67	76.52	79.13	**77.87** (0.9)	53.87	85.53
CART	64.64	67.88	**66.40** (0.9)	63.04	81.46	79.60	82.00	**80.52** (0.7)	55.41	87.85	79.60	82.00	**80.52** (0.7)	33.33	66.67	72.17	78.26	**75.87** (2.0)	53.24	85.06
NB	66.40	66.70	**66.54** (0.1)	69.95	84.50	78.20	80.80	**79.18** (0.8)	54.41	87.53	93.41	94.39	**93.90** (0.3)	35.30	68.12	73.91	76.96	**75.37** (0.8)	55.03	85.90
C5.0	57.50	68.20	**64.40** (3.2)	62.57	81.26	70.00	82.00	**74.80** (3.7)	50.79	83.81	92.70	97.60	**94.71** (1.9)	32.83	66.17	61.70	80.40	**71.66** (6.3)	49.62	82.17
boosted C5.0	64.40	77.70	**69.99** (3.7)	68.16	83.98	72.00	84.00	**78.80** (3.9)	53.54	86.62	92.70	97.60	**95.42** (1.8)	33.08	66.42	67.40	82.60	**75.97** (4.9)	53.64	84.30
AVERAGE	**64.6**	**71.1**	**67.8**	**66.4**	**83.1**	**74.9**	**83.0**	**78.9**	**54.3**	**87.0**	**78.3**	**93.6**	**90.5**	**33.5**	**66.8**	**69.7**	**80.8**	**74.0**	**53.4**	**84.9**

The distributions show that datasets *D1*, and *d3* contain mainly coil fragments, which is expected given the knowledge of overall protein structure. Additionally dataset *D3* contains a higher number of coil fragments due to including first three fragments in the sequences, i.e. in significant majority of the cases both 1st and 3rd fragments are coils. We also note that datasets *d2* and *d4* contain similar distribution, but with low number of coils, which agrees with the protein structure. Therefore we note that the dataset *DA* will be used to study goal 1, while goal 2 will be studied based on two sequences of datasets: *D2*, *D4*, and *DA*; *d2* and *d4*.

The *DA* datasets is suitable to provide results for goal 1 due to high quality of the data, and suitable distribution of the secondary structure. The *D2*, *D4*, and *DA* sequence will be studied to verify the goal 2. This is due to similar secondary structure distribution between these datasets. We note that since *D2* contains first two fragments in the sequence, *D4* first four, and *DA* all fragments, using this sequence of datasets will reveal if there is a relationship between position of the fragment and the quality of prediction. Similarly *d2* and *d4* sequences, which have similar distribution of classes, were selected.

To evaluate quality of the generated classifiers two aspect were considered. First, if they can archive accuracy above 88%, and second how they perform with respect to each other and so called default hypothesis. The default hypothesis is defined as the accuracy of classifier that would always choose the class (secondary structure) with the highest count of input examples. For instance, in case of *D1* is the coil class. Therefore, classifiers should generate models with accuracy significantly better than the default hypothesis accuracy to be evaluated as high quality. Therefore further analysis is based on *accuracy gain*, which equals to the difference between the achieved accuracy and the accuracy of the default hypothesis. The default hypothesis values for the selected datasets are shown in Table 7.

Table 7. Default hypothesis values

default hypothesis accuracy	**Datasets**							
	DA	D1	D2	D3	D4	d2	d3	d4
	46.6	95.91	45.12	60.91	46.38	53.29	96.13	55.84

3.1.1 Results for goal 1. The accuracy, sensitivity and specificity of the 8 classifiers, and the default hypothesis accuracy for ten-fold cross validation test on *DA* are shown in Figure 5.

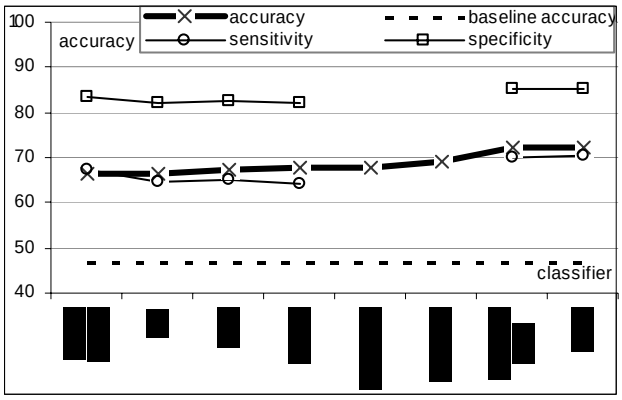

Figure 5. Summary of results for goal 1

The analysis shows that none of the classifiers was able to break the 88% accuracy limit. This shown that the described approach does not provide a suitable solution to provide a system that can be used to improve third generation secondary structure prediction methods. At the same time we note that all classifiers perform significantly better than the default hypothesis. This shows that the applied approach gives promising results that could be improved in the future. We also note that among the tested 8 classification systems the best results in terms of accuracy, sensitivity and specificity were achieved by MLP and boosted C5.0 classification systems. The two systems achieved 72.4% accuracy. A paired T-test at 5% confidence shows that the two classifiers performed significantly better, in terms of the accuracy, than all other classification systems.

3.1.2 Results for goal 2. Two studies: for *D2*, *D4*, and *DA* datasets, and for *d2* and *d4* datasets, were performed. The accuracy of the 8 classification systems for the selected datasets is shown in Figures 6 and 7. The figures show the accuracy gain values for each of the classification system, and the linear trend for the average accuracy gain.

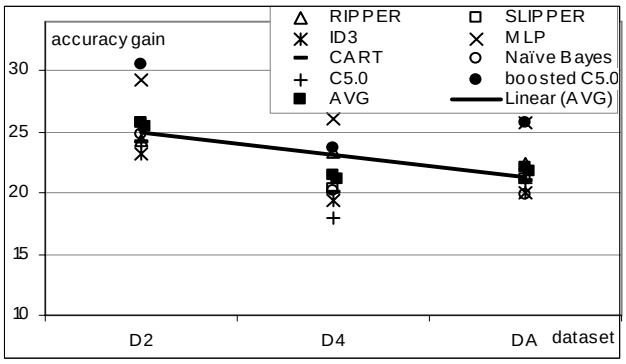

Figure 6. Summary of results for goal 2 for *D2*, *D4*, and *DA*

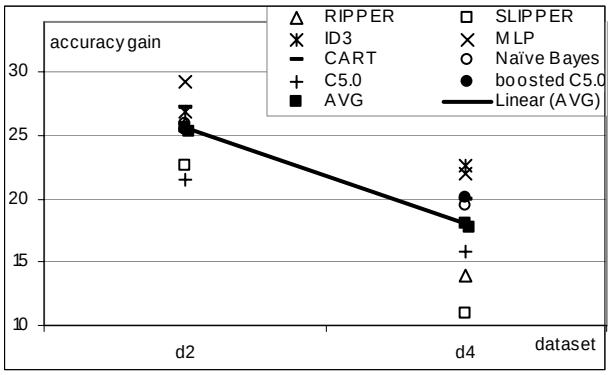

Figure 7. Summary of results for goal 2 for *d2* and *d4*

The linear trend shows that the prediction accuracy gain, which is independent of the class distribution, decreases when fragments farther from the protein head are considered. We also note that the trend is relatively weak, showing only about 3.5% accuracy gain difference in case of the *D1*, *D2*, and *DA* datasets, and about 7% in case of the *d2* and *d4* datasets.

Difference between average sensitivity and specificity for the selected datasets are shown in Figure 8.

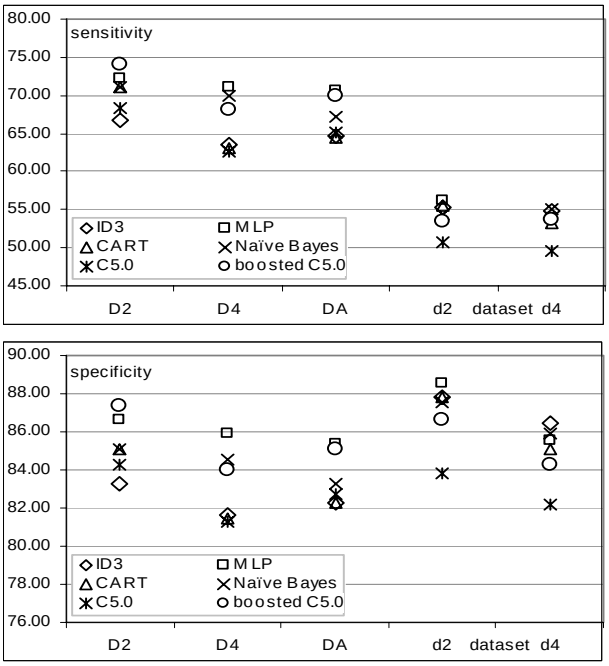

Figure 8. Summary of average sensitivity and specificity results for goal 2

The average sensitivity and specificity show the same trend as observed for accuracy gain. We observe that the MLP and boosted C5.0 classification systems are once again characterized by the best results. The lowest quality, on average, is exhibited by the ID3 system.

In a nutshell, we conclude that although the results show that higher accuracy prediction is achieved for the protein head fragments, more evidence should be collected to substantiate the claim. We also note that results of all classification systems for all datasets are characterized by high specificity of about 85%, when compared to sensitivity and accuracy. High specificity shows that generated classification models are characterized by low false positive scores, which means that the models for a specific secondary structure, say helix, very rarely classify fragments of the other secondary structures, i.e. coil and strand, as helices. This shows that the generated models are very selective, and further improvements are possible.

4 Summary and Future Work

The paper concerns feasibility study for development of a novel system for secondary protein structure prediction from primary protein sequences. It investigates two hypotheses. First, it investigates if the prediction accuracy for uniform protein fragments can break current 88% limit of predictive accuracy of the third generation prediction methods. Second, it investigates relation between position of the uniform proteins fragments, with the respect to the protein head, and the quality of secondary structure prediction.

The experimental study, which used a novel protein representation, and worked on high quality large protein dataset, shows that the prediction of the secondary structure for uniform protein fragments is limited to 72% accuracy. This shows that this approach cannot be used to develop the high quality prediction system for entire proteins. At the same time, we note that high specificity of the classification results and their relatively high accuracy shows that further improvements are possible and should be pursued. We also note that among the considered eight state of the art classification systems, the multiple layer perceptron neural networks and boosted C5.0 decision tree performed best and significantly better than other six considered systems.

The study also shows that higher accuracy, sensitivity, and specificity of the prediction are achieved when predicting secondary structure for uniform fragments closer to the beginning of the protein. This indicates that those fragments are of bigger importance and possibly have more

characteristic structure. We note that the discovered trend is relatively weak and further investigation is required to provide more evidence.

The future work will include enhancing the future representation of a primary sequence, including scoring matrices based approach for prediction, and performing additional studies that would reveal relations between the uniform proteins fragments and the corresponding secondary structure.

Acknowledgments

The authors would like to thank Dr. Jishou Ruan from Nankai University in China for useful comments and pointers. We would also like to acknowledge Leila Homaeian for help with preparation of the data, and discussion on feature representation.

This research was supported in part by the Natural Sciences & Engineering Research Council of Canada.

References

[1] Altschul, S., Madden, T., et al., Gapped Blast and PSI-Blast: a New Generation of Protein Database Search Programs, *Nucleic Acids Research*, 25, 3389-3402, 1997

[2] Berman, H.M., et al., The Protein Data Bank, *Nucleic Acids Research*, 28, 235-242, 2000

[3] Black, S.D., and Mould, D.R., Development of Hydrophobicity Parameters to Analyze Proteins which Bear Post- or Cotransitional Modifications, *Analytical Biochemistry*, 193, 72-82, 1991

[4] Breiman, L., Friedman, J., et al., *Classification and Regression Trees*, Chapman and Hall, 1984

[5] Cohen, W., Fast Effective Rule Induction, *Proc. of the 12th Intern. Conf. on Machine Learning*, 115-123, 1995

[6] Cohen, W., Learning Trees and Rules with Set-valued Features, *Proc. of the 13th National Conf. on Artificial Intelligence*, 709-716, 1996

[7] Cohen, W., and Singer, Y., A Simple, Fast and Effective Rule Learner, *Proc. of the 16th National Conf. on Artificial Intelligence*, 335-342, 1999

[8] Cornette, J.L., Cease, K., Margalit, H., et al., Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Protein, *J. of Molecular Biology*, 195, 659-685, 1987

[9] 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

[10] Duda, R.O., and Hart, P.E., *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973

[11] Eisenhaber, F., et al., Prediction of Secondary Structural Contents of Proteins from Their Amino Acid Composition Alone, I. New Analytic Vector Decomposition Methods, *Proteins*, 25:2, 157-168, 1996

[12] Engel R, Autoregressive Conditional Heteroskedastcity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, 50, 987-1008, 1982

[13] Ganapathiraju, M.K., et al., Characterization of Protein Secondary Structure, *IEEE Signal Processing Magazine*, 78-87, May 2004

[14] Garnier, J., Osguthorpe, D.J., and Robson, B., Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins, *J. of Molecular Biology*, 120:1, 97-120, 1978

[15] Gibrat, J.F., Garnier, J., and Robson, B., Further Developments of Protein Secondary Structure Prediction Using Information Theory. New Parameters and Consideration of Residue Pairs, *J. of Molecular Biology*, 198:3, pp.425-443, 1987

[16] Hargbo, J., and Elofsson, A., Hidden Markov Models that use Predicted Secondary Structures for Fold Recognition, *Proteins*, 36, 68-76, 1999

[17] Hobohm, U., Scharf, M., et al.., Selection of a Representative Set of Structures from the Brookhaven Protein Data Bank, *Protein Science*, 1, 409-417, 1992

[18] Hobohm, U., and Sander, C., Enlarged Representative Set of Protein Structures, *Protein Science*, 3, 522, 1994

[19] Hornik, K., Stinchcombe, M., and White, H., MLP's are Universal Approximators, *Neural Networks*, 2, 359-366, 1989

[20] Kihara, D., and Skolnick, J., The PDB is a Covering Set of Small Protein Structures, *J. of Molecular Biology*, 223, 793-802, 2003

[21] Kurgan, L., and Homaeian, L., Prediction of Secondary Protein Structure Content from Primary Sequence Alone - a Feature Selection Based Approach, submitted, 2005

[22] MacCallum, R.M., *Computational Analysis of Protein Sequence and Structure*, Ph.D thesis, Department of Biochemistry and Molecular biology, University College London, 1997

[23] Quinlan, J.R., Induction of Decision Trees, *Machine Learning*, 1, 81-106, 1986

[24] Rost, B., and Sander, C., Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure, *Proteins*, 19:1, 55-72, 1994a

[25] Rost, B., Sander, C., and Schneider, R., Redefining the Goals of Protein Secondary Structure Prediction, *J. of Molecular Biology*, 235, 13-26, 1994b

[26] Rost, B., and Sander, C., Third Generation Prediction of Secondary Structure, In: Webstar, D., (Ed.), *Protein Structure Prediction: Methods and Protocols*, Human Press Clifton, NJ, 71-95, 2000

[27] Rost, B., Review: Protein Secondary Structure Prediction Continue to Rise, *J. of Molecular Biology*, 134:2-3, 204-18, 2001

[28] Ruan, J., Wang, K., Yang, J., Kurgan, L., and Cios, K.J., Highly Accurate and Consistent Method for Prediction of

Helix and Strand Content from Primary Protein Sequences, *Artificial Intelligence in Medicine*, issue on *Computational Intelligence Techniques in Bioinformatics*, accepted, 2005

[29] RuleQuest Research, C5.0 at http://www.rulequest.com/see5-info.html, 2003

[30] Schapire, R.E, and Singer, Y., Improved Boosting Algorithms Using Confidence-rated Predictions, *Proc. of the 11th Annual Conf. on Computational Learning Theory*, 80-91, 1998

[31] Truhlar, D.G., et al, (Eds.), *Rational Drug Design*, The IMA Volumes in Mathematics and its Applications, vol.108, Springer-Verlag, 1999

[32] 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