The Protein Journal, Vol. 25, Nos. 7–8, December 2006 (© 2006)
DOI: 10.1007/s10930-006-9029-0
Published Online: November 11, 2006

Sequence Representation and Prediction of Protein Secondary Structure for Structural Motifs in Twilight Zone Proteins

Lukasz Kurgan,[1,2] and Kanaka Durga Kedarisetti[1]

Characterizing and classifying regularities in protein structure is an important element in uncovering the mechanisms that regulate protein structure, function and evolution. Recent research concentrates on analysis of structural motifs that can be used to describe larger, fold-sized structures based on homologous primary sequences. At the same time, accuracy of secondary protein structure prediction based on multiple sequence alignment drops significantly when low homology (twilight zone) sequences are considered. To this end, this paper addresses a problem of providing an alternative sequences representation that would improve ability to distinguish secondary structures for the twilight zone sequences without using alignment. We consider a novel classification problem, in which, structural motifs, referred to as structural fragments (SFs) are defined as uniform strand, helix and coil fragments. Classification of SFs allows to design novel sequence representations, and to investigate which other factors and prediction algorithms may result in the improved discrimination. Comprehensive experimental results show that statistically significant improvement in classification accuracy can be achieved by: (1) improving sequence representations, and (2) removing possible noise on the terminal residues in the SFs. Combining these two approaches reduces the error rate on average by 15% when compared to classification using standard representation and noisy information on the terminal residues, bringing the classification accuracy to over 70%. Finally, we show that certain prediction algorithms, such as neural networks and boosted decision trees, are superior to other algorithms.

KEY WORDS: Protein Secondary Structure; twilight zone; structural motifs; prediction.

1. INTRODUCTION

Determining protein functions and their interactions with other molecules, which primarily depend on protein structure, is a fundamental step in understanding many biological processes. Proteins are built out of folding units called domains, and at the same time these domains include smaller building blocks, called structural motifs (Boutonnet *et al.*, 1998; Unger and Sussman, 1993). Recent research shows that relatively small structural motifs, which usually consist of several α-helices and β-structures that include hydrogen-bonded β-strands, can be used to describe, classify and analyze protein structure (Szustakowski *et al.*, 2005; Taylor, 2002). Systematic characterization and identification of the structural motifs can help close the gap between secondary protein structures and functional protein folds, by providing building blocks that are assembled into the domains and by using these blocks to predict unknown structures from known sequences.

[1] Electrical and Computer Engineering Department, University of Alberta, Edmonton, Alberta, Canada T6G 2V4
[2] To whom correspondence should be addressed. E-mail: lkurgan@ece.ualberta.ca

Abbreviations: SF, structural fragment; DSSP, dictionary of secondary structures of proteins; PDB, protein data bank; AA, amino acid; MLP, multiple layer perceptron neural network; RIP, RIPPER; SLI, SLIPPER; NB, Naïve Bayes.

The three dominant computational structure prediction approaches include:

– Multiple-sequence alignment based methods. Sequence alignment is used to predict structure based on an observation that proteins with similar amino acid (AA) sequences have similar structure. In this case, 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 (Altschul *et al*., 1997; Cuff and Barton, 2000; Jones, 1999; Lin *et al*., 2005; Pollastri *et al*., 2002; Rost and Sander, 2000). The sequence alignment based methods provide superior results for the prediction of the secondary structure (McGuffin and Jones, 2003).
– Threading based methods. This approach compares 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 (Bowie *et al*., 1991; Jones, 1992; Rost, 1996, 1997). The main shortcoming of threading methods is that they are unable to recognize previously unencountered structures. In this case, threading methods will still rely on the database of known folds, and predict the most similar structure from the database. Recent research attempts to optimally combine sequence alignment and threading techniques to obtain more accurate fold predictions (Shan *et al*., 2001; Skolnick *et al*., 2004).
– Fragment assembly based methods (Bujnicki, 2006). They are based on an observation that the protein backbone structure can be accurately represented using short fragments taken from other proteins (Kim *et al*., 2004; Rohl *et al*., 2004). Another alternative are related hybrid methods, which combine fragment assembly, lattice based folding simulations and threading (Skolnick *et al*., 2001; Zhang and Skolnick, 2004).A very important aspect in case of the alignment-based methods is *sequence homology*, or in other words similarity between sequences, which is a result of inheritance from a common ancestor. Homology is defined as the percentage of amino acids in the protein sequences that are identical after aligning the sequence with other sequences from a given dataset (gaps between consecutive amino acids may be introduced during alignment, if necessary). The underlying assumption of the multiple alignment methods is that a minimum ~30% homology must exist between the query sequence and the sequences that are used to deduce its structure (Sander and Schneider, 1991). In 1999, Rost coined the term *twilight zone*, which relates to query sequences, which are characterized by low, 20–30% homology with sequences that are used to predict their structure (Rost, 1999). More than 95% of all sequence pairs detected in the twilight zone have different structures, which makes it very difficult to perform high quality structure prediction. To further illustrate this point we compare three state accuracy of secondary structure prediction for homologous (> 30%) and twilight zone sequences. In case of highly homologous sequences, i.e., when a query sequence can be aligned with high confidence to a set of sequences with known structure, the state-of-the-art alignment-based secondary structure prediction methods yield around 80% accuracy (Petersen *et al*., 2000; Pollastri and McLysaght, 2005). At the same time, in a recent study where twilight zone sequences were used, the state-of-the-art prediction methods gave substantially lower accuracies (Lin *et al*., 2005), i.e., 67.6% for PSIRED method (Jones, 1999), 67.4% for SSPro2 method (Pollastri *et al*., 2002), 66.4% for YASPIN method (Lin *et al*., 2005), 65.4% for JNET method (Cuff and Barton, 2000), and 65.0% for PHDpsi method (Przybylski and Rost, 2002). To this end, we concentrate on analysis of the twilight zone sequences without applying the multiple sequence alignment.

1.1. Motivation and goals

The current paper follows the theme of analyzing small structural motifs in order to expedite prediction of the overall protein topology for the twilight zone sequences. We focus on the most basic structural motifs: the individual α-helix, β-strand and coil fragments. Understanding how to distinguish between these structural motifs and how each of these basic building blocks of the protein structure can be represented and characterized with respect to the protein sequence is fundamental to advance our knowledge with respect to how higher-level structural motifs, such as super-secondary structures, domains and folds, are built. Here, protein sequences are divided into three sets of

structural fragments (motifs), defined as the longest fragments of a primary sequence that correspond to the same secondary structure. This investigation is motivated by two important factors:

1. The 13% wide gap between the accuracies for homologous and twilight zone sequences calls for studies that would aim at the latter case. The most recent contribution that analyzed structural similarities for short motifs include proteins with homology cut-off set at 95% (Szustakowski *et al.*, 2005). In contrast, we investigate structural motifs at the secondary structure level for sequences in the twilight zone.
2. The existing methods that investigate structural motifs assume primary sequences as the input information (Rohl *et al.*, 2004; Szustakowski *et al.*, 2005; Taylor, 2002). In contrast, since our paper concerns sequences with low sequence similarity, we concentrate on developing alternative sequence representation that is used as the input.This research draws on ideas from related structure prediction fields such as structural class, secondary structure content and protein function prediction, to develop a new, comprehensive and improved representation of protein sequences and to investigate which factors and prediction algorithms result in improved discrimination between the three secondary structures. Separate classification models are built for α-helix, β-strand, and coil fragments, and the classification of a given fragment is cooperatively determined by the three classifiers.

The main goal of this paper is not to propose yet another secondary structure prediction method, but rather to investigate how to represent twilight zone protein sequences in order to improve ability to differentiate between the three secondary structures. This work not only investigates a design of the alternative representation, but it also considers impact of other factors such as quality of the secondary structure information for terminal residues of the structural fragments and usage of different prediction algorithms. Results discussed in this paper show that when using a simple sequence representation based on AA propensities and similar classification methods to those used in the second generation secondary structure prediction methods (Gibrat *et al.*, 1987; Rost and Sander, 1994; Rost *et al.*, 1994) the prediction accuracy for the structural fragments is about 65%. At the same time, improving the sequence representation and disregarding terminal residues that are characterized by low quality secondary structure information results in significant improvements in accuracy. For the same classification methods, an accuracy of over 70% was achieved, reducing the error rates by 15%. Also, some prediction algorithms are shown to produce superior prediction results, and thus selection of a proper algorithm is an important consideration.

2. MATERIALS AND METHODS

2.1. Protein secondary structure

The Dictionary of Secondary Structures of Proteins (DSSP) (Kabsch and Sander, 1983) annotates each AA in the primary sequence as belonging to one of eight secondary structure states: H (alpha-helix), G (3-helix or 3_{10} helix), I (5-helix or π-helix), B (residue in isolated beta-bridge), E (extended strand), T (hydrogen bond turn), S (bend), and "_" (any other). Typically this annotation is reduced to three states: α-helix (H that includes "H" and "G"), β-strand (E that includes "E" and "B"), and coil (C that includes remaining types) (Moult *et al.*, 1997). The assignment of the secondary structure is usually performed in an automated fashion based on atomic coordinates. To date, the most popular method to assign the secondary structure is the DSSP (Kabsch and Sander, 1983), although a number of other methods, such as KAKSI, STRIDE, XTLSSTR, PSEA, and SECSTR can be also used (Martin *et al.*, 2005).

2.2. Protein sequence representation

While some methods that analyze and predict protein structure, including those based on the multiple alignments directly use the corresponding sequence, other methods first convert the sequence into an attributes-based representation, e.g. protein content and structural class prediction methods (Lin and Pan, 2001; Wang and Yuan, 2000). The attribute-based representation provides a viable alternative to improve secondary structure prediction for the twilight zone proteins, which by definition are characterized by relatively high sequence dissimilarity. Additionally, usage of an attribute representation for transforming protein sequences (or fragments) of different length into attribute vectors

of the same length, allows the analyst to use standard machine learning and data mining methods for prediction. The main drawback of these methods is that the existing attribute representations of proteins are inadequate.

Researchers have recognized that sequence representation based on the commonly used composition vector is not sufficient for prediction purposes, and therefore alternative representations have been sought (Cai *et al.*, 2002; Chou and Cai, 2004; Dubchak *et al.*, 1997; Kurgan and Homaeian, 2005; Lin and Pan, 2001; Luo *et al.*, 2002; Ruan *et al.*, 2005; Zhang *et al.*, 2001). This paper draws on numerous recent studies to define a comprehensive set of attributes to represent sequences that constitute structural fragments. The attributes used in our representation are presented in Table 1. We also perform attribute selection to select a subset of attributes that are the most relevant to describe the structural fragments.

2.3. Problem definition

A structural fragment (SF) is defined as the longest fragments of a primary sequence that correspond to the same secondary structure. The motivation to select such short structural motifs, in contract to prior studies, in which structural motifs were composed of small arrangements of few secondary structures (usually few helices and strands) (Szustakowski *et al.*, 2005; Taylor, 2002), comes from the low homology assumption. While the overall sequence may be in the twilight zone,

Table 1. Attribute representations for a protein sequence; 1 (Lin and Pan, 2001), 2 (Muskal and Kim, 1992), 3 (Syed and Yona, 2003), 4 (Eisenhaber *et al.*, 1996), 5 (Zhang *et al.*, 1998), 6 (Zhang *et al.*, 2001), 7 (Wang and Yuan, 2000), 8 (Luo *et al.*, 2002), 9 (Cai *et al.*, 2003), 10 (Ganapathiraju *et al.*, 2004), 11 (Nelson and Cox, 2000), 12 (Wang *et al.*, 2000), 13 (Yang and Wang, 2003), 14 (Hobohm and Sander, 1995), 15 (Ruan *et al.*, 2005); the attributes are normalized with respect to the sequence length

Attribute set name	Description	Motivation	References (prediction task)
Length	# of residues in the primary sequence	May be related to content	
hydropho-bicity	Average and accumulated average hydrophobicity computed using Eisenberg's (Cornette, 1987) and Fauchere-Pliska's (Fauchere and Pliska, 1983) hydrophobic indices	Hydrophobic force is one of the strongest determinant factors of a protein structure	1,6 (content)
Molecular weight	Sum of molecular weights (Black and Mould, 1991) of neutral, free AAs	May be related to content and function	2 (content) 3 (function)
Composition vector	Normalized composition percentage of each AA in the primary sequence	Used in most content & structural class prediction methods	1,4,5,6 (content) 7,8,9 (struct class)
Composition moment vector	1st order composition vector, combines position and composition AAs in the sequence	Supplementing composition with position was shown to improve content prediction	15 (content)
Auto-correlation	Autocorrelation value computed using Fauchere-Pliska's hydrophobic index	Reflects profile of hydrophobic values along the primary sequence	1,5,6 (content)
Electronic group	Divides AAs into neutrals, electron donors or acceptors	Electrostatic forces stabilize structure	10 (structure)
R group	Combines hydropathy, molecular weight and pI	May be related to structure and content	11 (structure /content)
Exchange group	Some AAs can be substituted by other without impact on the structure	Represents conservative replacements through evolution	12 (family) 13 (structure)
Hydrophobic group	Divides AAs into hydrophobic and hydrophilic	The same as for hydrophobicity	2,14 (function)
Other groups	Considers the following classes: charged, polar, aromatic, small, tiny, bulky, and polar uncharged	May be related to function	3,14 (function)
Chemical group	19 chemical groups that compose the side chains of the AAs	May be related to structure	10 (structure)

its short segments are likely to exhibit higher degree of homology. Additionally, this definition of structural motifs allows us to study impact of the quality of secondary structure assignment for terminal residues of SFs, which is a common problem for the secondary structure assignment methods (Martin *et al.*, 2005). To illustrate the SF definition, we consider the primary sequence and the secondary structure derived using DSSP based on NMR 2D homonuclear technique for Mediterranean mussel defensin MGD-1 protein (protein ID 1FJN) from the Protein Data Bank (PDB)(Berman *et al.*, 2000):

GFGCPNNYQCHRHCKSIPGRCGGYCGG
WHRLRCTCYRCG
CCCCCCHHHHHHHHHHCCCCCEEEECC
CCCCCEEEECCC

The corresponding SFs for α-helix are NYQCHRHCKS, for β-strand are GGYC and CTCY, and for coil are GFGCPN, IPGRC, GGWHRLR, and RCG.

Structural fragment prediction is defined as prediction of secondary structure for a given primary sequence fragment based on models inferred from SFs for which the corresponding structure is known; see Fig. 1. Solid lines denote how the models are generated, while dotted lines show how they are used to perform prediction.

We emphasize that this paper is not concerned with how these SFs are extracted from a sequence in order to perform secondary structure prediction, but we rather use this classification problem to propose novel sequence representation and to investigate how using certain prediction algorithms can help to improve our ability to distinguish between different secondary structures. This, in turn, can help to better characterize these structural motifs and ultimately improve analysis of the overall protein topology.

For the SF prediction problem, the attribute representation discussed in section 2.2 was supplemented with the following two attributes:

- *number of duplicates*, which is the number of occurrences of a given SF among all SFs for the same secondary structure, e.g. how many time the GFGCPN fragment appears among the coil fragments in the training database. Higher values provide higher confidence that the SF is associated with a given secondary structure
- *relative position*, which approximates the position of a given structural fragment in the primary AA sequence. Each protein sequence was divided into four quarters, and the relative position corresponds to the quarter within which the majority of the structural fragment is contained, e.g. relative position of the GFGCPN fragment is equal to 1 since the fragment in contained in the first quarter of the 1FJN sequence. This attribute represents the relationship between SF positions with respect to the sequence termini.

2.4. Data preparation

Since the SF prediction task was not considered in past research and we specifically aim at analyzing the twilight zone sequences, we first create a suitable database of SFs. The data was extracted from the PDB, August 12, 2004 release. For proteins with multiple chains, the last chain was selected. Next, the proteins were filtered to eliminate errors and inconsistencies. Proteins with missing primary or secondary sequences, with sequence length < 5, with sequences containing unknown or incorrect residues, with helices of length < 3, and with strands of length < 2 were filtered out. After filtration, 5834 proteins remained. Among them, a subset of 539 high quality twilight zone proteins was selected using the 25% PDB SELECT list (Hobohm and Sander, 1994). The 25% PDB SELECT list includes only high quality low homology proteins, i.e. proteins scanned with high resolution and with about 25% sequence identity. The primary sequences for the 539 proteins were divided into

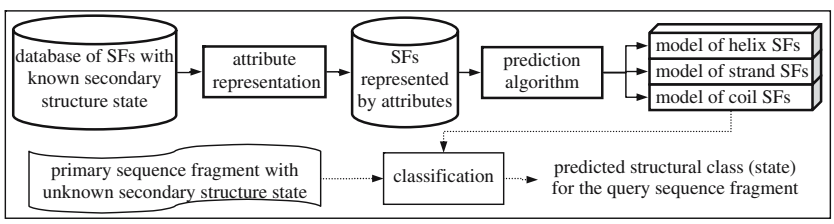

Fig. 1 Diagram of the structural fragment prediction task.

SFs, which were further converted into attribute representation and grouped by their corresponding secondary structure label. In each set duplicates and inconsistent SFs were counted and removed, and helix SFs of length ≤ 3, and strand and coil SFs of length ≤ 2 were eliminated to improve quality of the data (this was necessary due to the division of filtered proteins into SFs). A total of 7056 SFs were generated. A more detailed description of the filtering procedure can be found in (Kurgan and Kedarisetti, 2005).

We then created several datasets to explore specific goals (described later) related to the SF classification problem, as follows (the corresponding number of SFs is given in brackets):

– D (7056), which includes all SFs,
– D-2 (5284), D-4 (3590) and D-6 (2557), which includes all SFs, but with 2, 4, or 6 terminal residues removed in the SFs (1, 2, and 3 residues are removed on each side),The distribution of helix, strand, and coil SFs for each dataset is shown in Figure 2.

For example, our procedure would divide the 1FJN protein, shown earlier, and insert the resulting SFs into our datasets as in Table 2. Since the 1FJN protein contains relatively short SFs, no fragments for datasets D-l and D-vl were extracted.

2.5. Prediction algorithms

SF prediction can be performed using a wide range of prediction algorithms (classifiers). In this paper eight algorithms were considered. These may be divided into black-box algorithms, which generate a model that cannot be interpreted by a user, and white-box algorithms, which generate an interpretable model. The latter are further divided into rule-based, decision trees and probabilistic algorithms. Representative prediction algorithms for each of the categories are used (see Table 3). For the Naïve Bayes algorithm, attribute values were discretized using equal-frequency discretization.

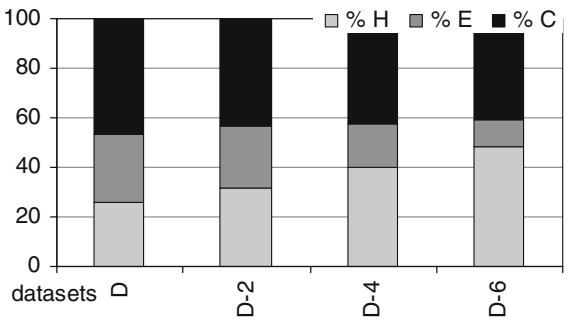

Fig. 2 Distribution of the helices (H), strands (E), and coils (C) for the considered datasets; dataset D includes all SFs, datasets D-2, D-4 and D-6 include SFs with 2, 4, and 6 terminal residues removed, respectively.

Table 2. Results of dividing 1FJN protein and inserting the resulting SFs into datasets D, D-2, D-4 and D-6

Dataset	Structural fragments
D	GFGCPN, NYQCHRHCKS, IPGRC, GGYC, GGWHRLR, CTCY, RCG
D-2	FGCP, YQCHRHCK, PGR, GWHRL
D-4	QCHRHC, WHR
D-6	CHRH

2.6. Detailed goals related to structural fragment prediction

The general problem of SF prediction was used to address a number of specific goals related to how the three secondary structures (and thus the corresponding structural motifs) can be distinguished based on the primary sequences represented using a comprehensive set of attributes described in section 2.2:

– GOAL 1: Investigate the quality of different prediction algorithms.
– GOAL 2: Investigate if terminal residues in the SFs suffer from decreased reliability of the secondary structure classification by discarding these residues and investigating the differences in prediction accuracy. This goal directly addresses the question of ill-defined secondary structures for terminal residues of SFs, which are generated by the secondary structure assignment methods (Martin *et al.*, 2005)
– GOAL 3: Selection of an optimal attribute representation. The comprehensive representation consisting of all attributes described in section 2.2 was used to select an optimal subset of attributes. This subset was compared with the most commonly used composition vector representation and with the original set of all attributes.Dataset D was used to address goal 1, datasets D, D-2, D-4, and D-6 for goal 2.2, and finally datasets D and D-6 for goal 3. In the next section, experimental results in support of each the defined goals are presented.

Table 3. Representative algorithms used to perform structural fragment prediction

Algorithm type		Algorithm name	Reference
Black-box		Multiple layer perceptron neural network (MLP)	(Hornik *et al.*, 1989)
White-box	Rule-based	RIPPER (RIP)	(Cohen, 1996)
		SLIPPER (SLI)	(Cohen and Singer, 1999)
	Decision trees	ID3	(Quinlan, 1986)
		CART	(Breiman *et al.*, 1984)
		C5.0	(RuleQuest, 2003)
		bC5.0 (C5.0 with boosting)	(RuleQuest, 2003)
	Probabilistic	Naïve Bayes (NB)	(Duda and Hart, 1973)

2.7. Testing and evaluation

To ensure statistical validity, experiments were performed using 10-fold cross validation, in which the original dataset is partitioned into 10 subsets. Of the 10 subsets, 1 is retained to test the prediction model, and the remaining 9 are used to generate the model. The cross validation process is repeated 10 times, with each of the 10 subsets used exactly once as the test data. The results from these 10-folds are averaged to produce a robust estimate of the quality of the model. The models were validated using accuracy, sensitivity and specificity, which are standard measures defined based on a confusion matrix that is shown in Table 4.

The *accuracy* is defined as ratio between the number of correct predictions and the total number of predictions: $accuracy = \frac{a+e+i}{a+b+c+d+e+f+g+h+i}$.

The *sensitivity* is the ratio between the correct and all predictions for a given secondary structure (H, E, and C): $sensitivity_{\mathrm{H}} = \frac{a}{a+b+c}$, $sensitivity_{\mathrm{E}} = \frac{e}{d+e+f}$, $sensitivity_{\mathrm{C}} = \frac{i}{g+h+i}$.

The *specificity* is the ratio between the correct and all predictions for fragments that should be excluded for a given secondary structure: $specificity_{\mathrm{H}} = \frac{e+f+h+i}{d+e+f+g+h+i}$, $specificity_{\mathrm{E}} = \frac{a+c+g+i}{a+b+c+g+h+i}$, $specificity_{\mathrm{C}} = \frac{a+b+d+e}{a+b+c+d+e+f}$.

Table 4. Confusion matrix for the SF prediction

Actual structure	Predicted structure		
	Helix (H)	Strand (E)	Coil (C)
Helix (H)	a	b	c
Strand (E)	d	e	f
Coil (C)	g	h	i

a, e and i are # of correct predictions for helix, strand, and coil fragments, respectively; b is the number of incorrect predictions where helix fragment is identified as strand, c is the number of incorrect predictions where helix fragments is identified as coil, etc.

We report the average, over the 10-folds, accuracy and weighted, by the relative number of SFs that belong to the three secondary structures, average sensitivity and specificity. The accuracy gives only an overall evaluation, while a high confidence can be placed for results that give high values for all three measures (Cios and Moore, 2002).

3. RESULTS

Over 300 experiments were performed using the four datasets and the comprehensive sequence representation that includes all attributes listed in section 2.2 to represent SF sequences. Results report average accuracy and standard deviations. For all prediction algorithms (except the proprietary implementation of SLIPPER that does not report the confusion matrix) weighted average sensitivity and specificity were computed to give further insights. The results are summarized in Table 5.

3.1. GOAL 1: prediction algorithm selection

The average accuracy over all eight-prediction algorithms for D dataset is 68.5%. The eight algorithms were ranked using average accuracy. MLP is the most accurate and bC5.0 is the second best, (see Fig. 3). Average, over the four datasets, accuracy, sensitivity and specificity results from Table 5 also indicate that MLP and bC5.0 are superior. Based on the paired *t*-test with 5% significance level, MLP is significantly more accurate than all algorithms except bC5.0. All of the algorithms demonstrated good specificity (all above 79%, average of 81%), indicating few false positives. The sensitivity values,

Table 5. Summary of experimental results for SF prediction for the 8 datasets and 8 prediction algorithms; dataset D includes all SFs, datasets D-2, D-4 and D-6 include SFs with 2, 4, and 6 terminal residues removed, respectively; All SFs were encoded using the comprehensive sequence representation described in section 2.2; "average" is the average over the four datasets, while "avg" is the average over the eight algorithms

	dataset	MLP	RIP	SLI	ID3	CART	C5.0	bC5.0	NB	avg
accuracy (stand. dev.)	D	72.6 ± 0.5	68.8 ± 2.1	67.6 ± 2.3	66.7 ± 0.6	67.6 ± 0.5	66.5 ± 2.8	72.0 ± 1.4	65.8 ± 0.2	68.5
	D-2	73.5 ± 0.4	69.0 ± 3.0	67.0 ± 2.3	67.3 ± 0.7	68.0 ± 0.5	68.0 ± 2.5	72.3 ± 2.8	67.8 ± 0.1	69.1
	D-4	73.3 ± 0.2	68.8 ± 2.8	67.2 ± 2.8	67.5 ± 0.7	67.5 ± 0.7	67.1 ± 3.6	73.9 ± 1.8	67.1 ± 0.2	69.1
	D-6	75.0 ± 0.6	71.2 ± 3.8	70.5 ± 3.6	69.5 ± 0.7	70.1 ± 0.7	68.0 ± 1.8	73.7 ± 2.1	67.7 ± 0.3	70.7
	average	73.6	69.5	68.1	67.8	68.3	67.4	73.0	67.1	
sensitivity /specificity	D	73 / 84	69 / 79	—	67 / 80	68 / 79	66 / 80	72 / 83	66 / 81	69 / 81
	D-2	74 / 84	69 / 80	—	67 / 81	68 / 82	68 / 82	72 / 84	68 / 82	69 / 82
	D-4	73 / 84	69 / 80	—	67 / 82	67 / 81	67 / 81	74 / 84	67 / 81	69 / 82
	D-6	75 / 83	71 / 79	—	70 / 79	70 / 80	68 / 81	74 / 83	68 / 81	71 / 81
	average	74 / 84	70 / 80	—	68 / 81	68 / 81	67 / 81	73 / 84	67 / 81	

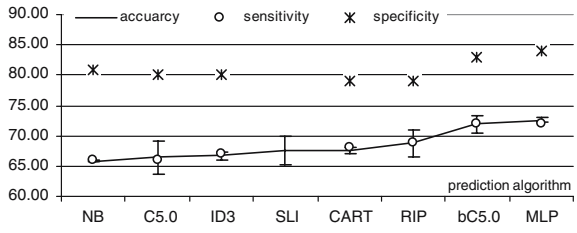

Fig. 3 Ranking of prediction algorithms on the D dataset; vertical bars represent standard deviation for the accuracy; the SFs were encoded using the comprehensive sequence representation described in section 2.2.

however, are relatively lower, averaging 69%. This means that the algorithms generate very selective models, which could be further improved by relaxing some constraints (e.g. pruning) to shrink the gap between sensitivity and specificity and thus increase accuracy.

3.2. GOAL 2: impact of terminal residues on prediction accuracy

We compare the predictive accuracy of dataset D against D-2, D-4 and D-6 to determine if terminal residues in the SFs have a detrimental effect on the reliability of SF structure prediction (see Fig. 4). The four datasets use all SFs, but with some terminal residues removed. The results show that, for all prediction algorithms, the prediction accuracy improves when deleting residues at the edges of SFs. The average, over the eight classifiers, accuracy increases by 0.6–2.2% when the terminal residues are removed. The same trend is also true for the best performing algorithms, i.e., 2.4% and 1.7% improvement when deleting three residues is achieved for MLP and

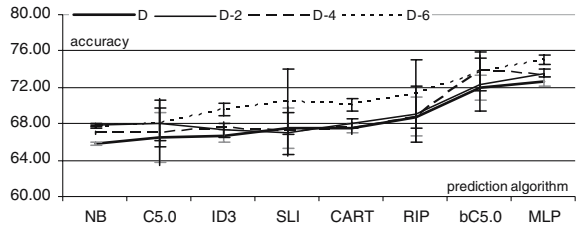

Fig. 4 Results of prediction when removing terminal SF residues; vertical bars represent standard deviation for the accuracy; the SFs were encoded using the comprehensive sequence representation described in section 2.2.

bC5.0, respectively. A paired *t*-test with 5% significance level was performed to investigate if the differences in accuracy are significant. The results for all three datasets with removed residues are significantly more accurate than the results obtained with dataset D. At the same time, the results for dataset D-6, which incorporated the most aggressive removal of terminal residues, are significantly more accurate at the 0.5% significance level than results for all three other datasets (D, D-2 and D-4). This indicates that there is a negative impact on the reliability of predictions when residues at SF edges are kept. In other words, the quality of the secondary structure assignment for the terminal residues is worse when compared with the residues located inside the SFs, which agrees with (Martin *et al.*, 2005).

3.3. GOAL 3: attribute representation of protein sequences

We have conducted over 3000 experiments using the D and D-6 datasets, three representative prediction algorithms (ID3, MLP, and NB), and 10-fold cross validation tests. The goal is to find a subset of the attribute representation described in section 2.2 that still yields comparably good prediction accuracy. Attribute selection was performed iteratively, where in each step, each of the attribute sets was individually tested, and the best one was selected. The first iteration resulted in selection of the composition moment vector. The second best attribute set was the composition vector, which gave 0.5% and 0.2% lower accuracy for datasets D and D-6, respectively. During the second and third iterations chemical group and autocorrelations based on hydrophobicity were selected; see the summary of results in Table 6. The results show average accuracy over the three algorithms for both datasets, how many times each of the attribute sets gave the best results, and relative rank for each of the attribute sets.

The selection process was stopped after the third iteration since the selected attribute sets were already comparable in accuracy to using all attributes, i.e. average accuracy for the selected three attribute sets was only 0.7% and 0.3% lower than average accuracy for the same three algorithms when all attributes were used for datasets D and D-6, respectively. The attribute set rank in all iterations shows that composition vector and electronic group also contribute to improved accuracy. For virtually all experiments the hydrophobicity attribute set computed using Fauchere's index was superior to Eisenberg's index. The composition moment vector gave on average better results than the commonly used composition vector, which confirms results in (Ruan *et al.*, 2005). The chemical group was also recently shown to increase accuracy of protein structural class prediction (Kurgan and Homaeian, 2006). In short, the results show that only a handful

Table 6. Attribute set selection results (selected attribute sets are in grey); 1 (SF length), 2 (# duplicates), 3 (relative position), 4 (Eisenberg's hydrophobicity), 5 (Fauchere's hydrophobicity), 6 (mol weight), 7 (comp vector), 8 (comp moment vector), 9 (autocorrelation), 10 (electronic gr), 11 (R gr), 12 (exchange gr), 13 (hydrophobic gr), 14 (other gr), 15 (chemical gr); "avg accuracy D" and "avg accuracy D-6" is the average accuracy over the three algorithms (ID3, MLP and NB) for datasets D and D-6, respectively (values in bold denote the best results); "avg" is the average accuracy over the attribute sets; "# times best" shows how many times using the corresponding attribute set gives the best, over the eight algorithms, accuracy (6 best results are recorded in total for each iteration – 2 datasets and 3 algorithms); "rank" denotes relative rank for each attribute sets in each iteration

iteration		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Avg
1	Avg accuracy D	48.0	44.3	46.3	50.7	53.2	45.3	65.1	**65.6**	58.3	57.2	53.3	54.0	44.6	55.0	62.6	**53.6**
	Avg accuracy D-6	49.8	44.8	44.4	47.8	54.5	51.9	69.2	**69.4**	58.6	62.2	60.0	62.4	54.6	57.2	68.1	**57.0**
	# times best	0	0	0	0	0	0	3	**3**	0	0	0	0	0	0	0	
	Rank	12	14	15	11	9	13	2	**1**	5	4	7	6	10	8	3	
2	Avg accuracy D	66.1	65.3	65.2	65.4	66.1	65.6	66.0	N/A	66.4	66.3	65.1	65.5	65.7	64.8	**66.6**	**65.7**
	Avg accuracy D-6	69.7	68.9	68.1	68.1	69.4	68.4	70.0	N/A	68.3	69.5	68.7	69.8	69.0	68.5	**70.3**	**69.0**
	# times best	1	0	0	0	0	0	0	N/A	2	0	0	0	0	0	**3**	
	Rank	4	9	13	12	5	10	2	N/A	7	3	11	6	8	14	**1**	
3	Avg accuracy D	67.6	66.8	66.4	66.2	66.9	66.4	67.0	N/A	**67.7**	66.9	66.4	66.5	66.8	66.6	N/A	**66.8**
	Avg accuracy D-6	70.0	70.2	70.2	69.9	69.8	69.5	**70.9**	N/A	70.4	70.2	69.3	70.0	70.0	69.5	N/A	**70.0**
	# times best	1	1	0	0	1	0	0	N/A	**2**	1	0	0	0	0	N/A	
	Rank	3	5	8	10	7	12	2	N/A	**1**	4	13	9	6	11	N/A	

Table 7. Comparison of prediction with different attribute representations; dataset D includes all SFs and dataset D-6 includes SFs with 6 terminal residues removed; "average" is the average accuracy over the eight algorithms

Prediction algorithm	Attribute representations for dataset D-6			Attribute representations for dataset D		
	Composition vector	Selected best attributes	All attributes	Composition vector	Selected best attributes	All attributes
MLP	71.3	73.3	75.0	68.0	70.9	72.6
RIP	69.6	70.8	71.2	65.0	67.6	68.8
SLI	67.2	69.7	70.5	62.8	66.1	67.6
ID3	67.2	69.4	69.5	63.6	65.6	66.7
CART	66.6	69.3	70.1	64.0	67.1	67.6
C5.0	67.1	68.6	68.0	64.1	66.4	66.5
bC5.0	69.9	72.5	73.7	68.5	70.8	72.0
NB	68.5	68.6	67.7	65.3	66.7	65.8
Average	68.4	70.3	70.7	65.2	67.7	68.5

of attributes are needed to distinguish between the three types of SFs, but at the same time those "best" attributes are different than the commonly used attribute representations. As such, the results provide useful guidelines to develop sequence representations for the twilight zone proteins.

Additionally, prediction accuracies when using (1) all attributes, (2) the selected subset consisting of composition moment vector, chemical group, and hydrophobic autocorrelations (3) the most commonly used representation including the composition vector, were compared. Experiments were performed using all eight algorithms, and two datasets previously mentioned (see Table 7). Our results show that on average 3.2%, 2.6% and 2.2% accuracy was gained by removing the terminal residues when using the composition vector only, the selected best attributes and all attributes, respectively. The 1.9% and 2.5% improvement in accuracy was obtained when the selected best attributes were used instead of the composition vector for the D-6 and D datasets, respectively. The combined improvement, when both best proposed sequence representation is used and the terminal residues are removed, equals 5.1%, which translates into reduction of the error rate by 15% (from 35% to 30%). Similar improvements are observed for the best performing algorithms, i.e., MLP and bC5.0. The significance of the differences in accuracy over the eight algorithms was investigated using a paired t-test. For both datasets using the selected best attributes gives significantly better accuracy at the 0.05% significance level when compared with the commonly used composition vector. Similarly, when all attributes are used the accuracy is significantly better at the 0.05% and 0.5% levels when compared to composition vector for the D and D-6 datasets, respectively. This demonstrates that improved sequence representation helps to obtain statistically significant improvements in prediction of secondary structure for the SFs.

4. DISCUSSION

The paper studies regularity among twilight zone protein structures at the level of structural fragments (SFs), which are defined as structural motifs that correspond to the longest fragments of a primary sequence that correspond to the same secondary structure. SFs are the basic building blocks of larger scale and more complex protein structures, such as super secondary structures, basic forms and domains (Szustakowski *et al.*, 2005; Taylor, 2002). The main goal was to answer two questions: first, how to represent protein sequences to better differentiate between different SFs (secondary structures), and second, what algorithms should be considered to improve this ability and if structure prediction for terminal residues in the SFs suffers from decreased reliability?

Based on comprehensive experimental studies, our results provide several interesting insights into characterization of the SFs. First, terminal residues are characterized by decreased quality of the secondary structure assignment when compared with the residues located inside the SFs. We have shown that removing these residues results in significant improvement in discrimination between the secondary structures. Second, the SFs should be described by a carefully designed set of attributes to allow for better differentiation between the three secondary

structures. The results show that attribute-based representations of a sequence corresponding to SF should include the composition moment vector, chemical group, hydrophobic autocorrelation, composition vector and electronic group information. These attributes describe the composition and location of AA that constitute the SF's sequence, the composition of individual chemical groups in the AA's side chains, and finally the hydrophobicity profile of the SF sequence. The experiments show that based on this information, over 70% of the SFs can be correctly classified into their corresponding secondary structure. Finally, we show that some prediction algorithms, such as neural networks and boosted decision trees, generate significantly better results than some other methods. Therefore, selection of a suitable prediction method will result in better quality of the SF models.

The characterization of the structural motifs was performed for low homology (twilight zone) sequences. It provides useful insights for the difficult problem of secondary structure prediction for such sequences. Additionally, the procedures described in this paper are not dependent on the sequence alignment and thus are complementary to the current mainstream secondary structure prediction methods.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ruan for fruitful comments and discussions. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). *Nucleic Acids Res*. **25:** 3389–3402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). *Nucleic Acids Res*. **28:** 235–242.

Black, S., and Mould, D. (1991). *Anal. Biochem*. **193:** 72–82.

Bowie, J., Luthy, R., and Eisenberg, D. (1991). *Science* **253:** 164–170.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). In: Classification and Regression Trees, Chapman and Hall.

Boutonnet, N., Kajava, A., and Rooman, M. (1998). *Proteins* **30:** 193–212.

Bujnicki, J. (2006) *Chembiochem* **7:**(1), 19–27.

Cai, Y., Liu, X, and Chou, K. C. (2002). *J. Comput. Chem*. **24:**(6), 727–731.

Cai, Y., Liu, X., Xu, X., and Chou, K. C. (2003). *J. Theor. Biol*. **221:** 115–120.

Chou, K-C., and Cai, Y-D. (2004). *Biochem. Bioph. Res. Co*. **321:** 1007–1009.

Cios, K. J., and Moore, G. W (2002). *Artif. Intell. Med*. **26:** 1–24.

Cohen, W. (1996). In: Proc. 13th Nat Conf. on Artificial Intelligence, Portland, Oregon, pp. 709–716.

Cohen, W., and Singer, Y. (1999). In: Proc 16th Nat Conf. on Artificial Intelligence, Orlando, Florida, pp. 335–342.

Cornette, J., Cease, K., Margalit, H., Spouge, J., Berzofsky, J., and DeLisi, C. (1987). *J. Mol. Biol*. **195:** 659–685.

Cuff, J. A., and Barton, G. J. (2000). *Proteins* **40:** 502–511.

Dubchak, I., Muchnik, I., and Kim, S-H. (1997). Protein Folding Class Predictor for SCOP: Approach Based on Global Descriptors, Proc of 5th Intelligent Systems for Molecular Biology (ISMB) Conference, Halkidiki, Greece, pp. 104–107.

Duda, R., and Hart, P. (1973) Pattern Classification and Scene Analysis. New York: John Wiley and Sons.

Eisenhaber, F., Imperiale, F, Argos, P., and Frommel, C. (1996). *Proteins* **25:**(2), 157–168.

Fauchere, J. L., and Pliska, V. (1983). *Eur. J. Med. Chem*. **18:** 369–375.

Ganapathiraju, M. K., Klein-Seetharaman, J., Balakrishnan, N., and Reddy, R. (2004). *IEEE Signal Proc. Mag*. **15:** 78–87.

Gibrat, J. F., Garnier, J., and Robson, B. (1987). *J. Mol. Biol*. **198:**(3), 425–443.

Hobohm, U., and Sander, C. (1994). *Protein Sci*. **3:** 522.

Hobohm, U., and Sander, C. (1995). *J. Mol. Biol*. **251:** 390–399.

Hornik, K., Stinchcombe, M., and White, H. (1989). *Neural Networks* **2:** 359–366.

Jones, D. T. (1992) *J. Mol. Biol*. **287:** 797–815.

Jones, D. T. (1999) *J. Mol. Biol*. **292:** 195–202.

Kabsch, W., and Sander, C. (1983). *Biopolymers* **22:**(12), 2577–2637.

Kim, D. E., Chivian, D., and Baker, D. (2004). *Nucleic Acids Res*. **32:** W526–W531.

Kurgan, L., and Homaeian, L. (2005) Proc of Inter Conf on Machine Learning and Data Mining (MLDM2005). Germany: Leipzig 334–345.

Kurgan, L., and Kedarisetti, K. (2005) Proc of Symposium on Human-Centric Computing. Canada: Banff 26–36.

Kurgan, L., and Homaeian, L. (2006). *Pattern Recognition* **39:**(12), 2323–2343.

Lin, Z., and Pan, X-M. (2001). *J. Protein Chem*. **20:**(3), 217–220.

Lin, K., Simossis, V. A., Taylor, W. R., and Heringa, J. (2005). *Bioinformatics* **21:**(2), 152–159.

Luo, R., Feng, Z., and Liu, J. (2002). *Eur. J. Biochem*. **269:** 4219–4225.

Martin, J., Letellier, G., Marin, A., Taly, J., Brevern, A.de, and Gibrat, J. (2005). *BMC Struct. Biol*. **5:** 17.

McGuffin, L., and Jones, D. (2003). *Proteins* **52:**(2), 166–175.

Moult, J., Hubbard, T., Bryant, S., Fidelis, K., and Pedersen, J. T. (1997). *Proteins* **29:** 2–6.

Muskal, S. M., and Kim, S-H. (1992). *J. Mol. Biol*. **225:** 713–727.

Nelson, D., and Cox, D. (2000) Lehninger Principles of Biochemistry. 3New York: Worth.

Quinlan, J. R. (1986) *Mach. Learn*. **1:** 81–106.

Petersen, T., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G., and Lund, O. (2000). *Proteins* **41:** 17–20.

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). *Proteins* **47:** 228–235.

Pollastri, G., and McLysaght, A. (2005). *Bioinformatics* **21:**(8), 1719–1720.

Przybylski, D., and Rost, B. (2002). *Proteins* **46:** 197–205.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). *Method. Enzymol*. **383:** 66–93.

Rost, B, Sander, C., and Schneider, R. (1994). *J. Mol. Biol*. **235:** 13–26.

Rost, B., and Sander, C. (1994). *Proteins* **19:**(1), 55–72.

Rost, B. (1996) *Method. Enzymol*. **266:** 525–539.

Rost, B. (1997) *J. Mol. Biol.* **270:** 1–10.
Rost, B. (1999) *Protein Eng.* **12:** 85–94.
Rost, B., and Sander, C. (2000). In: Webstar, D., (ed.), Protein Structure Prediction: Methods and Protocols, Human Press Clifton, pp.71–95.
Ruan, J., Wang, K., Yang, J., Kurgan, L., and Cios, K. (2005). *Artif. Intell. Med.* **35:**(1–2), 19–35.
RuleQuest Research (2003). C5.0 rule learner at www.rulequest.com/see5-info.html.
Sander, C., and Schneider, R. (1991). *Proteins* **9:** 56–68.
Shan, Y. B., Wang, G. L., and Zhou, H. X. (2001). *Proteins* **42:** 23–37.
Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M. R., Rotkiewicz, P., and Boniecki, M. (2001). *Proteins* **5:** 149–156.
Skolnick, J., Kihara, D., and Zhang, Y. (2004). *Proteins* **56:** 502–518.
Syed, U., and Yona, G. (2003). In: Proc of Annual Conference on Research in Computational Molecular Biology (RECOMB 2003), Berlin, Germany, pp. 224–234.
Szustakowski, J., Kasif, S., and Weng, Z. (2005). *Bioinformatics* **21:**(Suppl.2), ii66–ii71.
Taylor, W. (2002) *Nature* **416:**(6881), 657–660.
Unger, R., and Sussman, J. (1993). *J. Comput. Aid. Mol. Des.* **7:**(4), 457–472.
Wang, Z-X., and Yuan, Z. (2000). *Proteins* **38:** 165–175.
Wang, J., Ma, Q., Shasha, D., and Wu, C. (2000). In: Proc of the 6th ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining, Boston, MA, pp. 305–309.
Yang, X., and Wang, B. (2003). In: Proc of the 8th ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery, San Diego, CA, pp. 80–87.
Zhang, C. T., Lin, Z., Zhang, Z., and Yan, M. (1998). *Protein Eng.* **11:**(11), 971–979.
Zhang, Z. D., Sun, Z. R., and Zhang, C. T. (2001). *J. Theor. Biol.* **208:** 65–78.
Zhang, Y., and Skolnick, J. (2004). *P. Natl. A. Sci.* **101:** 7594–7599.