Impact of the Predicted Protein Structural Content on Prediction of Structural Classes for the Twilight Zone Proteins

Lukasz A. Kurgan, Mandana Rahbari, and Leila Homaeian
Department of Electrical and Computer Engineering, University of Alberta, Canada
{lkurgan, mandana, leila}@ece.ualberta.ca

Abstract

This paper addresses in silico prediction of protein structural classes as defined in the SCOP database. The SCOP defines total of 11 classes, while majority of proteins are classified to the 4 classes: all-α, all-β, α/β, and $\alpha+\beta$. The main goals of this paper are to experimentally evaluate the impact of predicted protein secondary structure content on the structural class prediction and to develop a novel protein sequence representation. The experiments include application of three protein sequence representations and four classifiers to prediction of both 4 and 11 structural classes. The predictions are performed using a large dataset of low homology (twilight zone) sequences. The proposed sequence representation includes the predicted structural content, which provides the strongest contribution towards classification, composition and composition moment vectors, hydrophobic autocorrelations, chemical group composition and molecular weight of the protein. The predicted content values are shown on average to improve the prediction accuracy by 3.3% and 4.2% for the 4 and 11 classes, respectively, when compared to sequence representation that does not utilize this information. Finally, we propose a very compact, 20 dimensional sequence representation that is shown to improve the prediction accuracy by 5.1-8.5% when compared with recently published results.

1. Introduction

At a basic level, a protein is composed of an ordered chain of amino acids (AAs), which locally folds into three types of secondary structures: α-helices, β-sheets and coils. In 1976, Levitt and Chothia defined four structural classes for globular proteins [1]: 1) all-α class, which contains proteins with small amount of strands, 2) all-β class with proteins with small amount of helices, 3) α/β class that includes proteins with helices and mostly parallel strands, and 4) $\alpha+\beta$ class with proteins that include helices and mostly antiparallel strands. Nowadays, proteins are manually assigned to one of the eleven Structural Classification of Proteins (SCOP) structural classes based on their structural information [5].

Protein structure prediction based on its sequence receives significant attention due to the increasing difference between the number of known protein sequences (nearly 2.5 millions) and experimentally determined structures (about 37,000). A priori knowledge of structural classes allows to improve protein secondary structure prediction [2] and reduces the search space of possible configurations of the tertiary structure [3,4].

Although majority of the structure prediction methods utilize multiple sequence alignment, the structural class prediction is performed based on classification of the sequences, which are converted into a feature-based representation [6-20] [29,30]. The prediction of structural classes for low homology sequences is of special importance since sequence alignment requires at least ~30% homology between the query protein and protein(s) used to predict its structure [31]. The proteins characterized by lower, 20-30% homology with sequences that are used to predict their structure are called *twilight zone* proteins [32]. More than 95% of all sequence pairs detected in the twilight zone have different structures [32], which significantly impacts quality of the structure prediction. For instance, prediction of the secondary structure for homologous sequences by the state-of-the-art alignment secondary structure prediction methods yields about 80% accuracy [33], while for the twilight zone sequences it drops to 65-68% [34]. Similarly, in case of the structural class prediction accuracies for highly homologous protein datasets reach over 90%, while they drop to 57% in case of the twilight zone sequences [20].

To this end, this paper aims to improve accuracy of structural class prediction for the twilight zone proteins. The proposed approach is based on a novel idea that uses a predicted secondary structure content to improve the accuracy of protein structural class prediction. The secondary structural content is defined as a percentage amount of α-helices and β-strands in a protein, and can be accurately predicted using multiple linear regression models [23,24].

2. Related Work

Prediction of the protein *structural classes* is performed in two steps. First, the protein sequences of various lengths are converted into a fixed size feature vector. Second, the feature vectors are fed into a classifier to obtain predicted class. Majority of prediction methods use a simple composition vector (20-dimensional vector that represents the occurrence frequencies of the 20 AAs) as the feature based sequence representation. These methods apply a wide range of classifiers including maximum component coefficient algorithm [6], least correlation angle algorithm [7], fuzzy clustering [8,29], artificial neural network [9,10,11], vector decomposition [12], component coupled geometric algorithm [13], Bayesian classification [14], and most recently support vector machines [15] and boosting [30]. The most noticeable progress among these algorithms was obtained by including the coupling effect among different AAs [3,

Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA'06)
0-7695-2735-3/06 $20.00 © 2006 IEEE

13]. Some of the most recent works apply alternative sequence representation that include auto-correlation functions based on non-bonded AA energy [16], polypeptide composition [17,18], functional domain composition [19] and most recently chemical composition and hydrophobic autocorrelations [20].

Similarly to structural class prediction, the *content prediction* is performed using the same two steps. The sequence representation usually consists of the composition vector and hydrophobic autocorrelations. Only two prediction algorithms, i.e., multiple linear regression method (MLR) [21] and neural networks [22], were applied. Recent research shows that MRL gives the most accurate results [23,24].

We note that to the best of our knowledge these two prediction methods were never combined together.

3. Background

3.1. Dataset of the Twilight Zone Sequences

To evaluate the classification accuracy, a dataset of twilight zone sequences was selected based on the 25%PDBSELECT list [25]. This list includes proteins that were scanned with high resolution and with low, on average 25% homology (the homology ranges between 22% and 45%). Using PDB release as of February 2005, 2340 sequences and domains were extracted based on this list. Among them there are 443 all-α, 443 all-β, 346 α/β and 441 $\alpha+\beta$ sequences, while for the remaining 246 sequences the SCOP classes are missing and 421 sequences belong to the remaining seven SCOP classes. Two datasets, one that includes the 4 major structural classes and another that includes all 11 classes were created. The final datasets with 4 classes (denoted as *25PDB-4*) contains 1673 proteins/domains. The second dataset (denoted as *25PDB-11*) consists of 2094 sequences/domains, which in addition to the 4 main classes include 26 multi-domain proteins (denoted as e), 52 membrane and cell surface proteins (denoted as f), 227 small proteins (denoted as g), 40 coiled coils proteins (denoted as h), 7 low resolution proteins (denoted as i), 62 peptides (denoted as j) and 7 designed proteins (denoted as k).

3.2. Sequence Representation

The proposed sequence representation includes features introduced in a recent structural class prediction method [20], which are combined with features used to predict structural content [23,24]. The dataset that includes sequences encoded using the corresponding 67 dimensional feature vector is denoted as *25PDB67*. This sequence representation was enhanced by adding predicted secondary structure content. Two prediction methods [23,24], which predict the α-helix and β-strand content, were used to compute the four additional features. The corresponding datasets that includes total of 71 features is denoted as *25PDB71* and includes:

- Composition vector (*CV*)
- First order composition moment vector (*CMV*)
- Autocorrelations based on Fauchere and Pliska's hydrophobicity index (*ACH*)
- Autocorrelations based on side-chain mass (*ACM*)
- Molecular weight of the protein (*MW*)
- Chemical group composition (*CG*)
- Secondary structural content for α-helix (H) and β-strand (E) based on method by Zhang, Sunt and Zhang [23] and Lin and Pan [24] (*CE-ZSH*, *CE-LP*, *CH-ZSH*, *CH-LP*)

The composition and composition moment vectors are defined as [26]:

$$CMW_i^{(k)} = \frac{\sum_{j=1}^{C_i} n_{ij}^k}{\prod_{d=0}^{k}(N-d)}$$

where i=1,2,…,20 is the AA index, k is the order of the composition moment vector (for k=0 it reduces to CV), N is the length of the protein sequence, n_{ij} is the j^{th} position of the i^{th} AA, and c_i is the count (composition) of the i^{th} AA in a sequence.

Autocorrelation function $AC_{n,}$ which is calculated based on Fauchere and Pliska's hydrophobicity index is defined as [27]:

$$ACH_n = \frac{\sum_{j=1}^{N-n} h_{ij} \cdot h_{ij+n}}{N-n}$$

where h_{ij} is the index value (shown in Table 1) for the i^{th} AA at the j^{th} position in the sequence, and n=1,2,…,10 is the lag that equals to the number of autocorrelations.

Similarly, autocorrelations of relative side-chain masses are defined as [23]:

$$ACM_n = \frac{\sum_{j=1}^{N-n} sm_{ij} \cdot sm_{ij+n}}{N-n}$$

where n=1,2,…,6 and sm_{ij} values are shown in Table 1.

The molecular weight of the protein is defined as:

$$MW = \frac{\sum_{j=1}^{N} m_{ij}}{N}$$

where m_i is the atomic weight of AAs in the sequence, see Table 1.

Chemical group composition is defined based on the chemical composition of the side chains. There are 19 chemical groups, and of them are associated with multiple different side chains – for details see [20].

The four features related to the predicted content were computed based on four MLR models (two methods [23,24] were used to generate models for α-helix and for β-strand).

Table 1. Hydrophobic index, relative side-chain masses and molecular weight of AAs

AA_i	A_1/M_{11}	C_2/N_{12}	D_3/P_{13}	E_4/Q_{14}	F_5/R_{15}	G_6/S_{16}	H_7/T_{17}	I_8/V_{18}	K_9/W_{19}	L_{10}/Y_{20}
Fauchere & Pliska hydrophobic index (h_i)	0.42 1.68	1.34 -0.82	-1.05 0.98	-0.87 -0.3	2.44 -1.37	0.00 -0.05	0.18 0.35	2.46 1.66	-1.35 3.07	2.32 1.31
Relative side-chain masses (sm_i)	0.115 0.577	0.36 0.446	0.446 0.323	0.55 0.55	0.7 0.777	0.00076 0.238	0.63 0.346	0.13 0.33	0.48 1	0.13 0.82
Molecular weight (m_i)	71 131	103 114	115 97	129 128.1	147.1 156.1	57 87	137.1 101	113.1 99.1	128.1 186.1	113.1 163.1

4. Proposed Prediction System and Goals

4.1. System Diagram

Figure 1 shows the overall prediction process. The prediction performed in prior works is shown using white boxes and solid arrows (prior works used only a subset of 67 features). Our contribution consists of adding the predicted content values to the feature representation, which is shown using grayed boxes and dotted arrows.

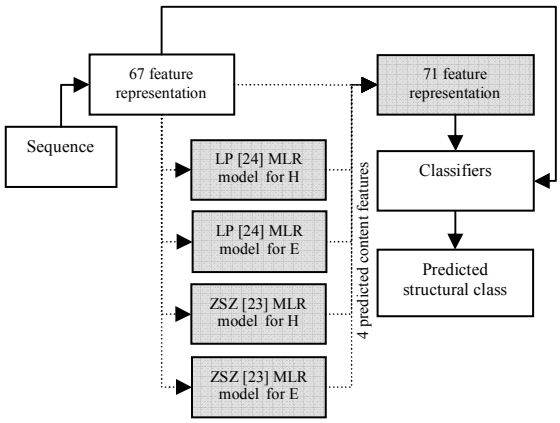

Figure 1. Diagram of the proposed prediction system

4.2. Goals

The paper addresses the two following goals:

- *Goal 1*. To design a compact feature based sequence representation and to verify if the predicted secondary structure content has an impact on the protein structural class prediction.
- *Goal 2*. To perform an empirical evaluation of the impact of the predicted content on the accuracy of the 4-classes and 11-classes protein structural class prediction. The prediction accuracies on the twilight zone dataset that is encoded using three sequence representations (25PDB67, 25PDB71 and the proposed compact representation) and using four different classifiers are compared. The comparison includes the overall accuracies, as well as the accuracies for individual structural classes.

5. Experiments and Results

The experiments were performed using the WEKA 3.4 environment [28]. All experiments, including feature selection to address Goal 1, and computations of the predicted content values and classification to address Goal 2 were performed using 10 fold cross validation.

The overall classification accuracy that was computed for Goal 2 is defines as:

$$accuracy = \frac{a+b+c+d}{n}$$

while the accuracy for all-α class (and by analogy for the remaining structural classes) is defined as:

$$accuracy_{all-\alpha} = \frac{n-(ba+ca+da+ab+ac+ad)}{n}$$

where a, b, c, d values are defined in Table 2.

Table 2. Confusion matrix for the 4-classes prediction

Actual structural class	Predicted structural class			
	All-α	All-β	α/β	α+β
all-α	a	ab	ac	ad
all-β	ba	b	bc	bd
α/β	ca	cb	c	cd
α+β	da	db	dc	d

5.1. Experimental Results for Goal 1

The design of a new, compact feature based sequence representation and evaluation of the value-added of the features related to the predicted content was performed based on a feature selection. The 25PDB71-4 and 25PDB71-11 datasets were used to establish separate representations for the 4-classes and 11-classes predictions. The representations were developed based on a consensus among three feature selection methods:

- *Chi*^{2} feature selection, which ranks attributes by computing the χ^2 test with respect to the class [35].
- *Information gain* feature selection, which ranks attributes using the information gain (introduced in decision trees) with respect to the class [35].
- ReliefF, which ranks attributes by repeatedly sampling examples and considering the value of a given attribute for the nearest example of the same and different classes [36].

The proposed representation consists of the best 20 features that selected based on the average rank among the three feature selection methods. The number of dimensions is equal to the most popular composition vector based representation.

Table 3. Feature selection results (number of selected features for each feature set)

Dataset	Selection method	# of extracted features from each feature set							% of extracted features from each feature set						
		CV	CMV	ACH	ACM	MW	CG	CE/CH	CV	CMV	ACH	ACM	MW	CG	CE/CH
25PDB71-4	Chi^2	6	3	4	0	1	2	4	30	15	40	0	100	20	100
	Information Gain	5	4	4	0	1	2	4	25	20	40	0	100	20	100
	ReliefF	6	2	2	0	1	5	4	30	10	20	0	100	50	100
25PDB71-11	Chi^2	1	1	10	0	1	3	4	5	5	100	0	100	30	100
	Information gain	3	2	7	0	1	3	4	15	10	70	0	100	30	100
	ReliefF	6	3	1	0	1	5	4	30	15	10	0	100	50	100

Table 4. Feature selection results (list of features that constitute the proposed 20 dimensional sequence representation)

Feature selection on 25PDB71-4

	Rank for a given method			
Features	Chi^2	Info Gain	ReliefF	Average
CE-LP	1	1	1	1.0
CE-ZSZ	3	2	2	2.3
CH-ZSZ	2	3	4	3.0
CH-LP	4	4	3	3.7
ACH_2	5	5	7	5.7
MW	6	6	5	5.7
CG_4	7	7	14	9.3
CV_5	8	8	13	9.7
CV_9	10	10	9	9.7
CMV_{18}	9	9	12	10.0
CMV_9	11	12	11	11.3
CV_{17}	13	13	8	11.3
CV_1	17	16	10	14.3
ACH_4	14	14	20	16.0
CG_{10}	15	15	18	16.0
CV_{20}	24	22	16	20.7
CV_{16}	16	18	30	21.3
CMV_8	21	20	28	23.0
ACH_3	18	17	37	24.0
CV_8	23	24	26	24.3

Feature selection on 25PDB71-11

	Rank for a given method			
Features	Chi^2	Info Gain	ReliefF	Average
CH-ZSZ	1	1	3	1.7
CV_2	2	4	2	2.7
CH-LP	5	2	4	3.7
MW	3	3	6	4.0
CE-ZSZ	7	5	1	4.3
CMV_1	4	7	7	6.0
CE-LP	10	6	5	7.0
CG_6	6	8	8	7.3
ACH_2	8	9	21	12.7
CG_8	18	11	13	14.0
ACH_4	13	14	23	16.7
CG_3	25	12	14	17.0
CV_9	27	17	9	17.7
CMV_9	30	22	10	20.7
CV_{17}	29	21	15	21.7
ACH_7	17	18	31	22.0
ACH_1	12	16	45	24.3
CV_1	31	26	17	24.7
ACH_3	23	24	39	28.7
ACH_5	16	19	52	29.0

Table 3 summarizes the selected best 20 features with respect to their corresponding feature sets; for each set the number and percentage of the selected features is given. The selected features include all four predicted secondary content values, molecular weight and some of the features from the remaining feature sets, except the autocorrelations based on side-chain masses. This shows that almost all feature sets are useful with respect to the prediction of the structural class, both in case of 4- and 11-classes.

Table 4 shows the ranking of the best selected features for both, 4- and 11-classes predictions; features selected in both cases are shown in bold. The results show that the most valuable features are the predicted content values; they occupy top 4 positions for 4 class problem and 4 out top seven for the 11 class problem. The second best feature set is the molecular weight that was ranked 6^{th} and 4^{th}, respectively. The other features that are included for both predictions are autocorrelations based on hydrophobicity with the lag equal 2, 3, and 4, composition vector for alanine (A), lysine (K), and threonine (T), and composition moment vector for lysine. The remaining features, which include composition for cysteine (C), isoleucine (I), serine (S) and tyrosine (Y) and composition moment for alanine (A), isoleucine and valine (V), hydrophobic autocorrelations with lag equal 1, 5 and 7 and finally six out of ten chemical composition groups, were included in one of the two representations.

In short, the results strongly indicate that the predicted content values are among the most useful features for prediction of the secondary structural class. The other useful feature sets include molecular weight, composition and composition moment vectors, low lag hydrophobic autocorrelations, and chemical composition groups. At the same time, the autocorrelations based on side-chain masses, which were originally used to predict the structure content [24], and higher lag hydrophobicity autocorrelations provide relatively smaller amount of useful information for the structural class prediction.

The new datasets, for the 4- and 11-classes prediction, that use the proposed 20 features representations are denoted as *25PDB20-4* and *25PDB20-11*, respectively.

5.2. Goal 2

The prediction of the structural classes for both 4-classes and 11-classes problems was performed using three sequence representations (25PDB67, 25PDB71, and 25PDB20) and four classifiers:

- Support Vector Machine (SVM) [37] with a second degree polynomial kernel.
- Multinomial logistic regression (LR) [38].
- Random Forest (RF) [39].

– Instance based (IB1) which is lazy learner based on the nearest neighbor algorithm [40].

The first three classifiers were selected based on their superior performance in prior works on structural class prediction [20]. The lazy learner was selected to provide contrast for the best performing algorithms. Each of the classifiers was optimized with respect to its parameters (e.g. kernel type, ridge value, number of trees) based on the 10 fold cross-validation.

Table 5. Summary of the structural class prediction results

Classifiers	Accuracy for 4 classes					Accuracy for 11 classes		
	This paper			Results after [20]		This paper		
	25PDB67-4	25PDB71-4	25PDB20-4	CV	66	25PDB67-11	25PDB71-11	25PDB20-11
RF	54.6%	56.4%	57.2%	47.6%	51.0%	52.2%	57.2%	56.7%
LR	60.5%	**62.2%**	60.0%	51.0%	56.7%	56.9%	58.4%	58.0%
IBI	41.2%	45.3%	47.0%	37.8%	39.2%	37.5%	42.0%	47.1%
SVM	56.8%	*56.2%*	58.2%	52.0%	55.1%	56.1%	**58.9%**	57.9%
average	53.3%	55.0%	**55.6%**	47.1%	50.5%	50.7%	54.1%	**54.9%**

Table 6. Structural class prediction results for individual classes

SCOP class	dataset	Accuracy$_{class\ i}$ for i = all-α, all-β, ….					signifi cance	dataset	Accuracy$_{class\ i}$ for i = all-α, all-β, ….					signifi cance
		RF	LR	IB1	SVM	mean			RF	LR	IB1	SVM	mean	
All-α	25PDB67-11	79.1	84.2	76.7	83.0	**80.7**	N/A	25PDB67-4	80.6	84.9	77.6	81.5	**81.2**	N/A
	25PDB71-11	82.7	84.7	80.0	82.7	**82.5**	+	25PDB71-4	82.8	85.0	81.6	83.9	**83.3**	+
	25PDB20-11	83.0	84.1	81.0	84.2	**83.1**	+	25PDB20-4	84.8	84.5	80.5	84.2	**83.5**	+
All-β	25PDB67-11	77.8	83.4	73.2	81.7	**79.0**	N/A	25PDB67-4	76.8	80.6	71.4	78.7	**76.9**	N/A
	25PDB71-11	81.1	84.0	76.7	81.6	**81.0**	+	25PDB71-4	78.7	82.3	74.0	80.6	**80.6**	++
	25PDB20-11	82.0	82.6	79.2	82.5	**81.6**	+	25PDB20-4	79.4	80.8	74.0	80.1	**78.9**	+
α/β	25PDB67-11	84.5	86.7	74.5	85.6	**82.8**	N/A	25PDB67-4	73.9	73.0	64.2	73.0	**71.0**	N/A
	25PDB71-11	85.1	86.6	75.2	86.0	**83.2**	+	25PDB71-4	73.4	72.7	64.4	73.6	**71.0**	~
	25PDB20-11	85.4	85.8	79.7	86.3	**84.3**	+	25PDB20-4	72.5	73.6	69.6	73.3	**72.2**	+
α+β	25PDB67-11	74.6	77.0	68.1	76.8	**74.1**	N/A	25PDB67-4	68.0	72.0	64.6	70.0	**68.7**	N/A
	25PDB71-11	75.5	78.2	70.2	76.9	**75.2**	+	25PDB71-4	69.3	73.2	65.1	72.1	**70.0**	++
	25PDB20-11	74.7	76.0	70.7	75.1	**74.1**	+	25PDB20-4	68.8	70.7	63.1	69.1	***68.0***	-
e	25PDB67-11	98.8	98.1	97.9	98.3	**98.3**	N/A							
	25PDB71-11	98.8	98.2	97.7	98.3	***98.2***	-							
	25PDB20-11	98.8	98.5	98.0	98.8	**98.5**	+							
f	25PDB67-11	98.6	97.3	98.3	97.9	**98.0**	N/A							
	25PDB71-11	98.7	97.2	98.1	98.1	**98.0**	~							
	25PDB20-11	98.5	98.0	97.9	98.3	**98.2**	+							
g	25PDB67-11	96.3	95.7	93.0	96.2	**95.3**	N/A							
	25PDB71-11	96.2	96.1	92.8	96.0	**95.3**	-							
	25PDB20-11	96.0	96.7	94.7	96.7	**96.0**	+							
h	25PDB67-11	98.1	97.6	97.6	97.8	**97.8**	N/A							
	25PDB71-11	98.4	97.7	97.5	98.0	**97.9**	+							
	25PDB20-11	98.3	98.0	97.6	98.1	**98.0**	+							
i	25PDB67-11	99.7	98.5	99.5	99.1	**99.2**	N/A							
	25PDB71-11	99.7	98.6	99.3	99.1	**99.2**	-							
	25PDB20-11	99.7	99.7	99.5	99.7	**99.7**	+							
j	25PDB67-11	97.3	95.9	96.7	96.3	**96.6**	N/A							
	25PDB71-11	97.1	96.3	96.7	96.4	**96.6**	+							
	25PDB20-11	97.4	96.8	96.5	96.6	**96.8**	+							
k	25PDB67-11	99.7	99.4	99.6	99.5	**99.6**	N/A							
	25PDB71-11	99.6	99.3	99.6	99.4	***99.5***	-							
	25PDB20-11	99.7	99.5	99.5	99.6	**99.6**	+							

The overall (across all classes) structural class prediction accuracies for the four classifiers, three representations and 4- and 11-classes problems are shown in Table 5. The best accuracies (shown in bold) were achieved by LR in case of the 4-classes prediction and by SVM in case of 11-classes prediction. The average (across all representations and both prediction) accuracy ranks the LR classifiers first (59.3%), with SVM second (57.4%), RF third (55.7%) and IB1 with a distant last position (43.4%), which confirms results from [20]. The poor performance of IB1 learner is due to low homology among sequences.

Comparison between 25PDB67 and 25PDB71, and 25PDB67 and 25PDB20 datasets shows the impact of using the secondary structure content on the prediction accuracy. Comparison of average accuracies between 25PDB71-4 and 25PDB67-4 datasets shows that adding structural content features improves the accuracy by 2.7%. At the same time, the difference in case of the 25PDB20-4 dataset is 3.3%, and demonstrates that the structural content helps and that the designed representation provides not only reduced dimensionality, but also improvements in accuracy. Similarly for the 11-classes prediction, the 71 dimensional representations

gives 3.4% improvement in accuracy and the proposed 20 dimensional representation gives 4.2% improvement when compared to representation that does not use the predicted content. In short, the results clearly demonstrate that adding the four structural content prediction based features increases the prediction accuracy for both 4- and 11-classes predictions and that the proposed, compact representation results in best, on average, results.

The obtained results were compared with recent results obtained for the same datasets, which were reported in [20]. This paper reports classification accuracies the four major structural classes when using the 20 dimensional composition vector representation and a custom designed 66 dimensional representation that did not use the predicted content values. Both average results and results for individual classifiers demonstrate superiority of the proposed solution. The proposed 20 dimensional representation gives on average 5.1% and 8.5% improvements when compared with the 66 and 20 dimensional representation proposed in [20]. The best obtained results show 5.5% improvement in favor of the proposed method.

Table 6 shows the impact of adding the predicted content on the accuracy for individual structural classes. The results show that the structural content features increase the prediction accuracy for all four major structural classes. For the 11-classes prediction the only two cases when adding structural content does not provide improvements are multi-domain proteins (class e) and designed protein (class k) (shown in italics in the *mean* column). Both of these classes combined include only 33 sequences, which constitutes only about 1.5% of the proteins in the entire dataset.

The *significance* column in Table 6 shows the statistical significance (based on a paired t-test at the 95% significance) of the differences between the 25PDB67 and 25PDB71, and 25PDB67 and 25PDB20 datasets for the four classifiers. The following annotation is used:

- + (-) denotes that results for 25PDB71 or 25PDB20 are better (worse) than that for 25PDB67, but the difference is not statistically significant.
- ++ (--) shows that the prediction results for 25PDB71 or 25PDB20 are statistically significantly better (worse) than for 25PDB67.
- ~ shows that the results for 25PDB71 or 25PDB20 and 25PDB67 are equal.

The results show that for 4-classes prediction adding structural content features results in statistically significant improvements for the all-β (3.7% improvement) and α+β classes (1.3% improvement). The results for the remaining two classes are improved, but the differences are not statistically significant. For the 11-classes prediction the structural content improves all results, but none of the improvements is statistically significant. Finally, the proposed 20 dimensional representation provides improvements in 3 out of 4 classes in the 4-classes prediction and all classes in the 11-classes prediction.

6. Summary and Conclusions

Protein structural class prediction from its sequence is a very challenging problem. The common factor among the past attempts was poor performance when considering prediction for the twilight zone proteins. The best reported past results show 48% [14] and 57% [20] accuracy. Other higher reported accuracies were shown to be a result of methodological errors [20,41]. At the same time, this problem provides the true values to the community, while prediction for sequences with higher homology should be performed using multiple sequence alignment [14]. To this end, this paper proposes a novel structural class prediction method for the twilight zone proteins. The method is the first to use the predicted secondary structure content values and to design a comprehensive and compact 20 dimensional protein representation.

Based on extensive experimental study several interesting finding and conclusions are made:

- The results clearly show that the predicted content values are among the most useful features for prediction of the secondary structural class. Adding these features on average increases the prediction accuracy for the 4 major classes by 3.3% and for the 11-classes by 4.2%.
- The proposed 20 dimensional feature based sequence representation includes predicted content values, molecular weight of the protein, composition and composition moment vectors, low lag hydrophobic autocorrelations, and chemical composition groups. A separate representation was proposed for prediction of the 4 and the 11 structural classes. For the 4-classes prediction, the proposed representation gives on average 5.1% and 8.5% improvements when compared with the best published results that applied the 66 and 20 dimensional representations, respectively [20]. At the same time, the proposed representation results in improvements for all classes in the 11-classes prediction and for 3 out of 4 classes for the 4-classes prediction when compared with a comprehensive representation that uses 67 features.
- Among many classifiers that were used in this prediction task, the multinomial logistic regression and support vector machine are shown to provide superior results.
- The best prediction results for the twilight zone proteins were obtained with the logistic regression (62.2%) and the support vector machine (58.9%) for the 4 and 11-classes predictions, respectively.

In short, we conclude that structural class prediction benefits from including the predicted secondary structure content, and that the proposed sequence representation

can be successfully used to improve accuracy for this challenging prediction task.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] M. Levitt and C. Chothia (1976), Structural patterns in globular proteins, *Nature,* 261:552–557

[2] M. Gromiha and S. Selvaraj (1998), Protein secondary structure prediction in different structural classes, *Protein Engineering,* 11:249–251

[3] K.C. Chou and C.T. Zhang (1995), Prediction of protein structural classes, *Crit. Rev. Biochem. Mol. Biol.,* 30:275-349

[4] I. Bahar, A.R. Atilgan, R.L., Jernigan and B. Erman (1997), Understanding the recognition of protein structural classes by amino acid composition, *Proteins,* 29:172–185

[5] A. Murzin, S. Brenner, T. Hubbard and C. Chothia (1995), SCOP: a structural classification of protein database for the investigation of sequence and structures, *Journal Molecular Biology,* 247:536–540.

[6] C.T. Zhang and K.C. Chou (1992), An optimization approach to predicting protein structural class from amino-acid composition, *Protein Science,* 1:401–408

[7] K.C. Chou and C.T. Zhang (1993), A new approach to predicting protein folding types, *Journal Protein Chemistry,* 12:169–178

[8] C.T. Zhang, K.C. Chou and G.M. Maggiora (1995), Predicting protein structural classes from amino acid composition: application of fuzzy clustering, *Protein Engineering,* 8:425–435

[9] B.A. Metfessel, P.N. Saurugger, D.P. Connelly and S. Rich (1993), Cross-validation of protein structural class prediction usingstatistical clustering and neural networks, *Protein Science,* 2:1171–1182

[10] I. Dubchak, I. Muchnik, S.R. Holbrook and S.H. Kim (1995), Prediction of protein-folding class using global description of amino-acid sequence, *Proc. Nat. Acad. Sci.,* 92:8700–8704

[11] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk and S.H. Kim (1999), Recognition of a protein fold in the context of the SCOP classification, *Proteins,* 35:401–407

[12] F. Eisenhaber, C. Frömmel and P. Argos (1996), Prediction of secondary structural content of proteins from their amino acid composition alone, II the paradox with secondary structural class, *Proteins,* 25:169–179

[13] K.C. Chou and G.M. Maggiora (1998), Domain structural class prediction, *Protein Engineering,* 11:523–538

[14] Z.-X. Wang and Z. Yuan (2000), How good is the prediction of protein structural class by the component-coupled method? *Proteins,* 38:165–175

[15] Y.D. Cai, X.J. Liu, X.B. Xu and K.C. Chou (2003), Support vector machines for prediction of protein domain structural class, *Journal of Theoretical Biology,* 221:115–120

[16] W.S. Bu, Z.P. Feng, Z. Zhang and C.T. Zhang (1999), Prediction of protein structural classes based on amino acid index, *European Journal of Biochemistry,* 266:1043–1049

[17] L. Jin, W. Fang and H. Tang (2003), Prediction of protein structural classes by a new measure of information discrepancy, *Comput. Biol. Chem.,* 27:373–380

[18] W.S. Bu, Z.P. Feng, Z. Zhang and C.T. Zhang (1999), Prediction of protein structural classes based on amino acid index, *European Journal of Biochemistry,* 266:1043–1049

[19] K.C. Chou and Y.D. Cai, Prediction protein structural class by functional domain composition, *Biochemical and Biophysical Research Communications,* 321 (2004), 1007–1009.

[20] L.A. Kurgan and L. Homaeian (2006), Prediction of structural classes for protein sequences and domains, impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, *Pattern Recognition,* 39(12): 2323-43, 2006

[21] W. R. Krigbaum and S. P. Knutton (1973), Prediction of the amounts of secondary structure in a globular protein from its amino acid composition, *Proc. Natl Acad. Sci.*, 70:2809-2813

[22] S. M. Muskal and S. H. Kim (1992), Predicting protein secondary structural content: a tandem neural network approach, *J. Molecular Biology*, 225:713-717

[23] Z. Zhang, Z. Sunt and C. Zhang (2001), A new approach to predict the helix/strand content of globular proteins, *Journal of Theoretical Biology,* 208:65-78

[24] Z. Lin and X. Pan (2001), Accurate prediction of protein secondary structural content, *Protein Chemistry,* 20(3):217-20

[25] U. Hobohm and C. Sander (1994), Enlarged representative set of protein structures, *Protein Science,* 3(3):522-524

[26] J. Ruan, K. Wang, J. Yang, L. Kurgan and K. Cios (2005), Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences, *Artificial Intelligent Medicine,* 35(1-2):19-35

[27] J. L. Fauchere and V. Pliska (1983), Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides, *European Journal of Medical Chemistry*, 18:369-375

[28] I.H. Witten and E. Frank (2005), *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann

[29] H.B. Shen, J. Yang, X-J. Liu and K.C. Chou (2005), Using supervised fuzzy clustering to predict protein structural classes, *Biochemical and Biophysical Research Communications*, 334:577–581

[30] K.Y. Feng, Y.D. Cai and K.C. Chou (2005), Boosting classifier for predicting protein domain structural class, *Biochemical and Biophysical Research Communications*, 334:213-217

[31] C. Sander, and R. Schneider (1991), Database of homology-derived structures and the structural meaning of sequence alignment, *Proteins*, 9:56-68

[32] B. Rost (1999), Twilight Zone of Protein Sequence Alignments, *Protein Engineering*, 12:85-94

[33] G. Pollastri and A. McLysaght (2005), Porter: A New, Accurate Server for Protein Secondary Structure Prediction, *Bioinformatics*, 21(8):1719-1720

[34] K. Lin, V.A. Simossis, W.R. Taylor, and J. Heringa (2005), A Simple and Fast Secondary Structure Prediction Method using Hidden Neural Networks, *Bioinformatics*, 21(2):152-9

[35] Y. Yang and J. O. Pedersen (1997), A comparative study on feature selection in text categorization, *Proceedings of the 14th International Conference on Machine Learning*, 412-420

[36] M.R. Sikonja and I. Kononenko (1997), An adaptation of Relief for attribute estimation on regression, *Proceedings of 14th International Conference on Machine Learning*, 296-304

[37] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R. Murthy (2001), Improvements to Platt's SMO algorithm for SVM classifier design, Neural Computations, 13(3):637–649

[38] S. le Cessie, J.C. van Houwelingen (1992), Ridge estimators in logistic regression, *Applied Statistics*, 41(1):191–201

[39] L. Breiman (2001), Random forests, *Machine Learning*, 45(1):5-32

[40] D. Aha and D. Kibler (1991), Instance-based learning algorithms, *Machine Learning*, 6:37–66

[41] K. Kedarisetti, L. Kurgan and S. Dick (2006), A Comment on 'Prediction of protein structural classes by a new measure of information discrepancy', *Computational Biology and Chemistry*, Elsevier, 30(5):393-394, 2006