Prediction of Secondary Protein Structure Content from Primary Sequence Alone – A Feature Selection Based Approach

Lukasz Kurgan and Leila Homaeian

University of Alberta, Department of Electrical and Computer Engineering,
Edmonton, Alberta, Canada T6G 2V4
{lkurgan,leila}@ece.ualberta.ca

Abstract. Research in protein structure and function is one of the most important subjects in modern bioinformatics and computational biology. It often uses advanced data mining and machine learning methodologies to perform prediction or pattern recognition tasks. This paper describes a new method for prediction of protein secondary structure content based on feature selection and multiple linear regression. The method develops a novel representation of primary protein sequences based on a large set of 495 features. The feature selection task performed using very large set of nearly 6,000 proteins, and tests performed on standard non-homologues protein sets confirm high quality of the developed solution. The application of feature selection and the novel representation resulted in 14-15% error rate reduction when compared to results achieved when standard representation is used. The prediction tests also show that a small set of 5-25 features is sufficient to achieve accurate prediction for both helix and strand content for non-homologous proteins.

1 Introduction

In the recent years increasing knowledge of protein structure accelerated medical research. Research in protein structure and interactions is of paramount importance to modern medicine, as it enhances general understanding of biological processes, and protein functions in particular. One of the most important related applied research and development areas is rational drug design, which aims to cut down costs and accelerate development process of drugs based on analytical models.

Protein structure can be learned by experimental and computational procedures. This paper develops a new computational method for prediction of protein secondary content. It proposes to perform prediction based on combination of feature selection procedure and data mining based approach. The proposed method extends the existing prediction methods by using a novel representation of primary protein structure. Comprehensive feature selection procedure performed with a very large set of almost 6,000 proteins resulted in development of an accurate prediction method that reduced error rates by 14-15% when compared to commonly used feature representation. Independent prediction on non-homogenous protein sets show that a small set of 5-25 features is sufficient to achieve high quality prediction models.

P. Perner and A. Imiya (Eds.): MLDM 2005, LNAI 3587, pp. 334 – 345, 2005.

In general, protein structure can be described on three levels: primary structure (Amino Acid (AA) sequence also called primer), secondary structure (folding of the primer into two-dimensional shapes, such as helices, strands, and various coils or turns), and tertiary structure (folding of the two-dimensional shapes into three-dimensional molecule). The Dictionary of Secondary Structures of Proteins annotates each AA as belonging to one of eight secondary structure types [4], which are typically reduced to three groups: helix, strand, and coil. The primary structure is currently publicly known for hundreds of thousands of proteins, e.g. NCBI protein database contains approximately 2 millions proteins, and SWISS-PROT database [3], stores over 159K primers. The secondary and tertiary structure is known for relatively small number of proteins, i.e. the Protein Data Bank (PDB) [1], currently contains about 30K proteins, out of which only a small portion have correct secondary structure and tertiary structure information. At the same time research in protein interactions and functions requires knowledge of tertiary structure. Experimental methods for discovery of secondary and tertiary structure such as X-ray crystallography and nuclear magnetic resonance spectroscopy are time consuming, labor expensive, and cannot be applied to some proteins [6]. Computational methods perform prediction of the tertiary structure with an intermediate step of predicting the secondary structure.

Computational methods for prediction of secondary structure from the primary sequence aim to close the existing gap between the number of known primary sequences and higher structures. One of the important pieces of information to perform prediction of secondary structure is protein content. While the secondary structure prediction aims to predict one of the three groups for each AA in the primary sequence, the secondary content prediction methods aim to predict amount of helix and strand structures in the protein. The secondary structure content can be learned experimentally by using spectroscopic methods, such as circular dichroism spectroscopy in the UV absorption range [13], and IR Raman spectroscopy [2]. Unsatisfactory accuracy and inconvenience of the experimental methods in some cases makes the computational approaches worth pursuing [20]. Computational methods have long history, and usually used statistical methods and information about AA composition of proteins to perform prediction.

This paper describes a novel approach that considers two aspects of content prediction task: quality of primary sequence representation and design of a prediction method. The existing methods, one the other hand, applied different prediction methods, but concentrated only on one dominant AA sequence representation. Secondary content prediction consists of two steps. First, primary sequence is converted into feature space representation, and next the helix and strand content are predicted using the feature values. A typical feature space representation consists of composition vector, molecular weight, and structural class, which are explained later. The first content prediction effort was undertaken in 1973 and used Multiple Linear Regression (MLR) method to predict content based on the composition vector [8]. A number of approaches, which used some combination of the composition vector, molecular weight, and structural class representation and neural network [10], analytic vector decomposition technique [5], and MLR method [17] [18] [19] [20] to predict the content were developed. A novel method that uses both composition vector and composition moment vector and a neural network was recently developed [12].

2 Proposed Prediction Method

The main difference between proposed and existing methods lies in the feature space representation used for prediction. The new method considers a large and diverse set of features, and performs feature selection to find optimal, in terms of quality of prediction and number of used features, representation. The existing methods consider very limited feature representation. After optimal representation is selected, the new method uses the most popular MLR for prediction of the content, see Figure 1.

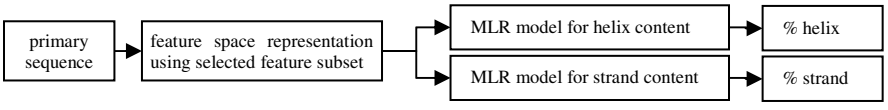

Fig. 1. Procedure for prediction of helix and strand content

The prediction is usually performed with an intermediate step when primary sequence is converted into feature space representation. The existing content prediction methods use a limited set of features while other methods, such as for prediction of protein structure or function, use a more diverse and larger number of features. This paper investigates if a more diverse set of features would help in content prediction. The considered set of features is summarized in Table 1 and later explained in detail.

Table 1. Features used to describe primary protein sequence and their applications

Feature	application type	reference(s)
Protein sequence length, avg molecular weight, avg isoelectric point	protein content and function prediction	[10] [14]
Composition vector	protein structure and content prediction	[5] [8] [10] [12] [17] [18] [19] [20]
1st order composition moment vector	protein content prediction	[12]
2nd order composition moment vector		
R-groups	protein structure and content prediction	[11]
Exchange groups	protein family and structure prediction	[15] [16]
Hydrophobicity groups	protein function prediction, structural and functional relationships	[7] [9] [14]
Electronic groups	protein structure prediction	[6]
Chemical groups	protein structure prediction	[6]
Other groups	protein function prediction, structural and functional relationships	[7] [14]
Dipeptides	protein function prediction	[14]

The properties include length, weight, and average isoelectric point. Protein length is defined as the number of AAs. To compute the molecular weight, the residue average weight values are summed and a water molecule mass is added. Average isoelectric point is computed using average isoelectric point values of all AAs in the primer; values are available at www.ionsource.com/virtit/VirtualIT/aainfo.htm. These features were used for protein content and function prediction [10] [14]. Composition vector is defined as composition percentage of each AA in the primary sequence. Composition moment vector takes into account position of each AA in the primary sequence [12]:

$$x_i^{(k)} = \frac{\sum_{j=1}^{K_i} n_{ij}^k}{N(N-1)....(N-k)} \quad (1)$$

where n_{ij} and x_i represent the j^{th} position of the i^{th} AA, and the composition of the i^{th} AA in the sequence, respectively; k is the order of the composition moment vector.

The 1^{st} and 2^{nd} orders were used, while the 0^{th} order reduces to the composition vector. Composition vector was used extensively for both protein structure and content prediction [5] [8] [10] [12] [17] [18] [19] [20], while composition moment vector was recently proposed for protein content prediction [12]. The property groups divide the AA into groups related to specific properties of individual AAs or entire protein molecule. Several different properties, such as hydrobhobicity, pI, electric charge, chemical composition, etc., that are summarized in Tables 2 and 3 are considered.

Table 2. Property based AA groups

Groups	Subgroups	AAs	Groups	Subgroups	AAs
R-group	Nonpolar aliphatic	AVLIMG	Hydropho-	Hydrophobic	VLIMAFPWYCG
	Polar uncharged	SPTCNQ	bicity group	Hydrophilic basic	KHR
	Positively charged	KHR		Hydrophilic acidic	DE
	Negative	DE		Hydrophilic polar with	STNQ
	Aromatic	FYW		uncharged side chain	
Exchange	(A)	C	Electronic	Electron donor	DEPA
group	(C)	AGPST	group	Weak electron donor	VLI
	(D)	DENQ		Electron acceptor	KNR
	(E)	KHR		Weak electron acceptor	FYMTQ
	(F)	ILMV		Neutral	GHWS
	(G)	FYW		Special AA	C
Other	Charged	DEKHRVLI	Other group	Tiny	AG
group	Polar	DEKHRNTQSYW		Bulky	FHWYR
	Aromatic	FHWY		Polar uncharged	NQ
	Small	AGST			

R-group combine hydropathy index, molecular weight and pI value together [11]. Exchange group represent conservative replacements through evolution. Hydrophobicity groups divide AAs into hydrophobic, which are insoluble or slightly soluble in water, in contrast with hydrophilic, which are water-soluble. Electronic group divides AAs based on their electronic properties, i.e. if they are neutral, electron donor or electron acceptor. Chemical group is associated with individual AAs. There are 19 chemical groups of which AAs are composed. Some of them are listed in Table 3. Other group considers the following mixed classes: charged, polar, aromatic, small, tiny, bulky, and polar uncharged. For each of the groups, the composition percentage of each subgroup in a protein sequence is computed. We note that these groups were extensively used for protein family, structure, function, prediction and to discover structural and functional relationships between proteins [6] [7] [14] [15] [16]. Finally, dipeptides are simply pairs of adjacent AAs in the primary sequence. The composition

Table 3. Chemical groups for AAs

AA	associated chemical groups
A	CH CO NH CH_3
C	CH CO NH CH_2 SH
D	CH CO NH CH_2 CO COO^-
E	CH CO NH CH_2 CH_2 CO COO^-

percentage of each pair is computed. They were previously used for protein function prediction[14].

2.1 Feature Selection for Protein Secondary Content Prediction

The above features were considered for prediction of protein secondary content. Initially correlation between features was investigated to find out if they are independent. The correlated features must be removed since they cannot be used with MLR model. Several correlated features were discovered. For example, some chemical subgroups were correlated with other features, such as composition vector, R-group, and other subgroups in the chemical group. The reason is that some chemical groups appear only in one AA or a group of AAs for which the composition percentage is computed in another feature. For example, COO– is found only in AAs D and E, which is identical to negative R-group, while some chemical groups always appear in the same AAs, such as C and NH_2. Table 4 shows final set of 495 features after removing overlapping and correlated features and provides abbreviation and indices that are used in the paper.

Table 4. List of features considered for feature selection

Feature	Abbr.	Indices
Protein sequence length	SL	1
Average molecular weight	MW	2
Average isoelectric point	IP	3
Composition vector (in alphabetical order)	CV	4-23
1st order composition moment vector (alphabetically)	MV1	24-43
2nd order Composition moment vector (alphabetically)	MV2	44-63
R-groups (*AVLIMG, SPTCNQ, KHR, DE, FYW*)	RG	64-68
Exchange groups (*AGPST, DENQ, ILM*)	XG	69-71
Hydrophobicity groups (*VLIMAFPWYCG, STNQ*)	HG	72-73
Electronic groups (*DEPA, LIV, KNR, FYMTQ, GHWS*)	EG	74-78
Chemical groups (*C, CAROM, CH, CH_2,, CH_2RING, CH_3,, CHAROM, CO, NH, OH*)	CG	79-88
Other groups (*DEKHRVLI, DEKHRNTQSYW, FHWY, AGST, AG, FHWYR NQ*)	OG	89-95
Dipeptides (alphabetically)	DP	96-495

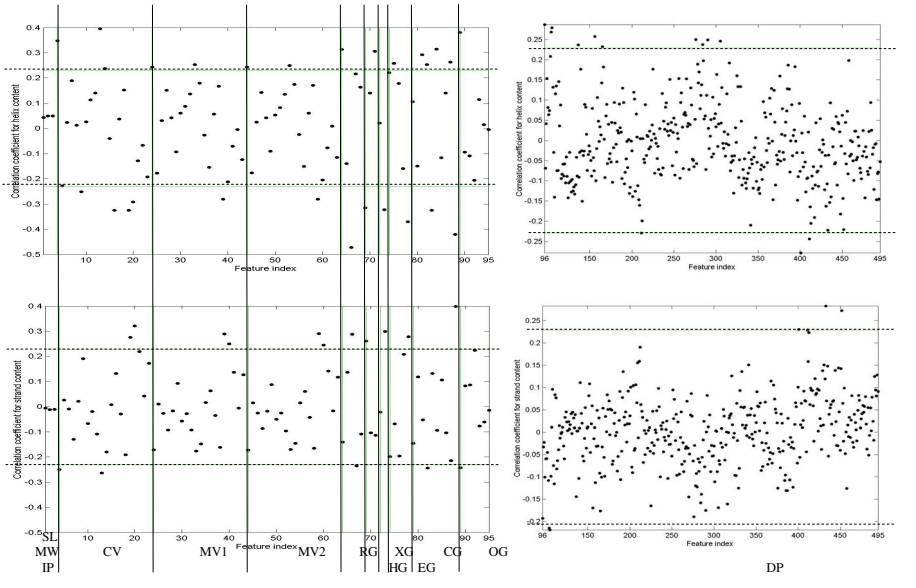

Fig. 2. Results of correlation test for the set of 495 features

Naïve Correlation Based Feature Selection. The simplest feature selection involves computing correlation between a given and the predicted feature and selection of a given number of features with highest correlations. Correlation values of 495 features and the helix and strand content were computed and are summarized in Figure 2.

Analysis of the figure shows that there are no strong correlations that can be used to select a suitable subset of features for content prediction. The strongest correlation values were in 0.3-0.4 range. None of the feature sets, i.e. physical properties, composition and composition moment vectors, property groups and dipeptides, can be evaluated as better or best correlated. The strongest correlated features, using correlation thresholds of 0.229 and -0.229, shown in Figure 2, are given in Table 5.

Table 5. The best correlated feature

structure	correlation	feature /values														
helix	negative	RG_2 0.47	CG_{10} 0.42	EG_5 0.37	CV_P 0.32	CG_5 0.32	CV_S 0.32	HG_2 0.32	XG_1 0.31	CV_T 0.29	$MV2_S$ 0.28	$MV1_S$ 0.27	DP_{SG} 0.27	CV_G 0.25	DP_{SS} 0.24	DP_{GS} 0.22
	positive	CV_L 0.39	OG_1 0.38	CV_A 0.34	CG_6 0.31	RG_1 0.31	XG_3 0.30	CG_3 0.29	DP_{AA} 0.28	DP_{AL} 0.27	DP_{AK} 0.26	CG_9 0.26	EG_2 0.25	DP_{EA} 0.25	$MV1_L$ 0.25	CG_4 0.25
		DP_{LA} 0.25	$MV2_L$ 0.24	DP_{LR} 0.24	DP_{ML} 0.24	$MV2_A$ 0.24	$MV1_A$ 0.24	DP_{LK} 0.23	CV_M 0.23	DP_{DA} 0.23	DP_{EL} 0.23					
strand	negative	CV_L 0.26	CV_A 0.25	CG_4 0.24	OG_1 0.24	RG_3 0.23										
	positive	CG_{10} 0.39	CV_T 0.32	HG_2 0.29	$MV2_S$ 0.28	$MV1_S$ 0.28	RG_2 0.28	DP_{TY} 0.28	EG_5 0.27	CV_S 0.27	DP_{VT} 0.27	XG_1 0.25	$MV1_T$ 0.24	$MV2_T$ 0.24	DP_{SG} 0.23	

Correlation accommodates only for correlation between individual features and the predicted values, while more complex correlation that include multiple features together exits. Therefore regression based correlation feature selection was performed.

Regression-Correlation Based Feature Selection. The feature selection was performed according to the procedure shown in Figure 3.

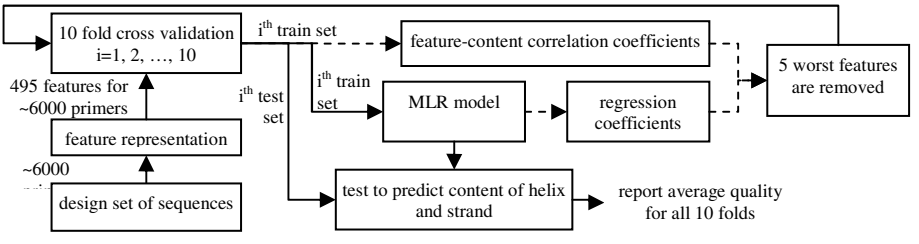

Fig. 3. Feature selection procedure performed independently for helix and strand content

Feature selection is performed independently for helix and strand content prediction. It uses dataset of about 6000 proteins extracted from PDB (described later) to investigate two selection procedures. Each of the 6000 primers is first converted into 495 features representation. Next, the dataset is split in the 10-fold cross validation (10CV) manner, and MLR model is computed for each fold. The model is tested with the test set and average quality over 10 folds is reported. Next, five worst features are

selected and removed, and remaining 490 features are used again to perform MLR. The process repeats until 5 features are left. In each iteration the worst 5 features are selected according to two independent criterions: 1) smallest values of corresponding regression coefficients, 2) smallest values of correlation coefficients between a given feature and the helix/strand content. Both sets of coefficients are recomputed and averaged for each cross-validation fold. The regression coefficients are constants of the MLR model. We assume that the lower the coefficient value the lesser the corresponding feature's impact on the predicted helix or strand content, and therefore the less useful it is for prediction. Similarly correlation coefficients express correlation between a given feature and the predicted helix or strand values. Again, the lower the correlation values the less useful the feature is. The main difference between the coefficient sets is that the MLR considers all features together while the correlation considers each feature independently. The results are discussed in the next section.

3 Experiments

Experiments apply 3 datasets, one for feature selection and two for validation of the developed method and comparison with other prediction methods. The feature selection dataset was extracted from PDB (release as of August 12th 2004) to cover wide range of known proteins. For proteins that have isotopes, the last one was selected. The proteins were filtered according to a set of rules shown in Table 6 to eliminate errors and inconsistent data. Also, sequences with identical primer and different secondary sequences were eliminated. Lastly, sequences with ambiguous AAs in the primer, i.e. B or Z, were removed resulting in a dataset with 5834 sequences that include homologous sequences. The length of the shortest sequence is 6 and of the longest sequence is 1295. The test datasets include:

Table 6. Filters used to derive feature selection dataset

Type of the Problem	# seq	Type of the Problem	# seq
Sequence length < 4	455	Helix indexed out of sequence	10038
Illegal AA	11540	Strand indexed out of sequence	8023
residue called *UKN*	25	Coil indexed out of sequence	219
More/less residues than	9	Overlap of helix and strand	782
the sequence length		Overlap of helix and coil	1342
Helix of length < 3	1291	No secondary structure	9972
Strand of length < 2	19022	No primary structure	13

- 210 non-homologous proteins set described in [20]. Although these proteins satisfy criteria defined in Table 6, 11 proteins were excluded from experiments, since they include unknown AA X in their primer in the newest PDB release. Therefore 199 proteins were used. The excluded proteins are: 1MBA_, 1MDC_, 1OPAA, 4SBVA, 1FBAA, 1ETU_, 1GP1A, 3ADK_, 1CSEI, 1ONC_, 1FUS_.
- 262 non-homologous proteins set described in [5]. Among the original set only 52 proteins were found in newest PDB release and satisfied criteria from Table 6.

Feature selection was performed using two approaches to select worst performing features for deletion, one based on correlation and the other based on regression coefficients. The content prediction quality was evaluated using two measures [20]:

$$e = \frac{\sum_{k=1}^{N} | F_k - D_K |}{N}, \quad \sigma = \sqrt{\frac{\sum_{k=1}^{N} (e - | F_k - D_K |)^2}{N-1}} \tag{2}$$

where e is an average error, σ is standard deviation, F_K is the predicted helix or strand content, D_K is the known content, and N is number of predicted proteins.

Results are shown in Figure 4. Each experiment involves 10CV. Feature selection results are based on computation of about 4000 MLR models. The optimal, in terms of trade-off between error e and number of features, subsets are shown by dashed lines. For both prediction of strand and helix content 4 subsets were selected: for the lowest error value (L), for the last five features (F), for a feature subset of small size (S), and for the best relative ratio between error and feature subset size (M).

The results for selected 4 datasets for both correlation and regression coefficient based approaches and helix and strand prediction are given in Table 7. It shows that minimum error for helix and strand content prediction is 11.28% and 8.67% respectively, and was achieved for regression based selection for dataset L. The maximum error when using just last 5 features is 15.16% and 11.48% for helix and strand

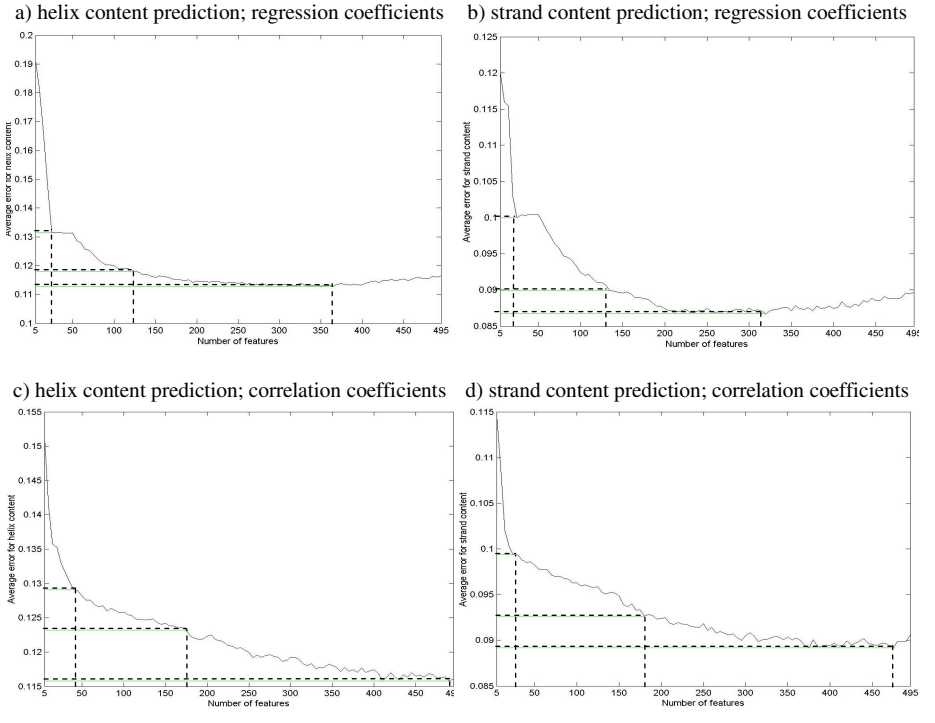

Fig. 4. Results for feature selection experiments using regression and correlation coefficients

prediction respectively, and was achieved for correlation based selection. Therefore 25% error reduction is achieved by using dataset L instead of F. Datasets S and M give relatively good tradeoff between prediction error and number of features. Dataset S with just 25 features for helix prediction gives 12.91% error, while for strand it gives 10% error. Similarly for M dataset, 125 features are used to predict helix content with 11.81% error while 135 features to predict strand content with 8.99% error.

The regression coefficients based selection gives better results for M and L datasets, while correlation coefficients based selection is better for small subsets. This results agrees with our expectations, since regression method benefits from relationships between features, while correlation based method considers each feature independently.

Table 7. Summary of feature selection results (best and worst results are shown in bold; baseline results achived when composition vector is used are shown in italics)

dataset name			5Features (F)	Small (S)	Medium (M)	Large (L)
regression	helix	# features	5	25	125	365
coeff.	prediction	e(σ)	**0.1917** (.0171)	0.1315 (.0116)	0.1181 (.0106)	**0.1128** (.0102)
selection	strand	# features	5	25	135	320
	prediction	e(σ)	**0.1203** (.0080)	0.1000 (.0067)	0.0899 (.0059)	**0.0867** (.0057)
correlation	helix	# features	5	40	175	490
coeff.	prediction	e(σ)	0.1516 (.0139)	0.1291 (.0117)	0.1232 (.0111)	0.1158 (.0115)
selection	strand	# features	5	25	180	475
	prediction	e(σ)	0.1148 (.0074)	0.0994 (.0067)	0.0926 (.0063)	0.0892 (.0063)
composition	helix	# features	*20*	strand	# features	*20*
vector	prediction	e(σ)	*0.1329 (.0117)*	prediction	e(σ)	*0.1011 (.0067)*

Another experiment, which involves 10CV prediction using 20 features composition vector to predict the content, was performed, see Table 8. Since composition vector is the most utilized feature set for prediction (all published results use it for prediction [5] [8] [10] [12] [17] [18] [19] [20]) this results gives a baseline to verify that feature selection procedure improves the existing prediction approaches. For the helix prediction a slight improvement of about 0.5% (which translates into 3% error rate reduction) was achieved by using S subsets consisting of 40 features. The 2% error rate improvement (which translates into 15% error rate reduction) was achieved

Table 8. Comparision of error rates for prediction of secondary structure content for different methods and for different considered feature subsets (best results shown in bold; baseline results shown in italics)

method	test dataset (reference)	feature subset	Resubstitution e(σ)		Jackknife e(σ)	
			helix	strand	helix	strand
this paper	199 out of 210 [20]	F_{reg}	0.176 (.017)	0.125 (.007)	0.181 (.018)	0.128 (.008)
MLR		S_{reg}	**0.143** (.013)	0.106 (.006)	**0.166** (.018)	0.123 (.008)
		M_{reg}	0.092 (.005)	0.052 (.001)	0.263 (.044)	0.178 (.020)
		F_{corr}	0.171 (.016)	0.126 (.007)	0.176 (.017)	0.130 (.007)
		S_{corr}	0.144 (.012)	**0.103** (.006)	0.185 (.021)	**0.119** (.008)
		M_{corr}	0.051 (.001)	0.030 (.000)	0.514 (.167)	0.354 (.065)
		CV	*0.148 (.014)*	*0.110 (.005)*	*0.167 (.018)*	*0.123 (.007)*
this paper	52 out of 262 [5]	F_{reg}	0.164 (.001)	0.156 (.016)	0.190 (.024)	0.179 (.023)
MLR		S_{reg}	0.115 (.006)	0.098 (.004)	0.240 (.029)	0.208 (.024)
		F_{corr}	**0.164** (.014)	**0.154** (.012)	**0.189** (.021)	**0.175** (.017)
		S_{corr}	0.085 (.004)	0.095 (.005)	0.448 (.126)	0.193 (.022)
		CV	*0.118 (.005)*	*0.109 (.005)*	*0.211 (.022)*	*0.194 (.021)*
AVDM-1	262 [5]	CV	0.144 (.117)	0.118 (.096)	0.145 (.017)	0.120 (.097)
AVDM-2		CV	0.132 (.109)	0.114 (.096)	0.142 (.115)	0.124 (.105)
MLR	210 [20]	CV	0.122 (.089)	0.108 (.082)	0.135 (.103)	0.120 (.097)

by using L subset consisting of 265 features. For the strand prediction, the 0.2% error rate improvement was achieved for the 25 features subset, while 1.4% improvement (14% error rate reduction) was achieved when 365 features were used. Although the achieved improvement seem small, the 15% and 14% error rate reduction in medically related field should be perceived as a significant result, especially that it is backed up by a study that considers a large and comprehensive set of proteins.

Prediction tests were performed to test selected feature subsets. Subsets F_{reg}, S_{reg}, and M_{reg} for regression coefficients and F_{corr}, S_{corr}, and M_{corr} for correlation coefficients based selection were used to perform independent test on the test datasets. Prediction of the secondary content was performed using MLR method. In case of regression number of data points (proteins in the dataset) should be larger than number of features. Therefore for 52 protein dataset only F and S subsets were considered.

Test consists of resubstitution and jackknife procedures [20] The fist procedure trains and test on the same dataset, while the other is a leave-one-out test. Test results are summarized and compared with other methods in Table 8. The table also includes results for MLR based prediction when the standard composition vector (CV) feature set is used. Since resubstitution test trains and test on the same data, it is prone to overfitting. Thus analysis concentrates on jackknife test results. Baseline results that apply composition vector are always worse that the best results achieved by the generated feature subsets. Subset S_{reg} generates slightly better results for helix prediction, while subset S_{corr} is better in case of strand prediction for the set of 199 proteins. Similarly models generated using subset F_{corr} reduce error rates for both helix and strand prediction by over 10% in case of the 52 protein set (18.9% error rate was achieved for F_{corr} while 21.1% was achieved for composition vector for helix prediction, while 17.5% and 19.4% error rates were achieved for strand prediction respectively). Subsets F that contain only 5 features achieve better results that prediction using 20 features composition vector. The results justify feature selection as a useful method not only to improve prediction results, but also to possibly reduce the number of features necessary for the secondary content prediction. The selected subsets F and S for both correlation and regression based feature selection are listed in Table 9.

Table 9. Selected subsets of features

Data	Struct.	Features
F_{reg}	helix	CV_{12} CV_{14} CV_{20} OG_3 OG_7
	strand	CV_1 CV_{11} CV_{12} CV_{14} OG_7
S_{reg}	helix	CV_2 CV_5÷CV_{12} CV_{14} CV_{17} CV_{18} CV_{19} CV_{20} RG_1 XG_2 XG3 EG_1 EG_2 EG3 CG_6 OG_2 OG3 OG_6 OG_7
	strand	CV_1÷CV_{20} RG_1 RG_2 RG_4 RG_5 XG_1 XG_2 XG_3 HG_1 HG_2 EG_1÷EG_5 CG_1÷CG_{10} OG_1÷OG_7 DP_{13} DP_{20} DP_{22} DP_{29} DP_{30} DP_{31} DP_{32} DP_{34} DP_{35} DP_{38} DP_{46} DP_{58} DP_{66} DP_{67} DP_{71} DP_{73} DP_{78} DP_{81} DP_{82} DP_{87} DP_{89} DP_{92} DP_{93} DP_{95} DP_{97} DP_{99} DP_{100} DP_{108} DP_{114} DP_{122} DP_{132} DP_{135} DP_{139} DP_{162} DP_{170} DP_{173} DP_{178} DP_{179} DP_{193} DP_{214} DP_{226} DP_{238} DP_{244} DP_{256} DP_{257} DP_{266} DP_{267} DP_{270} DP_{273} DP_{277} DP_{278} DP_{279} DP_{285} DP_{286} DP_{290} DP_{293} DP_{298} DP_{304} DP_{308} DP_{321} DP_{326} DP_{327} DP_{330} DP_{331} DP_{334} DP_{338} DP_{339} DP_{352} DP_{353} DP_{362} DP_{364} DP_{368} DP_{369} DP_{372} DP_{374} DP_{375} DP_{376} DP_{378} DP_{379} DP_{380} DP_{384} DP_{391} DP_{398} DP_{399}
F_{corr}	helix	CV_{10} RG_3 EG_5 CG_{10} OG_1
	strand	CV_{17} $MV1_{16}$ $MV2_{16}$ HG_2 CG_{10}
S_{corr}	helix	CV_1 CV_5 CV_{10} CV_{11} CV_{13} CV_{16} CV_{17} $MV1_1$ $MV1_{10}$ $MV1_{16}$ $MV2_1$ $MV2_{10}$ $MV2_{16}$ RG_1 RG_3 XG_1 XG_3 HG_2 EG_2 EG_5 CG_3 CG_4 CG_5 CG_6 CG_9 CG_{10} OG_1 DP_1 DP_9 DP_{10} DP_{41} DP_{61} DP_{70} DP_{116} DP_{181} DP_{189} DP_{195} DP_{210} DP_{306} DP_{316}
	strand	CV_1 CV_{10} CV_{16} CV_{17} CV_{18} $MV1_{16}$ $MV1_{17}$ $MV2_{16}$ $MV2_{17}$ RG_3 RG_4 XG_1 HG_2 EG_5 CG_4 CG_{10} OG_1 OG_4 DP_9 DP_{10} DP_{306} DP_{316} DP_{318} DP_{338} DP_{357}

Results achieved for subset S_{corr} for strand prediction are better than results of both AVDM and MLR methods, while the existing methods are better in case of helix content prediction, see Table 9. The AVDM method uses more advanced predictive model called analytic vector decomposition technique [5]. The MLR method uses MLR method, as in [8], but tests on the set of all 210 proteins. We anticipate that using more advanced prediction model in combination with feature selection performed in this paper would results in system that surpasses the existing approaches.

4 Summary and Future Work

The paper presents a novel method for prediction of protein secondary structure content. The method is the first to consider alternative feature representation of primary protein sequences. It performs feature selection task to generate optimal, in terms of trade-of between prediction error rates and number of feature, feature representation and performs MLR based prediction of the helix and strand protein content. The results based on the leave-one-out test for non-homologous protein sets show that not only 5-25 features set can be used to predict the secondary content values, but that the representation based only on 5 features can reduce error rates by 10% when compared to standard 20 features representation based on composition vector. The results for a comprehensive set of 6000 mixed homologous and non-homologous proteins also show that error rate reduction of 14-15% can be achieved when the proposed feature representation is used instead of standard composition vector based representation.

Future work will design 2-layer prediction system. First, protein structural class (α, β, and $\alpha\beta$) will be predicted and next specialized prediction models for each class and predicted structure will be used. Design is similar to [17] [18] [19] [20], but considers that structural class will be predicted, not assumed, and utilizes feature sselection.

References

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

[2] Bussian B., & Sender, C., How to Determine Protein Secondary Structure in Solution by Raman Spectroscopy: Practical Guide and Test Case DNsae I, *Biochem.*, 28, 4271-77, 1989

[3] Boeckmann B., et al., The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003, *Nucleic Acids Research*, 31, 365-370, 2003

[4] Dwyer D., Electronic Properties of Amino Acids Side Chains Contribute to the Structural Preferences in Protein Folding, *J Bimolecular Structure & Dynamics*, 18:6, 881-892, 2001

[5] 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

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

[7] Hobohm U., & Sander C., A Sequence Property Approach to Searching Protein Databases, *J. of Molecular Biology*, 251, 390-399, 1995

[8] Krigbaum W., & Knutton S., Prediction of the Amount of Secondary Structure in a Globular Protein from its Amino Acid Composition, *Proc. of the Nat. Academy of Science*, 70, 2809-2813, 1973

[9] Lodish H., et al., *Molecular Cell Biology*, 4th ed., W.H. Freeman & Company, New York, 50-54, 2000

[10] Muskal S.M., & Kim S-H., Predicting Protein Secondary Structure Content: a Tandem Neural Network Approach, *J. of Molecular Biology*, 225, 713-727, 1992

[11] Nelson D. & Cox M., *Lehninger Principles of Biochemistry Amino*, Worth Publish., 2000

[12] Ruan J. et al., Highly Accurate and Consistent Method for Prediction of Helix and Strand Content from Primary Protein Sequences, *Artificial Intelligence in Medicine*, special issue on *Computational Intelligence Techniques in Bioinformatics*, accepted, 2005

[13] Sreerama N., & Woody, R.W., Protein Secondary Structure from Circular Dichroism Spectroscopy, *J Molecular Biology*, 242, 497-507, 1994

[14] Syed U., & Yona G., Using a Mixture of Probabilistic Decision Trees for Direct Prediction of Protein Function, *Proc. of RECOMB 2003 Conf.*, 224-234, 2003

[15] Wang, J., et al., Application of Neural Networks to Biological Data Mining: a Case Study in Protein Sequence Classification, *Proc. of 6th ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining*, 305-309, 2000

[16] Yang, X., & Wang, B., Weave Amino Acid Sequences for Protein Secondary Structure Prediction, *Proc. of 8th ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery*, 80-87, 2003

[17] Zhang, C.T., Zhang, Z., & He. Z., Prediction of the Secondary Structure of Globular Proteins Based on Structural Classes, *J. of Protein Chemistry*, 15, 775-786, 1996

[18] Zhang, C.T., et al., Prediction of Helix/Strand Content of Globular Proteins Based on Their Primary Sequences, *Protein Engineering*, 11:11, 971-979, 1998a

[19] Zhang C.T., Zhang Z., & He Z., Prediction of the Secondary Structure Contents of Globular Proteins based on Three Structural Classes, *J Protein Chemistry*, 17, 261-272, 1998b

[20] Zhang Z.D., Sun Z.R., & Zhang C.T., A New Approach to Predict the Helix/Strand Content of Globular Proteins, *J Theoretical Biology*, 208, 65-78, 2001