Prediction of Protein Structural Class Using PSI-BLAST Profile Based Collocation of Amino Acid Pairs

Ke Chen[1], Lukasz Kurgan*[1] and Jishou Ruan[2]

[1]University of Alberta, Department of Electrical and Computer Engineering, Edmonton, CANADA, T6G 2V4
[2]Chern Institute of Mathematics, College of Mathematical Science & LPMC, Nankai University, Tianjin, PRC 300071
*lkurgan@ece.ualberta.ca (corresponding author)

Abstract –**Knowledge of structural classes is useful in understanding folding patterns in proteins. Numerous structural class prediction methods were proposed in the past. Although virtually all state-of-the-art classifiers were already tried, many of these methods use very simple protein sequence representation that often includes amino acid (AA) composition. To this end, we propose a novel sequence representation, which is based on PSI-BLAST profile based collocation of AA pairs. We used two benchmark datasets constructed by Zhou (*J. of Prot. Chem*. 1998, 17(8):729–38) to test the proposed representation with five representative classifiers. The two best classifiers, which include a support vector machine and an instance base learner, achieved 88% and 96% accuracy on the two datasets, respectively. Our results were compared with five recently proposed methods. The comparison shows superiority of the proposed method, which reduces the error rates by 30% and 21% on the two datasets when compared with the best-performing ensemble of boosted logistic regression classifier. Finally, the new sequence representation is compared with AA composition when using support vector machine classifier. The error rate reduction due to application of the new representation equals 40% and 25% for the two datasets, respectively. In short, the PSI-BLAST profile based collocation of AA pairs is shown to be a promising feature-based sequence representation.**

Keywords*: domain structural class; collocation of AA pairs; PSI-BLAST; machine learning*

1. Introduction and Related Work

Knowledge of protein structure plays a crucial role in analysis of protein function, simulation of protein-ligand interaction, rational drug discovery and in many other applications. Prediction of the tertiary protein structure remains a challenge even though it is being researched for over two decades. There are numerous, supplementary aspects of protein structure, which include secondary structure, solvent accessibility, contact maps and structural class, which prediction is actively pursued by many structural biology and bioinformatics research labs.

The concept of protein structural class was proposed by Levitt and Chothia in 1976 [1]. They inspected and classified 31 globular proteins into four structural classes: all-α, all-β, α/β, and α+β. The all-α and all-β classes represent structures that mainly consist of α-helices and β-strands, respectively. The α/β and α+β classes contain both α-helices and β-strands; the α/β class includes mainly parallel β-strands, while α+β class includes anti-parallel strands. The most frequently used classifications of protein structural classes can be found in the SCOP (Structural Classification of Protein) database [2]. This database is organized as a hierarchy of known protein and protein domain structures where first level is based on the structural class.

The last twenty years observed significant efforts in automated prediction of protein structural classes due to large numbers (currently over 3 millions) of unclassified sequences. Several early attempts were made in late 1980s [3, 4]. Composition vectors, auto-correlation function based on non-bonded residue energy, polypeptide composition, pseudo amino acid composition and complexity measure factor were applied to represent protein sequence in later works [5-10]. Different classification algorithms, including the maximum component coefficient [11], least correlation angle [12], fuzzy clustering [13], neural network [14], Bayesian classification [15], rough sets [16], component-coupled [5] and support vector machine [17], have been already used. Recent works also explored application of complex classification models, such as ensembles [18], bagging [19] and boosting [20].

Since virtually all state-of-the-art classifiers have been already tried, we concentrate on development of a novel representation of protein sequences based on a PSI-BLAST profile [21]. This profile has been successfully applied in window based protein prediction tasks, including secondary structure and solvent accessibility predictions [22, 23], while it was never applied to predict the structural class. We propose a novel method that transforms the original profile ($N\times 20$ matrix where N is the sequence length), into a fixed length feature-vector based on a recently proposed collocation of AA pairs [24] that is calculated from the sequence. This novel representation is shown to substantially improve the accuracy of the structural class prediction.

2. Materials and Methods

2.1. Dataset

Two datasets used in this paper were originally generated by Zhou [5]. Both the datasets were used in several past studies [5, 14, 16, 17, 20] and were extracted from SCOP. They include 277 and 498 protein domains, respectively. Both datasets are balanced, i.e., each class includes similar number of sequences.

2.2. Proposed Sequence Representation

The new representation, which combines PSI-Blast profile and the concept of frequency of *collocation of AA pairs* in the sequence [24], was developed for the proposed prediction method.

The motivation to introduce the collocation of AA pairs comes from an insufficient sequence representation that is offered by the commonly used *composition vector*, i.e., it only counts the frequencies of individual AAs. At the same time, frequencies of AA pairs (dipeptides) provide more information since they may reflect local (with respect to the sequence) interaction between AA pairs. Based on this argument we should count all dipeptides in the sequence. Since there are 400 possible AA pairs (*AA*, *AC*, *AD*, …, *YY*), a feature vector of that size is used to represent occurrence of these pairs in the sequence. Since short-range interactions between AAs, rather than only interactions between immediately adjacent AAs, have impact of folding [25], the proposed representation also considers collocated pairs of AAs, i.e. pairs that are separated by p other AAs. These pairs can be understood as the dipeptides with gaps. Collocated pairs for p = 0, 1,…, 4 are considered, where for p=0 the pairs reduce to the dipeptides. There are 400 feature values for each value of p.

On the other hand, the successful applications of PSI-BLAST profile illustrate that it contains more information than a query sequence. PSI-BLAST aligns a given query sequence to a database of sequences, and searches for these that are similar to the query sequence. Using multiple alignment, PSI-BLAST generates the frequency of each AA at each position in the query sequence. The PSI-BLAST profile generates 20-dimensional vector of AA frequencies for each position in the query sequence, which can be used to identify the key positions of conserved AAs. In other words, the profile can help in identifying which residues (segments) are conserved and which undergo mutations.

Our approach combines the frequency of collocation of AA pairs and the PSI-BLAST profile into so called *PSI-BLAST profile based collocation of AA pairs*. The PSI-BLAST profile is the N×20 matrix, which is denoted as $[a_{i,j}]$, where i=1,2,…,N denotes position in the query sequence and j=1,2,…,20 denotes a given AA. After applying the substitution matrix and log function, a_{ij} values range between -9 and 11. The proposed representation is related to calculation of the frequency of AA pairs based on binary coding. The binary coding uses a 20-dimensional vector to encode each AA. The 20 AAs can be represented as AA_1, AA_2, …, AA_{19}, and AA_{20}. In binary coding, AA_i is encoded as (0,0,…,0,1,0,…,0,0), where only the i^{th} value is greater than 0. The binary coding matrix is denoted as $[b_{i,j}]$. The binary encoding and PSI-Blast profile matrices have the same dimensionality (N×20).

The frequency of AA pairs can be computed from the binary coding matrix. For a given protein sequence $A_1A_2...A_N$

A_iA_{i+1} is a AA_mAA_n dipeptide

$\Leftrightarrow A_i=AA_m$ and $A_{i+1}=AA_n$

$\Leftrightarrow b_{i,m}=1$, $b_{i+1,n}=1$, $b_{i,p}=0$, $b_{i+1,q}=0$, where $p\neq m$ and $q\neq n$

Given that $c_{s,t} = \min(b_{i,s}, b_{i+1,t})$, then

$$c_{s,t} = \begin{cases} 1 & (iff\ s=m, t=n) \\ 0 & (\text{else}) \end{cases}$$

which means that AA_mAA_n was counted once while other dipeptides were counted 0 times. Matrix $[c_{s,t}]$ stores the frequencies of all dipeptides. The count of the AA pairs along the entire sequence can be computed as

$$c_{s,t} = \sum_{i=1}^{N-1} \min(b_{i,s}, b_{i+1,t})$$

The PSI-BLAST profile based collocation of AA pairs is calculated in similar way. The only difference is that the binary coding matrix $[b_{i,j}]$ is replaced by the PSI-Blast profile $[a_{i,j}]$. The frequency of dipeptide AA_sAA_t is computed as $c_{s,t} = \sum_{i=1}^{N-1} \min(a_{i,s}, a_{i+1,t})$ and matrix $[c_{s,t}]$ stores the frequencies of all dipeptides.

Since the PSI-BLAST profile values can be negative, and the frequencies of AA pairs should not be negative, using $\min(a_{i,s}, a_{i+1,t})$ function to represent the frequency of AA pairs is unsound. Instead, we define

$$c_{s,t} = \sum_{i=1}^{N-1} \max(0, \min(a_{i,s}, a_{i+1,t}))$$

in which the negative value of $\min(a_{i,s}, a_{i+1,t})$ is replaced by 0. Similarly, the frequencies of p-collocated AA pairs are defined as

$$d_{s,t,p} = \sum_{i=1}^{N-p-1} \max(0, \min(a_{i,s}, a_{i+p+1,t}))$$

The matrixes $[c_{s,t}]$ and $[d_{s,t,p}]$, which correspond to the frequency of the PSI-BLAST profile based dipeptides and p-collocated AA pairs, respectively, constitute the proposed protein sequence representation. We generate PSI-BLAST profile based collocation of AA pairs for p = 0, 1, 2, 3, and 4, which results in 2000 features per each sequence. Since the proposed representation includes relatively large number of features, a feature selection method was used to reduce the dimensionality and potentially improve the prediction accuracy. An entropy based feature selection method, which evaluates each feature by measuring the information gain with respect to the class, was used.

Table 1. Comparison of jackknife test accuracy between different classifiers for the proposed sequence representation that includes the selected 50 features. The best results are shown in bold.

Dataset	Algorithm	Jackknife accuracy [%]				
		all-α	all-β	α/β	α+β	overall
277 domains	SVM (3rd degree polynomial)	91.18	91.38	93.42	76.92	**87.73**
	Logistic Regression	76.47	79.66	87.01	64.62	77.32
	IB1	89.71	88.14	92.21	80.00	**87.73**
	C4.5	73.53	74.58	79.22	73.85	75.46
	Naïve Bayes	67.65	77.97	85.71	66.15	74.72
498 domains	SVM (3rd degree polynomial)	97.98	93.33	95.62	93.43	**94.93**
	Logistic Regression	95.96	95.83	94.16	90.51	93.91
	IB1	94.95	95.83	97.81	94.16	**95.74**
	C4.5	89.90	89.17	94.89	91.24	91.48
	Naïve Bayes	80.81	92.50	94.89	82.48	88.03

Table 2. Comparison of jackknife test accuracy between the two best classifiers (using the proposed representation with 50 features) and other reported methods. The best results are shown in bold.

Dataset	Algorithm	Reference	Feature-based sequence representation	Jackknife accuracy [%]				
				all-α	all-β	α/β	α+β	overall
277 domains	Rough sets	[16]	AA composition & physicochemical properties	77.1	77.0	93.8	66.2	79.4
	Component-Coupling	[5]	AA composition	84.3	82.0	81.5	67.7	79.1
	Neural Network	[14]	AA composition	68.6	85.2	86.4	56.9	74.7
	SVM	[17]	AA composition	74.3	82.0	87.7	72.3	79.4
	LogitBoost	[20]	AA composition	81.4	88.5	92.6	72.3	84.1
	SVM	(this paper)	PSI-BLAST based p-collocated AA pairs	91.2	91.4	93.4	76.9	**87.7**
	IB1	(this paper)	PSI-BLAST based p-collocated AA pairs	89.7	88.1	92.2	80.0	**87.7**
498 domains	Rough sets	[16]	AA composition & physicochemical properties	87.9	91.3	97.1	86.0	90.8
	Component-Coupling	[5]	AA composition	93.5	88.9	90.4	84.5	89.2
	Neural Network	[14]	AA composition	86.0	96.0	88.2	86.0	89.2
	SVM	[17]	AA composition	88.8	95.2	96.3	91.5	93.2
	LogitBoost	[20]	AA composition	92.5	96.0	97.1	93.0	94.8
	SVM	(this paper)	PSI-BLAST based p-collocated AA pairs	98.0	93.3	95.6	93.4	94.9
	IB1	(this paper)	PSI-BLAST based p-collocated AA pairs	95.0	95.8	97.8	94.2	**95.7**

The entropy of a feature X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

where $\{x_i\}$ is a set of values of X and $P(x_i)$ is the prior probability of x_i. The conditional entropy of X, given another feature Y (in our case the structural class) is defined as

$$H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j))$$

where $P(x_i| y_j)$ is the posterior probability of X given the value y_i of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain

$$IG(X \mid Y) = H(X) - H(X \mid Y)$$

According to this measure, Y has stronger correlation with X than with Z if $IG(X|Y)>IG(Z|Y)$. The feature selection was performed using 10-fold cross validation to avoid overfitting. Among the original set of 2000 features, the best 50 features, which give the highest IG values, were selected.

3. Results and Discussion

The classification systems used to develop and compare the proposed method were implemented in Weka [26]. The proposed new feature representation was tested with several state-of-the-art classifiers such as Support Vector Machine (SVM) [27], Multiple Logistic Regression [28], instance learning based IB1 algorithm [29], Naïve Bayes [30], and C4.5 decision tree [31], using the selected 50 features to represent sequences. We also compare these results with previous studies that used the same datasets, and different sequence representations and classifiers. All experiments were performed using jackknife test.

Table 1 shows the results when using the proposed sequence representation and the five selected classifiers. For both datasets, IB1 and SVM classifiers provide comparable and highest overall accuracy, i.e., 88% and 95-96% for the 277 and 498 datasets, respectively. The other three classifiers are substantially worse, i.e., over 10% and 1-6% worse accuracy for the 277 and 498 datasets, respectively.

The best-performing IB1 and SVM classifiers were further compared with other recently reported methods, such as rough sets [16], component-coupling algorithm [5], neural network [14], SVM [17] and ensemble of boosted logistic regression classifiers [20], see Table 2. The proposed representation results in substantial error rate reduction for

both datasets when compared with best previously reported results for the LogitBoost method, i.e., 3.6/12.3 = 30% and 0.9/4.3 = 21% error rate reduction for the 277 and 498 datasets, respectively. Our predictions give comparably high accuracy for all four structural classes, i.e., IB1 does not produce accuracy below 80% for 277 sequences, while accuracy of both SVM and IB1 for the 498 dataset does not decrease below 93%, for any of the four classes.

We also compare the quality of the commonly used AA composition and the proposed sequence representations. For one of the best performing SVM classifier, the PSI- BLAST profile based collocation of AA pairs gives 8.2% and 1.7% higher accuracy for the 277 and 498 sequences datasets, respectively, i.e., this corresponds to the 40% and 25% error rate reduction, respectively.

Finally, our best classifiers were relatively simple (single SVM and an instance base method) when compared with the best among the reported methods ensemble based classifier. This leaves some room for future improvements.

4. Conclusion

The proposed PSI-BLAST profile based collocation of AA pairs is a novel and promising feature representation. Our empirical tests show that the accuracy of the protein structural class prediction can be substantially improved by applying this representation, i.e., relatively simple classifiers that use the proposed features provide better accuracy than more complex classifiers on two benchmark datasets. The new representation can be extended to other protein prediction tasks that currently apply AA composition to improve their accuracy.

5. References

[1] Levitt M, Chothia C. Structural patterns in globular proteins, *Nature* 1976, 261(5561):552–8

[2] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*. 1995, 247(4):536–40

[3] Klein P, Delisi C, Prediction of protein structural class from the amino-acid sequence, *Biopolymers* 1986, 25:1659–1672

[4] Nakashima H, Nishikawa K, Ooi T, The folding type of a protein is relevant to the amino acid composition*, J. Biochem*. 1986, 99:153–162

[5] Zhou GP, An intriguing controversy over protein structural class prediction, *J Protein Chem*. 1998, 17(8):729–38

[6] Cai YD, Feng KY, Lu WC, Chou KC, Using LogitBoost classifier to predict protein structural classes. *J Theor Biol*. 2006, 238(1):172–6

[7] Bu W-S, Feng Z-P, Zhang Z, Zhang C-T, Prediction of protein (domain) structural classes based on amino-acid index, *European J of Biochemistry* 1999, 266:1043–49

[8] Jin L, Fang W, Tang H, Prediction of protein structural classes by a new measure of information discrepancy, *Computational Biology and Chemistry*, 2003, 27:373–80

[9] Sun X-D, Huang RB, Prediction of protein structural classes using support vector machines, *Amino Acids* 2006, 30:469–75

[10] Xiao X, Shao S, Huang Z, Chou KC, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, *J Comput Chem*. 2006, 27(4):478–82

[11] Zhang CT, Chou KC, An optimization approach to predicting protein structural class from amino-acid composition, *Protein Science* 1992, 1:401–8

[12] Chou KC, Zhang CT, A new approach to predicting protein folding types, *J of Protein Chem.* 1993, 12:169–78

[13] Shen HB, Yang J, Liu X-J, Chou KC, Using supervised fuzzy clustering to predict protein structural classes, *Biochem Biophys Res Commun*. 2005, 334:577–81

[14] Cai Y, Zhou G, Prediction of protein structural classes by neural network, *Biochimie*. 2000, 82(8):783–5

[15] Z-X. Wang, Z. Yuan, How good is the prediction of protein structural class by the component-coupled method?, *Proteins* 2000, 38:165–75

[16] Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K, Prediction of protein structural class with Rough Sets, *BMC Bioinfor.* 2006, 7:20

[17] Cai YD, Liu XJ, Xu X, Zhou GP, Support vector machines for predicting protein structural class, *BMC Bioinfor.* 2001, 2:3

[18] Kedarisetti KD, Kurgan L, Dick S. Classifier ensembles for protein structural class prediction with varying homology. *Biochem Biophys Res Commun*. 2006, 348(3):981–8

[19] Dong L, Yuan Y, Cai Y, Using Bagging classifier to predict protein domain structural class, *J Biomol Struct Dyn*. 2006, 24(3):239–42

[20] Feng KY, Cai YD, Chou KC, Boosting classifier for predicting protein domain structural class, *Biochem Biophys Res Commun*. 2005, 334(1):213–7

[21] Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*. 1997, 17:3389–402

[22] Jones DT, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol.* 1999, 292:195–202

[23] Kim H, Park H, Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor, *Proteins* 2004, 54(3):557–62

[24] Chen K, Kurgan L, Rahbari M, Prediction of protein crystallization using collocation of amino acid pairs, *Biochem Biophys Res Commun*. 2007, 355(3):764–9

[25] Chen K, Kurgan L, Ruan J, Optimization of the Sliding Window Size for Protein Structure Prediction, *IEEE Symp on Comp Intelligence in Bioinformatics and Comp Biology,* 2006, 366–372

[26] Witten I, Frank E, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005

[27] Keerthi SS, Shevade SK, Bhattacharyya C and Murphy KRK, Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Computation* 2001, 13:637–49

[28] Le CS and Houwelingen JC, Ridge Estimators in Logistic Regression. *Applied Statistics* 1992, 41:191–201

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

[30] John GH, Langley P, Estimating Continuous Distributions in Bayesian Classifiers. *11th Conf on Uncertainty in Artificial Intelligence*, 1995, 338–345

[31] Ross Q, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993