Paper:

Protein Structural Class Prediction via *k*-Separated Bigrams Using Position Specific Scoring Matrix

Harsh Saini*, Gaurav Raicar*, Alok Sharma*,, Sunil Lal*, Abdollah Dehzangi**, Rajeshkannan Ananthanarayanan*, James Lyons**, Neela Biswas***, and Kuldip K. Paliwal****

*The University of the South Pacific, Fiji
Laucala Bay, Suva, Fiji
**Griffith University
Brisbane, Australia
***Royal Brisbane and Women's Hospital
Brisbane, Australia
E-mail: {saini_h, raicar_g}@usp.ac.fj

[Received December 20, 2013; accepted February 15, 2014]

Protein structural class prediction (SCP) is as important task in identifying protein tertiary structure and protein functions. In this study, we propose a feature extraction technique to predict secondary structures. The technique utilizes bigram (of adjacent and *k*-separated amino acids) information derived from Position Specific Scoring Matrix (PSSM). The technique has shown promising results when evaluated on benchmarked Ding and Dubchak dataset.

Keywords: SCOP, structural class prediction, PSSM, bigram, *k*-separated bigram, transition probabilities, SVM

1. Introduction

Determining the structure of a protein plays a very important role in fields like molecular biology, cell biology and medical science. The function of a protein is closely linked to its structure, thus making it easy to analyze the heterogeneity of proteins, protein-protein interactions and development of new drug designs [1]. Traditional techniques such as X-ray crystallography and Nuclear Magnetic Resonance are good experimental tools but require expensive equipment and are time consuming processes [2, 3]. To cope with the plethora of protein sequences generated by various scientific communities, computational methods are being explored to determine the structure of protein accurately and efficiently.

The concept of protein structural classes was first reported by Levitt and Chothia [4] after performing a visual inspection on polypeptide chain topologies in a dataset of 31 globular proteins. Out of the many classes explored, the biological community typically follows four primary structural classes labelled as all-α, all-β, α/β and $\alpha+\beta$; where all-α classes and all-β represent structures that primarily consist of α-helices and β-strands respectively, the α/β class includes a mixture of proteins with both α-helices and β-strands and the $\alpha+\beta$ class contains proteins with α-helices and antiparallel β-sheets. Accurate classifications of protein structural classes are found in the Structural Classification of Proteins (SCOP) database [5]. Predicting protein structural class is normally a two-step problem in which the amino acid sequence is first transformed to a fixed-length feature vector. The feature vector which is usually a good representation of the primary sequence is then given to a classifier and the prediction is noted.

In literature, various features have been extracted which can be broadly classified into three groups – sequential, physicochemical and evolutionary based features. Early prediction algorithms, mostly made use of composition based features [6] which shows poor performance on datasets with low sequence similarity also known as the twilight zone [7, 8]. To solve this problem, physicochemical based features were introduced which are based on physicochemical attributes, e.g. Hydrophobicity and Polarity [9–13]. Recently, evolutionary based features have become popular and are achieving good results [14, 15]. Evolutionary features are extracted from Position Specific Scoring Matrix (PSSM) and are basically a representation of a protein sequence which defines the probability of amino acids occurring at a particular position in the sequence. Representing a protein sequence by its PSSM solves problems such as having zero value components in bigram feature vector by Sharma et al. [16].

In previous studies, various classification techniques have been explored and employed such as Linear Discriminant Analysis [17], K-Nearest Neighbor [18], Bayesian Classifier [19], Support Vector Machine (SVM) [20, 21], Artificial Neural Networks (ANN) [22, 23] and Ensemble classifiers [10, 24]. Out of the previously mentioned classification techniques, SVM-based classifiers has shown promising results [25]. However, it is shown in literature that to further improve the protein structural class prediction accuracy, a good combination

Table 1. Summary of datasets used in this paper.

Class	DD-dataset		EDD-dataset
	Train Samples	*Test Samples*	*Samples*
α	54	61	556
β	109	117	967
α/β	115	143	1311
$\alpha+\beta$	33	62	584

of feature extraction technique as well as classification technique is needed for optimal structural class prediction as shown in [25, 26].

In this paper, relationships amongst amino acids pairs in a protein that need not necessarily be adjacent in the primary sequence are explored using amino acid occurrence probabilities present in PSSM. The proposed technique, called *k*-separated bigrams, models information directly from PSSM whereby it calculates the bigram transition probabilities between amino acids, however, these transitions need not be between adjacent amino acid positions. The *k*-separated bigrams are evaluated later on in this paper and their influence on classification is discussed.

2. Dataset

In this research, the benchmarked Ding and Dubchak (DD) protein sequence dataset is used [11]. The dataset consists of a training set for the creation of the model and an independent test set for testing queries against the model. The data set belong to the four major structural classes, α, β, $\alpha+\beta$, and α/β. The training set consists of 311 protein sequences where any given pair of sequences does not have more than 35% sequence identity for aligned subsequences longer than 80 residues and the test set consists of 383 protein sequences where the sequence identity between any two given proteins is less than 40%.

Additionally, the extended Ding and Dubchak (EDD) dataset was also used to evalute the performance of the proposed technique. This dataset is generated by populating the DD-dataset with additional protein samples, which contains 3418 samples that belong to one of the four structural classes [27]. It does not contain a separate train and test set, therefore, the dataset was split into set of 60% for training and 40% for test. A summary of the datasets is provided in **Table 1**.

3. Procedure

In this paper, PSSM for each protein sequence has been extracted from the DD and EDD datasets using PSI-BLAST tool with NCBI's non-redundant (NR) database [8, 15]. Upon extracting PSSM, the proposed feature extraction technique was applied to create the feature vector that was used for training a support vector machine (SVM) classifier. A summary of the method is illustrated in **Fig. 1** and it has been discussed in more detail in the later sections.

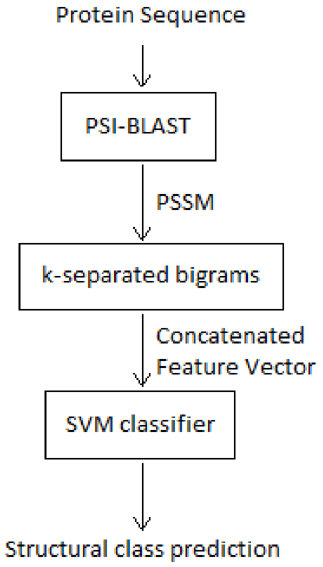


Fig. 1. A flowchart depicting classification using *k*-separated bigrams.

3.1. Feature Extraction Technique

We propose a technique called *k*-separated bigrams that models bigram transition probabilities between amino acid pairs that are not necessarily adjacent in the primary sequence using information present in PSSM. The value of *k* in *k*-separated bigrams determines the degree of separation (spatially) between amino acid pairs in the protein sequence. For instance, $k=1$ corresponds to amino acid pairs that are adjacent to each other whereas $k=2$ corresponds to amino acid pairs that are separated by 1 amino acid in between. Similarly, $k=3,4,\ldots,K$ corresponds to amino acid pairs that are separated by $2,3,\ldots,K-1$ amino acids respectively. Bigram transition probabilities are mathematically calculated based on the formula described in Eq. (1).

$$\lambda_{m,n}(k)=\sum_{i=1}^{L-k} P_{i,m}P_{i+k,n} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (1)$$

where $1\leq m\leq 20,\ 1\leq n\leq 20,\ 1\leq k\leq K$.

In Eq. (1) stated above, *P* represents PSSM matrix for a given protein that has *L* rows (where *L* is the number of amino acids in the protein sequence) and 20 columns (since there are only 20 unique amino acids). *k* represents the transition distance ranging from 1 to *K*, whose optimal value has to be discovered experimentally. For every value of *k*, a different feature vector of 400 dimensions is generated for the same protein sequence. These feature vectors include the bigram transition probabilities for amino acid pairs with separation of *k*, which are represented by $\lambda(k)$ in the equation. Upon applying the equation on a given protein sequence, $\lambda(k)$ for $k=1,2,\ldots,K$

Table 2. A listing of training accuracy via cross validation for various values of k on the DD and EDD datasets.

k	DD-dataset	EDD-dataset
1	83.9%	88.4%
2	83.0%	88.7%
3	83.0%	87.4%
4	82.3%	86.9%
5	83.6%	86.9%
6	83.6%	86.3%
7	79.7%	85.7%
8	78.8%	85.3%
9	–	84.2%

are concatenated to form a feature vector λ of $400 \times K$ dimensions.

3.2. Optimal Value Determination for K

The primary criteria used to determine the optimal value for K during the training phase was cross validation accuracy for $\lambda(k)$ with $k = 1,2,3,\ldots$. If a significant drop in accuracy was noticed, further evaluations of k are stopped. Other factors to take into consideration is the dimensionality of concatenated feature vector λ. As per general rule, as the dimensionality of λ increases, the computational requirements and degree of difficulty in classification also increases.

For the DD-dataset, the optimal range for k-separated bigrams was experimentally determined to be $k = 1,2,\ldots,6$ with $K = 6$. For each $\lambda(k)$, 10 fold cross validation with SVM classifier was applied upon the training set to determine the cross validation accuracy. Evaluations of $\lambda(k)$ continued till $k = 8$ upon which it was determined that the optimal value for K is 6. As per the experimentation results shown in **Table 2**, a significant drop in training accuracy is noticed with $k = 7$ and $k = 8$ and it was determined to use a threshold of 80.0% in the selection of k-separated bigrams for concatenation later to form λ. Therefore, further evaluations of $\lambda(k)$ were terminated and $k = 1,2,\ldots,6$ was determined to be the optimal range for applying k-separated bigrams.

Similar steps were taken to deduce the optimal value of K for the EDD-dataset. Through the experimentation on EDD-dataset (**Table 2**), it was determined to select all k-separated bigrams that produced more than 85.0% accuracy during training. Therefore, $K = 8$ was selected and the optimal range for k-separated bigrams was determined to be in $k = 1,2,\ldots,8$.

4. Results and Discussion

As previously stated, the proposed technique was evaluated using the benchmarked DD and EDD datasets. Initially, the classification model was created using the training set and it was evaluated using the test set. **Table 2** lists the cross validation accuracies on two training sets for different values for k-separated bigrams are quite high, which indicates that, relatively, there is significant amount of information present in the data that has been extracted.

Table 3. Sensitivity and specificity analysis of the proposed technique.

Class	DD		EDD	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*
α	92.9%	96.8%	94.5%	99.0%
β	92.9%	92.5%	96.7%	98.1%
$\alpha+\beta$	93.4%	93.6%	98.5%	98.0%
α/β	95.5%	97.5%	98.0%	97.5%

Table 4. Comparison of classification accuracy of the proposed technique with other reported works on DD and EDD datasets.

Method	DD *Test Accuracy*	EDD *Test Accuracy*
AAC+HXPZV	66.8%	68.9%
PF1	66.3%	68.1%
PF2	70.2%	68.9%
PSSM Bigram	79.6%	88.7%
***k*-separated bigrams**	**81.5%**	**93.7%**

However, according to the approach being described in this paper, the different $\lambda(k)$ were concatenated to form a single feature vector λ. We performed sensitivity and specificity analysis to evaluate the performance of the proposed technique relative to each structural class. The values for sensitivity and specificity were high, indicating a well balanced classification. The results are for the DD and EDD datasets are summarized in **Table 3**.

For purposes of comparison and evaluation against some techniques that have been used for structural class prediction, k-separated bigrams are compared with Amino Acid Composition (AAC) with Hydrophobicity (H), predicted secondary structure based on normalized frequency of α-helix (X), polarity (P), polarizability (Z), van der Waals volume (V) [11], pair-wise frequency type 1 and type 2 (PF1 and PF2) [28], and PSSM bigrams [16]. The results are summarized in **Table 4**. k-separated bigrams produced good results on both the benchmarked datasets achieving classification accuracies of 81.5% on the DD-dataset and 93.7% on the EDD-dataset.

In addition, the training and test sets were combined and the proposed technique was evaluated using cross validation paradigm for folds $n = 5,6,\ldots,10$. As per the results, it can be seen that the proposed technique has performed significantly better than other techniques that employ methods to predict structural classes for proteins. The highest accuracy for DD-dataset was recorded as 86.0% while 95.2% was most noteworthy classification for EDD-dataset. The results are highlighted in **Table 5**.

Table 5. Comparison of classification accuracy of the proposed technique with other reported works on DD and EDD datasets.

Dataset	Method	Accuracy for *n* Folds					
		n=5	*n*=6	*n*=7	*n*=8	*n*=9	*n*=10
DD	AAC+HXPZV	66.7%	67.0%	66.0%	65.1%	68.0%	67.6%
	PF1	66.7%	67.4%	67.3%	67.4%	67.7%	67.6%
	PF2	69.0%	69.3%	69.6%	69.6%	69.6%	69.6%
	PSSM Bigram	81.5%	81.7%	81.8%	81.8%	82.0%	82.0%
	***k*-separated bigrams**	**85.8%**	**85.7%**	**85.9%**	**85.9%**	**86.0%**	**86.0%**
EDD	AAC+HXPZV	41.5%	42.3%	42.8%	42.1%	42.7%	42.4%
	PF1	49.0%	48.8%	49.2%	49.2%	49.1%	49.3%
	PF2	49.1%	49.3%	48.4%	49.1%	49.5%	49.1%
	PSSM Bigram	89.0%	89.2%	89.4%	89.5%	89.6%	89.7%
	***k*-separated bigrams**	**94.7%**	**95.0%**	**95.0%**	**95.1%**	**95.1%**	**95.2%**

5. Conclusion

In this paper, a feature extraction technique has been proposed, which is based on *k*-separated bigrams using PSSM. The technique extracts transition probabilities between amino acid pairs with varying degree of separation to create the features. It was clearly shown that there is substantial information present in adjacent and non-adjacent amino acid transitions, which can be extracted using *k*-separated bigrams to help improve classification in protein structural class prediction. The proposed technique gave promising results and the accuracies noted via cross validation on the combined DD and EDD datasets were 86.0% and 95.2% respectively.

References:

[1] W. Chmielnicki, "A hybrid discriminative/generative approach to protein fold recognition," Neurocomputing, Vol.75, No.1, pp. 194-198, 2012.

[2] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. on Neural Networks, Vol.13, No.2, pp. 415-425, 2002.

[3] H.-B. Shen and K.-C. Chou, "Ensemble classifier for protein fold pattern recognition," Bioinformatics, Vol.22, No.14, pp. 1717-1722, 2006.

[4] M. Levitt and C. Chothia, "Structural patterns in globular proteins," Nature, Vol.261, No.5561, pp. 552-558, 1976.

[5] A. G. Murzin, S. E. Brenner, T. Hubbar, and C. Chothia, "Scop: A structural classification of proteins database for the investigation of sequences and structures," J. of Molecular Biology, Vol.247, No.4, pp. 536.540, 1995.

[6] G.-P. Zhou, "An intriguing controversy over protein structural class prediction," J. of Protein Chemistry, Vol.17, No.8, pp. 729-738, 1998.

[7] L. Kurgan and L. Homaeian, "Prediction of secondary protein structure content from primary sequence alone.a feature selection based approach," Machine Learning and Data Mining in Pattern Recognition, pp. 334-345, Springer, 2005.

[8] M. Mizianty and L. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," BMC Bioinformatics, Vol.10, No.1, p. 414, 2009.

[9] A. Dehzangi and S. Karamizadeh, "Solving protein fold prediction problem using fusion of heterogeneous classifiers," Information, Vol.14, No.11, pp. 3611-3621, 2011.

[10] A. Dehzangi, K. Paliwal, A. Sharma, O. Dehzangi, and A. Sattar, "A Combination of Feature Extraction Methods with an Ensemble of Different Classifiers for Protein Structural Class Prediction Problem," IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2013.

[11] C. Dubchak and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," Bioinformatics, Vol.17, No.4, pp. 349-358, 2001.

[12] A. Sharma, K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, and S. Miyano, "A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition," BMC bioinformatics, Vol.14, No.1, pp. 233, 2013.

[13] H. Zhang, T. Zhang, J. Gao, J. Ruan, S. Shen, and L. Kurgan, "Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility," Amino acids, Vol.42, No.1, pp. 271-283, 2012.

[14] T. Liu and C. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information," J. of Theoretical Biology, Vol.267, No.3, pp. 272-275, 2010.

[15] T. Liu, X. Geng, X. Zheng, R. Li, and J.Wang, "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles," Amino acids, Vol.42, No.6, pp. 2243-2249, 2012.

[16] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," J. of Theoretical Biology, 2012.

[17] P. Klein, "Prediction of protein structural class by discriminant analysis," Biochimica et Biophysica Acta (BBA) – Protein Structure and Molecular Enzymology, Vol.874, No.2, pp. 205-215, 1986.

[18] Y.-S. Ding and T.-L. Zhang, "Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier," Pattern Recognition Letters, Vol.29, No.13, pp. 1887-1892, 2008.

[19] A. Chinnasamy, W.-K. Sung, and A. Mittal, "Protein structure and fold prediction using tree-augmented naive bayesian classifier," J. of Bioinformatics and Computational Biology, Vol.3, No.4, pp. 803-819, 2005.

[20] A. Anand, G. Pugalenthi, and P. N. Suganthan, "Predicting protein structural class by SVM with class-wise optimized features and decision probabilities," J. of Theoretical Biology, Vol.253, No.2, pp. 375-380, 2008.

[21] Y.-D. Cai, X.-J. Liu, X.-b. Xu, and K.-C. Chou, "Prediction of protein structural classes by support vector machines," Computers & chemistry, Vol.26, No.3, pp. 293-296, 2002.

[22] Y.-D. Cai and G.-P. Zhou, "Prediction of protein structural classes by neural network," Biochimie, Vol.82, No.8, pp. 783-785, 2000.

[23] S. Jahandideh, P. Abdolmaleki, M. Jahandideh, and E. B. Asadabadi, "Novel two-stage hybrid neural discriminant model for predicting proteins structural classes," Biophysical Chemistry, Vol.128, No.1, pp. 87-93, 2007.

[24] K. Chen, L. A. Kurgan, and J. Ruan, "Prediction of protein structural class using novel evolutionary collocation based sequence representation," J. of Computational Chemistry, Vol.29, No.10, pp. 1596-1604, 2008.

[25] L. A. Kurgan, T. Zhang, H. Zhang, S. Shen, and J. Ruan, "Secondary structure-based assignment of the protein structural classes," Amino Acids, Vol.35, No.3, pp. 551-564, 2008.

[26] L. Kurgan and K. Chen, "Prediction of protein structural class for the twilight zone sequences," Biochemical and Biophysical Research Communications, Vol.357, No.2, pp. 453-460, 2007.

[27] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," Bioinformatics, Vol.25, No.20, pp. 2655-2662, 2009.

[28] P. Ghanty and N. R. Pal, "Prediction of protein folds: Extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers," IEEE Trans. on Nano Bioscience, Vol.8, No.1, pp. 100-110, 2009.

Name:
Harsh Saini

Affiliation:
School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Fiji

Address:
Laucala Bay, Suva, Fiji

Brief Biographical History:
2011 Received the B.S. degree, The University of the South Pacific
2014 Received the M.S. degree, The University of the South Pacific

Name:
Gaurav Raicar

Affiliation:
School of Computing Information and Mathematical Sciences, The University of the South Pacific, Fiji

Address:
Laucala Bay, Suva, Fiji

Brief Biographical History:
2011 Received the B.S. degree, The University of the South Pacific
2014 Received the M.S. degree, The University of the South Pacific

Name:
Alok Sharma

Affiliation:
Associate Professor, School of Engineering & Physics, The University of the South Pacific, Fiji
Adjunt Associate Professor, Griffith University

Address:
Laucala Bay, Suva, Fiji

Brief Biographical History:
2001 Received the M.Eng. degree, Griffith University
2006 Received the Ph.D. degree, Griffith University
2010-12- Research Fellow, the University of Tokyo
2013- Adjunct Associate Professor at the Institute for Integrated and Intelligent Systems (IIIS), Griffith University
2014- Associate Professor at the University of the South Pacific

Main Works:
- Various projects carried out in conjunction with Motorola (Sydney), Auslog Pty., Ltd. (Brisbane), CRC Micro Technology (Brisbane), the French Embassy (Suva), and JSPS (Japan). Research interests include pattern recognition, computer security, human cancer classification and protein fold and structural class prediction problems.

Membership in Academic Societies:
- The Institute of Electrical and Electronics Engineers (IEEE)

Name:
Sunil Lal

Affiliation:
Senior Lecturer, The University of the South Pacific, Fiji

Address:
Laucala Bay, Suva, Fiji

Brief Biographical History:
2001 Received the B.Sc. degree, University of the South Pacific, Fiji
2005 Received the M.Sc. degree, University of the South Pacific, Fiji
2009 Received the Ph.D. degree, University of the Ryukyus, Japan

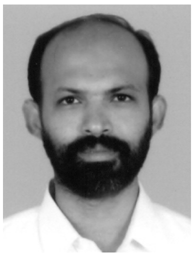
Membership in Academic Societies:
- The Institute of Electrical and Electronics Engineers (IEEE)
- Swedish AI Society
- South Pacific Computer Society

Name:
Abdollah Dehzangi

Affiliation:
Ph.D. Candidate at Griffith University
Researcher, National ICT Australia Ltd.

Address:
Brisbane, Australia

Brief Biographical History:
2007 Received the B.Sc. degree in Computer Engineering-Hardware from Shiraz University, Iran
2011 Received the M.Sc. degree in the area of bioinformatics from Multi Media University (MMU), Cyberjaya, Malaysia
2011- Ph.D. Candidate at Griffith University



Name:
Rajeshkannan Ananthanarayanan

Affiliation:
Senior Lecturer, The University of the South Pacific, Fiji

Address:
Laucala Bay, Suva, Fiji

Brief Biographical History:
2005- Joined as Lecturer in Mechanical Engineering Department, Thiagarajar College of Engineering, Madurai, India
2008- Joined as Lecturer in Mechanical Engineering Department, SEP, The University of South Pacific, Fiji
2012- Senior Lecturer in Mechanical Engineering Department, The University of South Pacific, Fiji

Name:
James Lyons

Affiliation:
Ph.D. Candidate, Griffith University

Address:
Brisbane, Australia

Brief Biographical History:
2007- BIT at Griffith University

Name:
Neela Biswas

Affiliation:
Medical Researcher, Royal Brisbane and Women's Hospital

Address:
Brisbane, Australia

Brief Biographical History:
2008-2013 MBBS at James Cook University
2014- Resident Medical Officer, Royal Brisbane and Women's Hospital

Main Works:

- "Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping," J. of Theoretical Biology, Vol.354, pp. 137-145, May 2014.

Name:
Kuldip K. Paliwal

Affiliation:
Professor, Griffith University

Address:
Brisbane, Australia

Brief Biographical History:
1971 Received the M.Sc. degree from Aligarh Muslim University, India
1978 Received the Ph.D. degree from Bombay University, India

Membership in Academic Societies:

- Founding Member (1991-1995), IEEE Signal Processing Society's Neural Networks Technical Committee
- Founding Member (1999-2003), IEEE Signal Processing Society's Speech Processing Technical Committee
- Associate Editor (1994-1997, 2003-2004), IEEE Trans. on Speech and Audio Processing
- Editor-in-chief (2005-2011), Speech Communication Journal