AUTHOR QUERY FORM

ELSEVIER	**Journal: YJTBI** **Article Number: 7171**	**Please e-mail or fax your responses and any corrections to:** **E-mail: corrections.essd@elsevier.macipd.com** **Fax: +44 1392 285878**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof. Click on the Q link to go to the location in the proof.

Location in article	**Query / Remark: click on the Q link to go** **Please insert your reply or correction at the corresponding line in the proof**
Q1	Please confirm that given names and surnames have been identified correctly and are presented in the desired order.
Q2	The reference given here is cited in the text but is missing from the reference list (Ding and Dubchack, 2001), please make the list complete or remove the reference from the text.
Q3	Please check whether the designated corresponding author is correct, and amend if necessary.
Q4	Please check author e-mail address.
Q5	Please check the telephone number of the corresponding author, and correct if necessary.
Q6	The citation “Chmielnicki et al., 2012” has been changed to “Chmielnicki and Stapor, 2012” to match the author name/date in the reference list. Please check here and in subsequent occurrences, and correct if necessary.

Thank you for your assistance.

Please check this box if you have no corrections to make to the PDF file □

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition

Alok Sharma[a,b,*], James Lyons[c], Abdollah Dehzangi[d,e], Kuldip K. Paliwal[c]

[a] Laboratory of DNA Information Analysis, University of Tokyo, Japan
[b] School of Engineering and Physics, University of the South Pacific, Fiji
[c] School of Engineering, Griffith University, Australia
[d] Institute for Integrated and Intelligent System, Griffith University, Brisbane, Australia
[e] National ICT Australia (NICTA), Brisbane, Australia

HIGHLIGHTS

- ▸ Performance of protein fold recognition has been improved.
- ▸ A new feature extraction method has been proposed.
- ▸ An improvement of around 10% has been observed.

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form
4 December 2012
Accepted 5 December 2012

Keywords:
Protein fold recognition
Bi-gram features
Position specific scoring matrix (PSSM)
Protein sequence

ABSTRACT

Discovering a three dimensional structure of a protein is a challenging task in biological science. Classifying a protein into one of its folds is an intermediate step for deciphering the three dimensional protein structure. The protein fold recognition can be done by developing feature extraction techniques to accurately extract all the relevant information from a protein sequence and then by employing a suitable classifier to label an unknown protein. Several feature extraction techniques have been developed in the past but with limited recognition accuracy only. In this work, we have developed a feature extraction technique which is based on bi-grams computed directly from Position Specific Scoring Matrices and demonstrated its effectiveness on a benchmark dataset. The proposed technique exhibits an absolute improvement of around 10% compared with existing feature extraction techniques.

1. Introduction

The identification of a three dimensional structure of a protein from its primary structure (which is a sequence of amino acids) is considered to be an important and challenging task in biological science and bioinformatics. An abundance of protein sequences are available which inherit significant biological information. The development of computational methods to decipher protein structure would help in understanding protein heterogeneity, protein–protein interactions and protein–peptide interactions. This would further help in disease diagnosis and drug design.

Proteins with different similarities and lengths can belong to the same fold. Similarly, proteins in the same fold can have the same major secondary structure in the same arrangement and with the same topology whether or not they have a common evolutionary origin (Craven et al., 1995; Yang et al., 2011). The categorization of protein folds from a protein sequence is an intermediate step in the recognition of protein structure. A wide range of techniques have been developed for protein fold recognition which address the issue of either classifier development or feature extraction development. For the former case, several classifiers have been developed or used including linear discriminant analysis (Klein, 1986), Bayesian classifiers (Chinnasamy et al., 2005), Bayesian decision rule (Wang and Yuan, 2000), k-nearest neighbor (Shen and Chou, 2006; Ding and Zhang, 2008), Hidden Markov model (Bouchaffra and Tan, 2006; Deschavanne and Tuffery, 2009), artificial neural network (Chen et al., 2007; Ying et al., 2009), support vector machine (SVM) (Ding and Dubchak, 2001; Shamim et al., 2007; Ghanty and Pal, 2009) and ensemble classifiers (Shen and Chou, 2009; Dehzangi et al., 2009, 2010; Yang et al., 2011; Dehzangi, 2011; Dehzangi and Karamizadeh, 2011). Among these classifiers, SVM (or SVM-based for ensemble strategy) classifier exhibits quite promising

* Corresponding author at: Laboratory of DNA Information Analysis, University of Tokyo, Japan. Tel.: +81 3 5449 5615.
E-mail address: aloks@ims.u-tokyo.ac.jp (A. Sharma).

0022-5193/$ - see front matter © 2012 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.jtbi.2012.12.008

results (Liu et al., 2012; Kurgan et al., 2008; Ghanty and Pal, 2009). For the latter case, several feature extraction techniques have been developed. Dubchak et al. (1997) have proposed syntactical and physicochemical-based features for protein fold recognition. They used amino acids' composition (AAC) as syntactical-based features and 5 following attributes of amino acids for deriving physicochemical-based features namely, hydrophobicity (H), predicted secondary structure based on normalized frequency of α-helix (*X*), polarity (*P*), polarizability (*Z*) and van der Waals volume (*V*). They used three descriptors (composition, transition and distribution) to compute the features. The AAC features comprise of 20 features and physicochemical-based features comprise of 105 features (21 features for each of the attributes used). The features proposed by Dubchak et al. (1997) have been widely used in the field of protein fold recognition (Chinnasamy et al., 2005; Krishnaraj and Reddy, 2008; Valavanis et al., 2010; Ding and Dubchak, 2001; Dehzangi et al., 2009; Kecman and Yang, 2009; Kavousi et al., 2011; Dehzangi and Amnuaisuk, 2011; Chmielnicki and Stapor, 2012). Apart from the above mentioned 5 attributes used by Dubchak et al. (1997), features are also extracted by incorporating other attributes of the amino acids. Some of the other attributes used are: solvent accessibility (Zhang et al., 2010), flexibility (Najmanovich et al., 2000), bulkiness (Huang and Tian, 2006), first and second order entropy (Zhang et al., 2008), size of the side chain of the amino acids (Dehzangi and Amnuaisuk, 2011). Taguchi and Gromiha (2007) proposed features which are based on amino acids' occurrence; Shamim et al. (2007) have extracted features from the structural information of amino acid residues and amino acid residue pairs; Ghanty and Pal (2009) proposed pairwise frequencies of amino acids separated by one residue (PF1) and pairwise frequencies of adjacent amino acid residues (PF2). There are 400 features each in PF1 and PF2. These pairwise frequency features (PF) are used as in the augmented form in the study conducted by Yang et al. (2011), thereby, having 800 features. Thus, the feature vector of PF has 800 features. Chou (2001) proposed pseudo-amino acid composition (A) based features to effectively represent protein sequence. Shen and Chou (2006), Kurgan et al. (2008) and Liu et al. (2012) have shown autocorrelation features for protein sequence, and Dehzangi and Amnuaisuk (2011) derived features by considering more physicochemical properties.

Since the extracted features play a crucial role in deciphering protein structure, in this paper we focus on developing feature extraction techniques and evaluate their recognition performance using SVM classifier. In Table 1, we summarize recognition performance of various existing feature extraction techniques using SVM classifier on the benchmark Ding and Dubchak (DD) dataset (Ding and Dubchak, 2001). It can be observed from the table that so far the highest recognition performance by a feature extraction technique on SVM classifier is 62.8%.

Table 1
Recognition accuracy for various feature extraction techniques using SVM classifier on DD-dataset.

Feature set	Recognition accuracy (%)
AAC+HXPZV (Ding and Dubchack, 2001)	56.0
Shamim et al. (2007)	60.5
Ghanty and Pal (2009)	59.2
Chmielnicki and Stapor (2012)	62.8
AHVPZ (Yang et al., 2011)	44.7
AX (Yang et al., 2011)	40.3
AHXPZV (Yang et al., 2011)	49.4
PF (Yang et al., 2011)	60.8
AHVPZ+PF (Yang et al., 2011)	51.2
AX+PF (Yang et al., 2011)	49.4
AHXPZV+PF (Yang et al., 2011)	52.7

In the literature, Ghanty and Pal (2009) have used bi-gram features for protein fold recognition. They have computed these features by counting the bi-gram frequencies of occurrences from the amino acid sequence representing the primary structure of a given protein.[1] Since all the primary protein sequences are made of 20 amino acids, there will be 400 different combinations of amino acids giving 400 bi-gram features. However, the number of amino acids in a protein sequence is limited and the dimensionality of the bi-gram feature vector is comparatively large. Therefore, many components in the bi-gram feature vector become equal to zero. Thus, the use of primary sequence for computing the bi-gram frequencies for feature extraction is not an effective way of capturing the information. As a result classification performance is expected to be low. Furthermore, if a protein sequence is updated by using position specific scoring matrix (PSSM) (Altschul et al., 1997) to obtain the consensus sequence and the bi-gram features are extracted from the consensus sequence, this problem of having mostly zeros in a bi-gram feature vector would still remain.

Instead of representing the given protein by its original primary sequence or by its consensus sequence, we propose to represent it by its PSSM directly. We compute the bi-gram feature vector by counting the bi-gram frequencies of occurrences from PSSM. Since PSSM provides information about the probability of 20 amino acids at each location of the protein sequence, we avoid zero components in the resulting bi-gram feature vector. Therefore, our procedure would retrieve more information useful for the protein fold recognition.

We investigate our feature extraction procedure on the benchmark DD-dataset and show that it achieves protein fold recognition accuracy of 69.5% (using SVM as a classifier). The obtained result is approximately 7% better than the result of the state-of-the-art feature extraction technique. We also merge the training set and test set of DD-dataset to perform *k*-fold cross-validation and obtain recognition accuracy around 10% better than the recognition accuracy of the existing feature extraction techniques.

2. Support Vector Machine for the evaluation of feature extraction techniques

SVM is considered to be the state-of-the-art machine learning and pattern classification algorithm (Vapnik, 1995). It has been extensively applied in classification and regression tasks. SVM aims to find maximum margin hyperplane to minimize classification error. SVM model is closely related to the neural network. In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network (Delashmit and Manry, 2005). A function called the kernel *K* is used to project the data from input space to a new feature space, and if this projection is non-linear it allows non-linear decision boundaries (Bishop, 2006).

To find a decision boundary between two classes, SVM attempts to maximize the margin between the classes, and choose linear separations in a feature space. The classification of some known point in input space $\mathbf{x}_i$ is y_i which is defined to be either -1 or $+1$. If $\mathbf{x}'$ is a point in input space with unknown classification then

$$y' = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}') + b\right) \quad (1)$$

where y' is the predicted class of point $\mathbf{x}'$. The function $K()$ is the kernel; n is the number of support vectors; α_i are adjustable

[1] We will call this sequence in the remainder of the text as the primary protein sequence, primary sequence, original protein sequence or the protein sequence interchangeably.

Please cite this article as: Sharma, A., et al., A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. (2012), http://dx.doi.org/10.1016/j.jtbi.2012.12.008

weights and b is a bias. In this study, the complexity parameter (C) (WEKA http://www.cs.waikato.ac.nz/ml/weka/; Chang and Lin, 2011; Keerthi et al., 2001; Platt, 1998) is set to be 1000. We use libsvm for training and testing with the radial basis function (RBF) kernel (Chang and Lin, 2011). The RBF kernel function can be given by $K(\mathbf{z}_i,\mathbf{z}_j)=\exp\left(-g^*\|\mathbf{z}_i-\mathbf{z}_j\|\hat{2}\right)$, where g is gamma parameter which is set to be 0.0038. These g and C parameters are tunable parameters which are determined here by doing cross-validation on the training set such that the classification accuracy is optimized. The g and C parameters are same for all the results except otherwise stated.

3. Dataset

In this study, the benchmark DD protein sequence dataset (Ding and Dubchak, 2001) have been employed. The DD-dataset consists of 311 protein sequences in the training set where two proteins have no more than 35% of sequence identity for aligned subsequence longer than 80 residues. The test set consists of 383 protein sequences where sequence identity is less than 40%. Both the sets belong to 27 SCOP folds (Murzin et al., 1995; http://scop.mrc-lmb.cam.ac.uk/scop/) which represented all major structural classes: α, β, α/β, and $\alpha+\beta$ (Ding and Dubchak, 2001). The summary of DD-dataset has been given in Table 2.

4. Feature extraction technique for protein fold recognition

In this section, we present the proposed bi-gram feature extraction technique using PSSM linear probabilities. It has been mentioned in the Introduction section that the computed bi-gram feature vector from the original protein sequence or the consensus sequence is very sparse as not all the combinations of amino acids are found in a protein sequence. In this paper, we do not represent the given protein by its primary sequence or by its consensus sequence. Instead we represent it by its PSSM and compute the bi-gram features using the probability information contained in PSSM. Let P be the matrix representing PSSM of a given protein. The matrix P will have L rows and 20 columns (where L is the length of the primary sequence). Its element at ith-row and jth-column is denoted by $p_{i,j}$ which can be interpreted as the relative probability of jth amino acid at the ith location of the primary protein sequence (with $\sum_{j=1}^{20} p_{i,j}=1$, for $i=1,2,\ldots,L$). The frequency of occurrence of transition from mth amino acid to nth amino acid is computed as follows:

$$B_{m,n}=\sum_{i=1}^{L-1} p_{i,m}p_{i+1,n},\ \text{where } 1\le m\le 20 \text{ and } 1\le n\le 20 \tag{2}$$

This equation gives 400 frequencies of occurrences $B_{m,n}; m=1,\ldots,20; n=1,\ldots,20$ for 400 bi-gram transitions. We call the matrix $\boldsymbol{B}$ as the bi-gram occurrence matrix and its 400 elements define our bi-gram feature vector F; i.e.,

$$F=[B_{1,1},B_{1,2},\ldots,B_{1,20},B_{2,1},\ldots,B_{2,20},\ldots,B_{20,1},\ldots,B_{20,20}]^{\mathrm{T}} \tag{3}$$

where the superscript T indicates the transpose of the vector. These bi-gram features can also be written in the form of pseudo amino acid composition (Chou, 2011). To do this, let us write the bi-gram feature vector as

$$F=[\psi_1,\psi_2,\ldots,\psi_u,\ldots,\psi_\Omega]^{\mathrm{T}} \tag{4}$$

where $\Omega=mn=400$ is the dimensionality of the feature vector F. The components of feature vector F can be expressed as the pseudo amino acid features as follows:

$$\psi_u=\begin{cases} B_{1,u} & (1\le u\le 20)\\ B_{2,u-20} & (21\le u\le 40)\\ \cdots & \\ \cdots & \\ B_{20,u-380} & (381\le u\le 400)\end{cases} \tag{5}$$

Since in the computation of feature vector F all the information of PSSM probability have been used, intuitively F contains more

Table 2
Summary of DD-dataset.

Fold	Number of training vectors	Number of test vectors
α		
Globin-like	13	6
Cytochromec	7	9
DNA-binding 3-helical bundle	12	20
4-Helical up-and-down bundle	7	8
4-Helical cytokines	9	9
Alpha; EF-hand	6	9
β		
	30	44
Cupredoxins	9	12
Viral coat and capsid proteins	16	13
ConA-like lectins/glucanases	7	6
SH3-like barrel	8	8
OB-fold	13	19
Trefoil	8	4
Trypsin-like serine proteases	9	4
Lipocalins	9	7
α/β		
(TIM)-barrel	29	48
FAD (also NAD)-binding motif	11	12
Flavodoxin-like	11	13
NAD (P)-binding Rossmann-fold	13	27
P-loop containing nucleotide	10	12
Thioredoxin-like	9	8
Ribonuclease H-like motif	10	12
Hydrolases	11	7
Periplasmic binding protein-like	11	4
$\alpha+\beta$		
β -Grasp	7	8
Ferredoxin-like	13	27
Small inhibitors, toxins, lectins	13	27

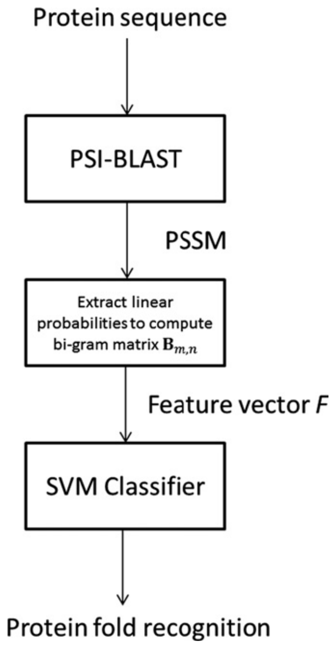

Fig. 1. A flow-diagram of protein sequence classification using bi-grams.

information useful for protein fold recognition task than computing bi-gram directly from the protein sequence (or from a consensus sequence). From biological perspective, proteins in the same fold often have amino-acid subsequences that are highly conserved. The bi-gram probabilities characterize the subsequence of amino acids in these conserved regions. If a certain subsequence is conserved in a fold, then each protein in that fold will have a group of bi-grams from that conserved region. This can help in discriminating folds that do not have the same amino acids subsequences. A flow-diagram showing the classification of a protein sequence into a protein fold has been depicted in Fig. 1.

In addition to bi-gram features, we can also compute mono-gram features from the probability information contained in PSSM. Since there are 20 amino acids, we will have 20 mono-gram features. These are computed as follows:

$$M_m = \sum_{i=1}^{L} p_{i,m}, \text{where } 1 \le m \le 20 \qquad (6)$$

Instead of computing mono-gram features from PSSM as done in Eq. (6), we can compute the frequency of occurrence of individual mono-grams (or amino acids) directly from the primary protein sequence itself. This procedure has been used in the past to compute the occurrence feature by Taguchi and Gromiha (2007).

5. An illustration of bi-gram feature computation using a toy problem

In order to illustrate the bi-gram feature extraction method, let us consider a toy example of a protein with primary sequence *RRARA* of length $L=5$. Note that we assume that the toy proteins are made of 3 amino acids *A*, *R* and *T*. Table 3 shows the PSSM of this protein.

Using the probability information in PSSM, we can find out the consensus sequence for this protein as *AARAR*. The bi-gram features computed from the original protein sequence *RRARA* and the consensus sequence *RRARA* are shown in Table 4.

Therefore, the bi-gram feature vectors of the original protein sequence and the consensus sequence will be {0,1,0,2,1,0,0,0,0} and {1,2,0,1,0,0,0,0}, respectively. From this we can see that both the primary sequence as well as the consensus sequence produce many zero components in the bi-gram feature vector. When we compute the bi-gram feature vector using Eq. (2) from PSSM (as given in Table 3), we obtain the bi-gram occurrence matrix as shown in Table 5. This gives the bi-gram feature vector as {0.4775, 0.6775, 0.445, 0.4125, 0.5, 0.3875, 0.31, 0.4725, 0.3175}.

Table 3
PSSM of the protein *RRARA*.

Amino acids	*A*	*R*	*T*
R	0.50	0.25	0.25
R	0.45	0.30	0.25
A	0.25	0.50	0.25
R	0.40	0.25	0.35
A	0.10	0.60	0.30

Table 4
Bi-gram feature vectors *F* computed from the original protein sequence and the consensus sequence. From column 2 to column 10 are the bi-gram frequencies.

F	*AA*	*AR*	*AT*	*RA*	*RR*	*RT*	*TA*	*TR*	*TT*
Frequency (original protein sequence)	0	1	0	2	1	0	0	0	0
Frequency (consensus sequence)	1	2	0	1	0	0	0	0	0

Table 5
Bi-gram occurrence matrix *B*.

0.4775	0.6775	0.4450
0.4125	0.5000	0.3875
0.3100	0.4725	0.3175

Thus, the bi-gram feature vector computed from PSSM using Eq. (2) does not show the sparsity as seen earlier when it was computed using the original primary sequence or the consensus sequence.

For completeness, we have also computed the mono-gram feature vector using Eq. (6) from PSSM (as given in Table 3). This gives the mono-gram feature vector as {1.7,1.9,1.4}.

6. Experimentation

We perform computational experiments on the benchmark DD-dataset to show the effectiveness of our proposed method. The DD-dataset has separated training set and test set (as shown in Table 2). We employ the SVM classifier from libsvm to find the accuracy of protein fold recognition where the accuracy is defined as the percentage of correctly recognized proteins to all the proteins of the test set. The SVM classifier is widely used in classification task. It finds maximum margin hyperplane to minimize classification error.

In statistical prediction, the following three procedures are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test. However, of the three test procedures, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in Chou and Shen (2010) and demonstrated by equations 28–30 in Chou (2011). Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors (see, e.g., Mohabatkar, 2010; Qiu and Wang, 2012; Sahu and Panda, 2010; Esmaeili et al., 2010; Hayat and Khan, 2011, 2012; Shi et al., 2012; Kandaswamy et al., 2011). However, to reduce the computational time, we adopted the independent dataset and *k*-fold cross-validation in this study as done by many investigators with SVM as the prediction engine. Thereby, the experiment has two parts. In the first part, we employ the training set to estimate the parameters of the classifier and use a separate test set to find the accuracy of protein fold recognition. The accuracies thereby obtained are compared with the accuracies reported in the literature. In the second part, we merge the training set and test set of DD-dataset together and perform *k*-fold cross-validation[2] on a number of feature extraction techniques.

For the first part, we employ Eqs. (2) and (6) to find the bi-gram and mono-gram feature vectors, respectively. Thereby, we applied the SVM classifier to compute the accuracy of protein fold recognition. The obtained accuracies are compared with other reported results and shown in Table 6.

It can be observed from Table 6 that the highest accuracy is obtained by bi-gram technique (69.5%) (as shown as the bold face in the table), however, mono-gram is also showing promising results.

In order to check the sparsity level of using PSSMs instead of the original protein sequence on the entire training set, we compute the average number of non-zero entries in the bi-gram feature vectors derived from the original protein sequence and

[2] For statistical stability we performed 100 times *k*-fold cross-validation in this paper.

Please cite this article as: Sharma, A., et al., A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. (2012), http://dx.doi.org/10.1016/j.jtbi.2012.12.008

Table 6
Recognition accuracy of the proposed bi-gram and mono-gram feature extraction techniques compared with various existing feature extraction techniques using SVM classifier on DD-dataset.

Feature set	Recognition accuracy (%)
ACC+HXPZV (Ding and Dubchack, 2001)	56.0
Shamim et al., (2007)	60.5
Ghanty and Pal (2009)	59.2
Chmielnicki and Stapor (2012)	62.8
AHVPZ (Yang et al., 2011)	44.7
AX (Yang et al., 2011)	40.3
AHXPZV (Yang et al., 2011)	49.4
PF (Yang et al., 2011)	60.8
AHVPZ+PF (Yang et al., 2011)	51.2
AX+PF (Yang et al., 2011)	49.4
AHXPZV+PF (Yang et al., 2011)	52.7
Mono-gram (this paper)	62.1
Bi-gram (this paper)	**69.5**

Table 7
Recognition accuracy by *k*-fold cross validation procedure for various feature extraction techniques using SVM classifier on DD-dataset.

Feature sets	$k=5$	$k=6$	$k=7$		$k=9$	$k=10$
PF1	48.6	49.1	49.5	50.1	50.5	50.6
PF2	46.3	47.0	47.5	47.7	47.9	48.2
PF[a]	51.2	52.2	52.6	52.9	53.4	53.4
O	49.7	50.4	50.8	50.8	51.1	51.0
AAC[b]	43.6	43.9	44.2	44.8	44.6	45.1
AAC+HXPZV[c]	45.1	46.2	46.5	46.8	46.9	47.2
PSSM+PF1	62.5	63.2	63.7	64.2	64.5	64.6
PSSM+PF2	62.7	63.3	64.1	64.2	64.6	64.7
PSSM+PF[a]	65.5	66.2	66.5	66.9	67.1	67.5
PSSM+O	62.5	62.1	62.5	62.9	63.4	63.5
PSSM+AAC[b]	57.5	58.1	58.4	58.7	59.1	59.2
PSSM+AAC+HXPZV[c]	55.9	56.9	57.1	57.7	58.0	58.2
Mono-gram (this paper)	67.7	68.4	68.6	69.1	69.4	69.6
Bi-gram (this paper)	**72.6**	**73.1**	**73.7**	**73.7**	**74.1**	**74.1**

[a] PF results use gamma=0.001.
[b] AAC results use gamma =11 QUOTE and $C=100$ QUOTE .
[c] AAC+HXPZV results use gamma=2 and $C=8$ QUOTE .

PSSMs, respectively. It was experimentally determined that 29% of non-zero entries are obtained for the bi-gram feature vectors extracted from the original protein sequences. On the other hand, 95% of non-zero entries are obtained for the bi-gram feature vectors extracted from PSSMs. This analysis indicates that about 70% of bi-gram feature vector is sparse when extracted from the original protein sequence and the sparsity level is reduced to only 5% when extracted from PSSMs. Therefore, it can be seen that bi-grams from PSSMs significantly reduce the sparsity level which help in improving the recognition performance.

Next, we have merged the training set and test set of DD-dataset to perform *k*-fold cross-validation procedure. The results are depicted in Table 7. The values of *k* are taken to be 5, 6, 7, 8, 9 and 10. For the classifier, SVM is used with RBF kernel. The RBF kernel parameters are gamma $=0.0038$ and $C=1000$. For some feature sets, these parameters are changed to obtain better recognition performance. The parameters that are changed are mentioned under Table 7. The following feature sets are considered for the experiment: PF1, PF2 (Ghanty and Pal, 2009), PF (Yang et al., 2011), Occurrence (O) (Taguchi and Gromiha, 2007), AAC and AAC+HXPZV (Ding and Dubchak, 2001). We have also updated the protein sequences to get the consensus sequence by using their corresponding PSSMs; i.e., each amino acid of a protein sequence is replaced by the amino acid that has the highest probability in PSSM. After this updating procedure, we have used the same feature extraction techniques (PF1, PF2, PF, O, AAC and AAC+HXPZV) again to obtain the recognition performance. In Table 7, we have placed the results for PSSM updated protein sequences (or the consensus sequence) in the columns 2–7 of the row of PSSM+*FEAT*, where *FEAT* is any feature extraction technique. The highest recognition accuracy of a particular *k*-fold cross-validation is mentioned in bold face.

It can be observed from Table 7 that when protein sequences are not updated by using PSSM then the PF feature shows better recognition accuracy than PF1, PF2, ACC, O and ACC+HXPZV for all $k=5$, 6, 7, 8, 9 and 10. Among the consensus sequences also, the feature set PF exhibits the best recognition performance. The mono-gram feature (of this paper) is showing better recognition performance than all the other existing feature extraction techniques with the highest accuracy of 69.6% when $k=10$. In particular, mono-gram is outperforming occurrence feature O for all the values of *k*. The bi-gram feature (of this paper) is showing the best recognition performance for all *k*. The highest accuracy obtained is 74.1% (at $k=9$ and $k=10$) which is around 10% better than the other feature extraction techniques in the literature. This is a significant improvement in terms of recognition accuracy when compared with existing feature extraction techniques.

Additionally, we have tried this algorithm to compute tri-gram features from PSSMs and observed that the recognition accuracy remains almost the same as that from the bi-gram features obtained by the present PSSM-based algorithm. Thus, the use of *n*-gram features (where $n>2$) is found not to be very useful here.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen, 2009), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

7. Conclusion

In this study, we have developed a feature extraction technique based on bi-gram. The proposed technique utilizes PSSM linear probabilities to compute features. The effectiveness of the technique was gauged against several existing feature extraction techniques on a benchmark dataset and very promising results have been obtained. It was observed that the proposed technique exhibits up to 10% improvement in recognition accuracy of protein fold.

We have also shown that instead of computing frequency of amino acid occurrence from primary protein sequence, we can compute mono-gram features from PSSM directly. This has also shown promising results.

Acknowledgment

We thank the Reviewers and the Editor for their constructive comments which appreciably improved the presentation quality of the paper.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 17, 3389–3402.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer Science, New York.

Bouchaffra, D., Tan, J., 2006. Protein fold recognition using a structural Hidden Markov model. In: Proceedings of the 18th International Conference on Pattern Recognition, pp. 186–189.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 27:1–27:27 ⟨http://www.csie.ntu.edu.tw/~cjlin/libsvm⟩.

Chen, K., Zhang, X., Yang, M.Q., Yang, J.Y., 2007. Ensemble of probabilistic neural networks for protein fold recognition. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE), pp. 66–70.

Chinnasamy, A., Sung, W.K., Mittal, A., 2005. Protein structure and fold prediction using tree-augmented naive Bayesian classifier. J. Bioinf. Comput. Biol. 3 (4), 803–819.

Chmielnicki, W., Stapor, K., 2012. A hybrid discriminative-generative approach to protein fold recognition. Neurocomputing 75, 194–198.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43, 246–255. (erratum: 2001, vol. 44, 60).

Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2, 63–92, http://dx.doi.org/10.4236/ns.2009.12011, openly accessible at⟨http://www.scirp.org/journal/NS/⟩.

Chou, K.C., Shen, H.B., 2010. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. 2, 1090–1103, http://dx.doi.org/10.4236/ns.2010.210136 Nature Protocols, 2008, 3, 153–162.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). J. Theor. Biol. 273, 236–247.

Craven, M.W., Mural, R.J., Hauser, L.J., Uberbacher, E.C., 1995. Predicting protein folding classes without overly relying on homology. ISMB 3, 98–106.

Dehzangi, A., Amnuaisuk, S.P., 2011. Fold prediction problem: the application of new physical and physicochemical-based features. Protein Pept. Lett. 18, 174–185.

Dehzangi, A., Amnuaisuk, S.P., Dehzangi, O., 2010. Enhancing protein fold prediction accuracy by using ensemble of different classifiers. Aust. J. Intell. Inf. Process. Syst. 26 (4), 32–40.

Dehzangi, A., Amnuaisuk, S.P., Ng, K.H., Mohandesi, E., 2009. Protein fold prediction problem using ensemble of classifiers. In: Proceedings of the 16th International Conference on Neural Information Processing, Part II, pp. 503–511.

Dehzangi, A., 2011. Karamizadeh, 2011. Solving protein fold prediction problem using fusion of heterogeneous classifiers. Information—an International Interdisciplinary Journal 14 (11), 3611–3622.

Delashmit, W.H., Manry, M.T., 2005. Recent developments in multilayer perceptron neural networks. In: Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC.

Deschavanne, P., Tuffery, P., 2009. Enhanced protein fold recognition using a structural alphabet. Proteins: Struct. Funct. Bioinf. 76, 129–137.

Ding, C., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17 (4), 349–358.

Ding, Y.S., Zhang, T.L., 2008. 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 Lett. 29, 1887–1892.

Dubchak, I., Muchnik, I., Kim, S.K., 1997. Protein folding class predictor for SCOP: approach based on global descriptors. In: Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, pp. 104–107.

Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.

Ghanty, P., Pal, N.R., 2009. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans. Nano Biosci. 8, 100–110.

Hayat, M., Khan, A., 2011. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J. Theor. Biol. 271, 10–17.

Hayat, M., Khan, A., 2012. MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM. J. Theor. Biol. 292, 93–102.

Huang, J.T., Tian, J., 2006. Amino acid sequence predicts folding rate for middle-size two-state proteins. Proteins: Struct. Funct. Bioinf. 63 (3), 551–554.

Kandaswamy, K.K., Chou, K.C., Martinetz, T., Moller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62.

Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N., Moosavi-Movahedi, A.A., 2011. A protein fold classier formed by fusing different modes of pseudo amino acid composition via PSSM. Comput. Biol. Chem. 35 (1), 1–9.

Kecman, V., Yang, T., 2009. Protein fold recognition with adaptive local hyper plane algorithm. Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 09 IEEE Symposium, 75–78.

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

Klein, P., 1986. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta 874, 205–215.

Krishnaraj, Y., Reddy, C.K., 2008. Boosting methods for protein fold recognition: an empirical comparison. IEEE International Conference on Bioinformatics and Biomedical Engineering, pp. 393–396.

Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J., 2008. Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551–564.

Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., 2012. Accurate prediction of protein structural class using autocovariance transformation of PSI-BLAST profiles. Amino Acids 42, 2243–2249.

Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept. Lett. 17, 1207–1214.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Najmanovich, R., Kuttner, J., Sobolev, V., Edelman, M., 2000. Side-chain flexibility in proteins upon ligand binding. Proteins: Struct., Funct. Bioinf. 39 (3), 261–268.

Platt, J., 1998. Fast training of support vector machines using sequential minimal optimization, In: Schoelkopf, B., Burges, Smola, A. (Eds.), Advances in Kernel Methods—Support Vector Learning.

Qiu, Z., Wang, X., 2012. Prediction of protein–protein interaction sites using patch based residue characterization. J. Theor. Biol. 293C, 143–150.

Sahu, S.S., Panda, G., 2010. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput. Biol. Chem. 34, 320–327.

Shamim, M.T.A., Anwaruddin, M., Nagarajaram, H.A., 2007. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 23 (24), 3320–3327.

Shen, H.B., Chou, K.C., 2006. Ensemble classier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.

Shen, H.B., Chou, K.C., 2009. Predicting protein fold pattern with functional domain sequential evolution information. J. Theor. Biol. 256, 441–446.

Shi, S.P., Qiu, J.D., Sun, X.Y., Suo, S.B., Huang, S.Y., Liang, R.P., 2012. A method to distinguish between lysine acetylation and lysine methylation from protein sequences. J. Theor. Biol. 310, 223–230.

Taguchi, Y.-h., Gromiha, M.M., 2007. Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinf. 8, 404.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Wang, Z.Z., Yuan, Z., 2000. How good is prediction of protein-structural class by the component-coupled method? Proteins 38, 165–175.

Yang, T., Kecman, V., Cao, L., Zhang, C., Huang, J.Z., 2011. Margin-based ensemble classifier for protein fold recognition. Expert Syst. Appl. 38, 12348–12355.

Ying, Y., Huang, K., Campbell, C., 2009. Enhanced protein fold recognition through a novel data integration approach. BMC Bioinf. 10 (1), 267.

Valavanis, I.K., Spyrou, G.M., Nikita, K.S., 2010. A comparative study of multi-classification methods for protein fold recognition, Int. J. Comput. Intell. Bioinf. Syst. Biol. 1 (3), 332–346.

Zhang, H., Zhang, T., Gao, J., Ruan, J., Shen, S., Kurgan, L.A., 2010. Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. Amino Acids, 1–13.

Zhang, T.L., Ding, Y.S., Chou, K.C., 2008. Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. J. Theor. Biol. 250, 186–193.

Please cite this article as: Sharma, A., et al., A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. (2012), http://dx.doi.org/10.1016/j.jtbi.2012.12.008