Dear Author,

Please, note that changes made to the HTML content will be added to the article before publication, but are not reflected in this PDF.

Note also that this file should not be used for submitting corrections.

Journal of Theoretical Biology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition

Harsh Saini [a,*,1], Gaurav Raicar [a,1], Alok Sharma [a,b], Sunil Lal [a], Abdollah Dehzangi [b], James Lyons [b], Kuldip K. Paliwal [b], Seiya Imoto [c], Satoru Miyano [c]

[a] University of the South Pacific, Fiji
[b] Griffith University, Brisbane, Australia
[c] Human Genome Center, University of Tokyo, Japan

HIGHLIGHTS

- Relationships between amino acid dimers that may be non-adjacent in sequence are explored.
- Features are extracted directly from PSSM instead of raw counts from primary sequence.
- SVM is used for classification.
- Achieved good results on Ding and Dubchak, Extended Ding and Dubchak, and Taguchi and Gromhia datasets.

ARTICLE INFO

Article history:
Received 1 March 2015
Received in revised form
28 April 2015
Accepted 21 May 2015

ABSTRACT

Background: Identification of the tertiary structure (3D structure) of a protein is a fundamental problem in biology which helps in identifying its functions. Predicting a protein's fold is considered to be an intermediate step for identifying the tertiary structure of a protein. Computational methods have been applied to determine a protein's fold by assembling information from its structural, physicochemical and/or evolutionary properties.
Methods: In this study, we propose a scheme in which a feature extraction technique that extracts probabilistic expressions of amino acid dimers, which have varying degree of spatial separation in the primary sequences of proteins, from the Position Specific Scoring Matrix (PSSM). SVM classifier is used to create a model from extracted features for fold recognition.
Results: The performance of the proposed scheme is evaluated against three benchmarked datasets, namely the Ding and Dubchak, Extended Ding and Dubchak, and Taguchi and Gromiha datasets.
Conclusions: The proposed scheme performed well in the experiments conducted, providing improvements over previously published results in literature.

1. Background

In the field of biological science, predicting the three-dimensional structure of a given protein is an important task since the structures relate closely to the biological function of the protein (Chmielnicki, 2012). This in turn, improves the understanding of the heterogeneity of proteins, proteinσprotein interactions and proteinσpeptide interactions and aids the development of drug designs. In drug design, knowing the tertiary structure of the target protein is crucial since drugs are created to bind with the active sites on the target protein (Dubchak et al., 1999; Sharma et al., 2013).

Although X-ray crystallography is a powerful tool in determining protein 3D structures, it is time-consuming and expensive. Particularly, not all proteins can be successfully crystallized. For example, membrane proteins are very difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane

* Corresponding author.
E-mail addresses: saini_h@usp.ac.fj (H. Saini), raicar_g@usp.ac.fj (G. Raicar), sharma_al@usp.ac.fj (A. Sharma), lal_s@usp.ac.fj (S. Lal), a.dehzangi@griffith.edu.au (A. Dehzangi), james.lyons@griffithuni.edu.au (J. Lyons), k.paliwal@griffith.edu.au (K.K. Paliwal), imoto@ims.u-tokyo.ac.jp (S. Imoto), miyano@hgc.jp (S. Miyano).
[1] These authors contributed equally to this work

http://dx.doi.org/10.1016/j.jtbi.2015.05.030
0022-5193/

protein structures have been determined. Although Nuclear Magnetic Resonance is indeed a very powerful tool in determining the 3D structures of membrane proteins as indicated by a series of recent publications, it is also time-consuming and costly (Berardi et al., 2011; Schnell and Chou, 2008; OuYang et al., 2013). Furthermore, these techniques are extremely time consuming, expensive and seemingly impractical for protein sequences of extremely large lengths (Hsu and Lin, 2002; Shen and Chou, 2006). To timely acquire useful information for developing novel drugs, one has to resort various bioinformatics tools and structural bioinformatics tools (Shen and Chou, 2006; Chou, 2015, 2004; Ding and Dubchak, 2001). In this regard, it would be certainly helpful to develop powerful methods for predicting protein structural classes and fold types.

Predicting the fold of a protein sequence is considered to be an intermediate step in identifying the tertiary structure of a protein. The primary structure (sequence of amino acids) of different protein sequences can vary in length and similarities; however, they can still belong to the same fold. The protein fold recognition problem can be defined as categorizing unknown protein sequences to its well defined folds. Majority of the folds are accurately represented and defined in the Structural Classification of Proteins (SCOP) database (Murzin et al., 1995).

Various techniques have been used for protein fold recognition which can be broadly grouped into feature extraction and classification development. For the former group, syntactical, physicochemical and evolutionary features have been used. Dubchak et al. (1997) suggested syntactical and physicochemical based features in which they used five attributes of amino acids namely – hydrophobicity (H), predicted secondary structure based on normalized frequency of α helix (X), polarity (P), polarizability (Z) and van der Waals volume (V). These features have in turn been widely adopted by other researchers in protein fold recognition (Ding and Dubchak, 2001; Dehzangi et al., 2013). Further to this, additional attributes have been used later on such as solvent accessibility, flexibility, and bulkiness (Dehzangi et al., 2013; Zhang et al., 2012; Najmanovich et al., 2000; Huang and Tian, 2006). Taguchi and Gromiha used syntactical based features (occurrence and composition) to do protein fold recognition (Taguchi and Gromiha, 2007). Ghanty and Pal (2009) used pairwise frequencies of amino acids and separated by one residue. These pairwise frequency features (PF) are concatenated in the study conducted by Yang et al. (2011), thereby, having 800 features. If in concatenation of features, the dimensionality is unmanageable then higher dimensionality of feature vector can be controlled by selecting a few important features (Sharma et al., 2006, 2011, 2012a, 2012b, 2012c, 2012d, 2013b, 2013c; Sharma and Paliwal, 2007, 2010, 2012a, 2012b, 2012c). Chou proposed pseudo-amino acid composition (PseAAC) based features to represent protein sequences and it has been applied successfully in a large variety of publications (Shen and Chou, 2006; Chou, 2001; Liu et al., 2012; Li et al., 2009; Sahu and Panda, 2010; Chen et al., 2012; Liao et al., 2012; Qin et al., 2012; Kong et al., 2014; Zhang et al., 2014; Shen and Chou, 2008; Cao et al., 2013; Du et al., 2012; Du et al., 2014; Chou, 2011). Recently, the use of evolutionary features for protein fold recognition is increasing as it is achieving good results (Sharma et al., 2013, 2014; Liu et al., 2012; Liu and Jia, 2010; Cai and Zhou, 2000; Dong et al., 2009; Paliwal et al., 2014; Dehzangi et al., 2013). 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.

Furthermore, over the years, many classification techniques have also been explored for protein fold recognition such as Linear Discriminant Analysis (Klein, 1986), K-Nearest Neighbor (Ding and Zhang, 2008), Bayesian Classifier (Chinnasamy et al., 2005), Support Vector Machine (SVM) (Sharma et al., 2013a, 2013c; Anand et al., 2008; Cai et al., 2002; Saini et al., 2014a, 2014b), Artificial Neural Networks (ANN) (Cai and Zhou, 2000; Jahandideh et al., 2007a, 2007b) and Ensemble classifiers (Shen and Chou, 2006; Dehzangi et al., 2013; Chen et al., 2008; Kedarisetti et al., 2006). Out of the previously mentioned classification techniques, SVM-based classifiers showed promising results (Kurgan et al., 2008). However, it is shown in literature that to further improve the protein folding accuracy, a good combination of features extraction technique as well as classification technique is needed (Kurgan et al., 2008; Kurgan and Chen, 2007).

As shown by a series of recent publications (Chen et al., 2014a, 2014b; Ding et al., 2014; Lin et al., 2014; Liu et al., 2014; Xu et al., 2014; Liu et al., 2015) in response to the suggestion (Dubchak et al., 1997), to propose a sequence-based statistical predictor for a biological system, the following guidelines are followed: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; and (d) properly perform validation tests to objectively evaluate the accuracy of the predictor.

In this paper, we propose a scheme to predict protein folds using probabilistic expressions of amino acid dimer occurrence that have varying degrees of spatial separation in the protein sequence. The primary reason for using a feature extractor, which is explained later in the paper, is to explore relationships amongst amino acids dimers in a protein that may be non-adjacent in the primary sequence. Additionally, these relationships modeled using evolutionary information present in PSSM to improve classifier performance. The extracted information for amino acid dimers is used for protein fold recognition using SVM on several datasets.

2. Methods

In a nutshell, the proposed scheme incorporates an extraction technique that computes features directly from the evolutionary information present in PSSM. Initially, PSSM is extracted from the protein sequences using PSI-BLAST. This is succeeded by calculating the probabilistic expressions of amino acid dimers to extract the feature sets $F(k)$ for varying degrees of spatial separation from $k=1, \ldots, K$. The optimal value of K is determined empirically and the various feature sets are concatenated to form F, which is processed

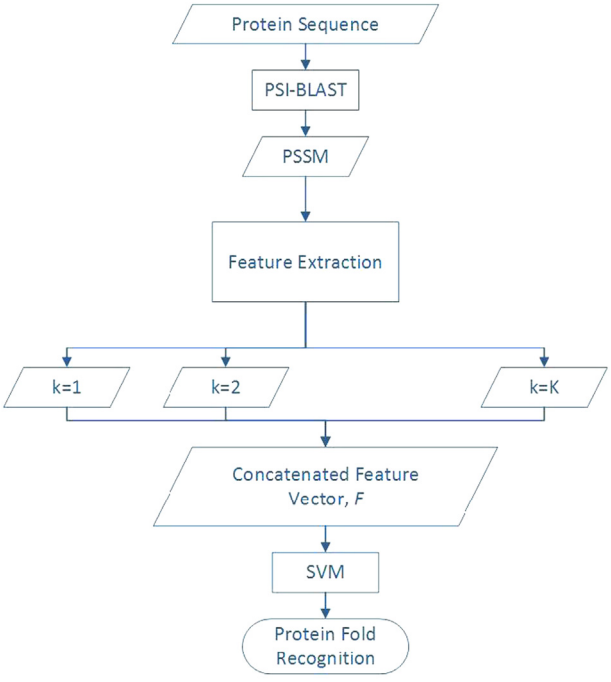

Fig. 1. Flow diagram of the proposed classification procedure.

Please cite this article as: Saini, H., et al., Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.05.030

by a SVM classifier for protein fold recognition. A flow-diagram of this scheme is illustrated in Fig. 1.

The proposed scheme can be summarized in an algorithm as shown in Algorithm 1. The algorithm highlights the feature extraction technique, the various operations used to modify the feature vector extracted and the method for determining the optimal value for the termination value K for the maximum degree of spatial separation when computing amino acid dimers. The equations and symbols used in the algorithm are described in the latter sections. It should be noted that the training set was used to determine K.

Algorithm 1. The training phase of the proposed scheme summarized in an algorithm

Step 1 Let $k:=1$

Step 2
$$F_{m,n}(k)=\sum_{i=1}^{L-k} P_{i,m}P_{i+k,n} \quad (1)$$

Step 3
$$F(k)=[\,F_{1,1}(k),F_{1,2}(k),\ \ldots,\ F_{1,20}(k),F_{2,1}(k),\ \ldots,F_{2,20}(k),\ \ldots,F_{20,1}(k),\ \ldots,F_{20,20}(k)\,] \quad (2)$$

Step 4
$$F\leftarrow\{\,F(k-1),F(k)\} \quad (3)$$
where $F(0)\leftarrow\varnothing$

Step 5 $A_k:=$ *classify*(F)[1] where *classify*(F) returns the cross-validation accuracy for feature vectors, F

Step 6
IF $A_k<A_{k-1}$ THEN
    GOTO Step 7
ELSE
    $k:=k+1$
    GOTO Step 2
ENDIF

Step 7 $K:=k$

[1]The metric used to evaluate the performance of various feature vectors was the n-fold cross-validation accuracy with $n=10$. SVM was used for classification with the kernel function as *radial basis function* and the C-parameter being 1000. The features were computed from the training set whereby the training set for DD-dataset was used, however, for EDD and TG datasets the samples were randomly divided into training and test sets in the ratio 3:2 and ensuring equal distribution of various classes.

The technique proposed in this study attempts to model relationships between amino acids that may or may not be adjacent in the amino acid chain, i.e., they may be separated by other amino acids, spatially, in the sequence whereby k determines the spatial distance between the dimers under consideration. In order to incorporate information from sequential evolution, probabilistic expressions of such dimers are extracted from the PSSM.

These spatial varied amino acid dimer occurrence probabilities have been mathematically summarized in Eq. 1. If P is the PSSM matrix representation for a given protein, P will have L rows and 20 columns, where L is the length of the protein sequence. The probabilistic expression of the mth amino acid to nth amino acid can be computed using Eq. 1 where $1\le m\le 20$ and $1\le n\le 20$.

Eq. 2 constructs a vector $F(k)$ for a particular value of k, which contains 400 elements representing the 400 amino acid dimers possible. As stated previously, k represents the distance between the amino acid positions that are used to compute the probabilistic expression. For $k=1$, the probabilities are computed between neighboring amino acids whereas, for $k=2$, the probabilities are computed between amino acids that are separated by 1 amino acid in the primary sequence. Therefore, for $k=K$, the amino acids used to calculate the probabilities are separated by $K-1$ amino acids.

Upon extracting the features, $F(k)$, for various values of k, it is possible to directly use these features for classification. The different feature sets for various values of k in $F(k)$ can be considered as independent feature sets since they model the probabilistic occurrence of amino acid dimers which are independent for different values of k. Initially during training, individual $F(k)$ for various values of k were evaluated using the SVM classifier (results in Appendix 1).

SVM is a supervised learning model linked to machine learning algorithms that is used for pattern recognition. It is widely used in classification and regression analysis. In its simplest form, SVM accepts a set of inputs and then predicts for each input which of the two possible classes it falls under. For multi-class problems, SVM can still be used by reducing the problem into multiple binary classification problems. SVM aims to construct a hyper-plane in infinite-dimensional space such that a good level of separation is achieved between the classes, thus lowering the generalization error of the classifier. In this paper, *libsvm* version 3.17 is used with kernel function as the *radial basis function* and the C parameter was set to 1000 leaving all other parameters to their respective default values.

Individually, these feature sets provide relatively good classification accuracies (see Appendix 1), indicating that there is discriminatory information present in these features that were extracted using feature extraction technique. Another observation from these results is that there is a gradual decline in the classifier performance as the value of k increases, indicating that there is relatively more information captured when the spatial distances between amino acids are smaller.

Although it is possible to achieve relatively good classification accuracy with the feature sets $F(k)$ individually, in this scheme, we propose to concatenate these various features. Therefore, the concatenated features, F, would signify a feature vector that comprises of the probabilistic expressions of amino acid dimers that have spatial separations from $k=1,2,\ldots,K$, whereby K denotes the upper bound of k in this scheme. This concatenation has been summarized as per Eq. 3 or, for simplicity, re-written as Eq. 4.

$$F=[\,F(1)\quad F(2)\quad \ldots\quad F(K)] \quad (4)$$

This concatenation provides information to explore dependencies between variables of different feature sets. Determining the optimal value of K for concatenation is a key challenge in this scheme. It is important to choose a value of K that leads to balanced classification. A large value of K, primarily, adds too many feature sets that may lead to a decline in classification accuracy whereas a small value of K would mean the loss of potentially discriminative features.

During the experimentation, the value of K was empirically determined by successively incrementing K and observing the classifier performance on the training set of the various datasets. This process was continued until a gradual decline in classification accuracies was observed. Upon performing such analysis, it was determined that the optimal values of K for the DD, EDD and TG datasets are $K=7$, $K=8$ and $K=6$ respectively. The results noted during the analysis are shown in Table 2.

It can be observed from the results in Table 2 that there is a steady increase in classification accuracies as the value of K increases from $K=1$ to a certain value where the classification accuracy reaches a peak for the particular dataset. Upon reaching this peak, any further increase in K leads to a gradual decline in the classification accuracy. Therefore, the identification of the optimal value of K was simplified greatly due to this trend observed during experimentation.

It should be noted that all evaluations up till this stage has been performed by using the training set of the various datasets and these results were analyzed to determine the various parameters. Upon finalizing optimal parameters, the model has been primed for evaluation.

Please cite this article as: Saini, H., et al., Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.05.030

3. Results and discussion

3.1. Datasets

In this research, the proposed scheme was evaluated using three benchmarked datasets, namely the Ding and Dubchak (DD) dataset, extended Ding and Dubchak (EDD) dataset, and Taguchi and Gromiha (TG) dataset.

The DD dataset consists of a training set for the creation of the model and an independent test set for testing queries against the model. The samples belong to 27 SCOP folds which further represent the four major structural classes – α, β, α+β, α/β. The training dataset consists of 311 protein sequences where any given pair of sequences do 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% (Ding and Dubchak, 2001).

The TG-dataset consists of 1612 protein sequences belonging to 30 different folding types of globular proteins from SCOP. The sequence similarity of proteins belonging to the TG dataset is no more than 25% (Taguchi and Gromiha, 2007). The EDD-dataset consists of 3418 proteins with less than 40% sequential similarity belonging to the 27 folds that originally used in DD-dataset. The EDD-dataset was extracted from SCOP in similar manner to Dong et al. (2009) in order to study our proposed method using a larger number of samples. A summary of the datasets has been provided in Table 1.

Since EDD and TG datasets do not have explicit train and test sets, the datasets were split, for the purposes of parameter selection and optimization during the training phase, into train and test sets with 60% of the samples belonging to the train set. It was ensured that the splits were random and equal distributions of folds were maintained.

3.2. Features

For the purposes of comparison and benchmarking, the proposed scheme was evaluated against several other accepted schemes in literature. A brief summary of these schemes is provided in these sections.

3.3. Amino acid composition

The fraction of each of the 20 amino acids within the protein sequence (Ding and Dubchak, 2001).

3.4. Amino acid composition+physicochemical properties

Dubchak used amino acid composition along with features extracted from physicochemical properties including hydrophobicity, predicted secondary structure, polarity, polarizability and normalized van der Waals volume. The size this feature vector is 125 Ding and Dubchak (2001).

Table 1
Summary of datasets.

Dataset	Number of folds	Train samples	Test samples	Total samples
DD	27	311	383	694
EDD[a]	27	–	–	3418
TG[a]	30	–	–	1612

[a] These datasets do not have benchmarked separate train and test sets.

3.5. Residue bigram probabilities

This feature incorporates the probabilities of the occurrence of all amino acid dimer pairs in the amino acid sequence. This is a 400 dimensional feature vector (Ghanty and Pal, 2009).

3.6. Alternate bigram probabilities

This feature vector consists of the probabilities of the occurrence of all possible pairs of amino acid which are separated by one residue in the protein sequence. This is also a 400 dimensional feature vector (Ghanty and Pal, 2009).

3.7. Occurrence

The amino acid occurrence, i.e., the un-normalized number of each amino acid is used instead of amino acid composition. Dubchak features other than amino acid composition are also used. The size of this feature vector is 125 (Taguchi and Gromiha, 2007).

3.8. Position specific Scoring matrix based bi-grams

These bigrams represent the probabilities of transition from one amino acid to another as determined by PSSM. The size of this feature is 400 (Sharma et al., 2013).

3.9. Position specific scoring matrix based tri-gram

These trigrams represent the probabilities of transitions of triplets of amino acids as determined by PSSM. The size of this feature is 8000 (Paliwal et al., 2014).

3.10. Alignment via dynamic time warping

This scheme predicts protein folds based on the alignment distance of the protein sequences using dynamic time warping (Lyons et al., 2014).

3.11. Experiment results

The experimentation was performed on the benchmarked datasets to evaluate the performance of the classification scheme described previously. In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling or *n*-fold crossover test, and jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset (Chou and Shen, 2010). However, to reduce the computational time, we adopted the *n*-fold cross-validation in this study, which has done by many investigators with SVM as the

Table 2
Performance of concatenated amino acid dimers during training with the terminating value for spatial separation from $K=1,2,\ldots,10$.

K	DD	EDD	TG
1	66.0	82.1	63.8
2	66.5	85.6	67.5
3	66.7	85.9	68.2
4	66.9	86.3	68.4
5	67.0	86.4	68.6
6	66.9	86.2	**68.7**
7	**67.6**	86.3	68.2
8	65.9	**86.6**	67.8
9	65.8	85.9	67.5
10	65.4	85.3	67.4

Please cite this article as: Saini, H., et al., Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.05.030

prediction engine. This strategy of performance evaluation is widely employed by researchers in literature. For statistical stability, *n*-fold cross-validation was repeated 100 times with random subsampling.

The proposed scheme was compared against various other schemes that use information structural and evolutionary information for fold recognition. These techniques included PF1 and PF2 (Ghanty and Pal, 2009), PF (Yang et al., 2011), Occurrence (O) (Taguchi and Gromiha, 2007), AAC and AAC+HXPZV (Ding and Dubchak, 2001), which compute feature sets from the original protein sequences. In addition, ACC (Dong et al., 2009), bi-gram (Sharma et al., 2013), tri-gram (Paliwal et al., 2014) and alignment (Lyons et al., 2014) are also included since they compute features directly from the evolutionary information present in PSSM. Moreover, features have been computed from the consensus sequences for PF1, PF2, O, AAC and AAC+HXPZV to obtain additional feature sets for comparison. In the tables that highlight the performance of these various techniques, a prefix of PSSM+ indicates that the features have been computed on the consensus sequence. These techniques were evaluated using n-fold cross-validation testing for $n=5, 6, \ldots, 10$.

These techniques have been evaluated using the DD, EDD and TG datasets using *n*-fold cross-validation for $n=5, 6, \ldots, 10$. It can be seen that *k*-separated bigrams outperform every other technique for all the three datasets. It can be seen from Table 3 that for the DD dataset, the proposed scheme yields the highest accuracy of 76.7%, an improvement of about 3%. Similarly, in Tables 4 and 5, the proposed scheme shows raw improvements of about 1–3% in both the TG and EDD datasets. It should be noted that the EDD-dataset is a much larger dataset that has a sequence similarity threshold of 40%, which may contribute for a better performance of the Alignment Method compared to the proposed technique.

Additionally, to further evaluate the performance of the proposed technique against other techniques, analysis of sensitivity and specificity of these techniques have been conducted. These measures highlight the relative performance of a classifier in a multi-class classification problem. The mathematical definitions of these measures are given as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}} \tag{6}$$

In the above equations, TP represents the true positives (members of the positive class correctly identified correctly), FP represents the false positives (members of the negative class incorrectly identified as positive), TN represents the true negatives (members of the negative class identified correctly) and FN represents the false negatives (members of the positive class incorrectly identified as negative).

Table 3
Performance (in % accuracy) of various feature sets on the DD-dataset using n-fold cross validation procedure.

Feature sets	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$
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	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	43.6	43.9	44.2	44.8	44.6	45.1
AAC+HXPZV	45.1	46.2	46.5	46.8	46.9	47.2
ACC	65.7	66.6	66.8	67.5	67.7	68.0
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	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	57.5	58.1	58.4	58.7	59.1	59.2
PSSM+AAC+HXPZV	55.9	56.9	57.1	57.7	58.0	58.2
Bi-gram	72.6	73.1	73.7	73.7	74.1	74.1
Tri-gram	72.1	72.6	73.0	73.2	73.7	73.8
Alignment (DTW)	72.6	73.5	73.8	74.2	74.7	74.7
This paper	**74.5**	**75.4**	**75.9**	**76.4**	**76.5**	**76.7**

Table 4
Performance (in % accuracy) of various feature sets on the TG-dataset using n-fold cross validation procedure.

Feature sets	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$
PF1	38.1	38.4	38.6	38.7	38.8	38.8
PF2	38.0	38.4	38.5	38.6	38.7	38.8
PF	42.3	42.6	42.7	43.0	43.0	43.1
O	35.8	36.1	36.2	36.1	36.3	36.3
AAC	31.5	31.5	31.7	31.8	31.9	32.0
AAC+HXPZV	35.7	36.0	36.1	36.2	36.3	36.3
ACC	64.9	65.4	65.9	66.2	66.4	66.4
PSSM+PF1	51.1	51.5	52.0	52.3	52.4	52.7
PSSM+PF2	50.2	50.4	50.7	50.8	51.0	51.1
PSSM+PF	57.2	57.8	58.0	58.3	58.5	58.8
PSSM+O	46.0	46.3	46.5	46.5	46.7	46.7
PSSM+AAC	43.2	43.5	43.6	43.8	43.8	44.0
PSSM+AAC+HXPZV	45.6	45.9	46.0	46.2	46.3	46.6
Bi-gram	67.1	67.5	67.6	67.8	68.1	68.1
Tri-gram	71.4	71.7	72.3	73.3	72.4	72.5
Alignment (DTW)	72.0	72.7	73.0	73.5	73.6	74.0
This paper	**73.1**	**73.6**	**73.9**	**74.2**	**74.3**	**74.5**

Table 5
Performance (in % accuracy) of various feature sets on the EDD-dataset using n-fold cross validation procedure.

Feature sets	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$
PF1	50.2	50.5	50.5	50.7	50.8	50.8
PF2	49.3	49.5	49.7	49.8	49.8	49.9
PF	54.7	55.0	55.2	55.4	55.5	55.6
O	46.4	46.6	46.6	46.7	46.7	46.9
AAC	40.3	40.6	40.7	40.7	40.9	40.9
AAC+HXPZV	40.2	40.4	40.6	40.7	40.9	40.9
ACC	84.9	85.2	85.4	85.6	85.8	85.9
PSSM+PF1	74.1	74.5	74.7	75.0	75.1	75.2
PSSM+PF2	73.7	74.1	74.5	74.6	74.7	74.9
PSSM+PF	78.2	78.6	78.8	79.0	79.1	79.3
PSSM+O	67.6	68.0	68.1	68.3	68.3	68.5
PSSM+AAC	60.9	61.3	61.5	61.6	61.7	61.9
PSSM+AAC+HXPZV	66.7	67.2	67.4	67.7	67.8	67.9
Bi-gram	83.6	84.0	84.1	84.3	84.3	84.5
Tri-gram	85.7	85.9	86.0	86.1	86.2	86.2
Alignment (DTW)	89.4	89.7	89.9	90.0	90.1	90.2
This paper	**89.1**	**89.5**	**89.6**	**89.7**	**89.8**	**89.9**

The results of this analysis, via 10-fold cross validation over hundred iterations for statistical stability, have been illustrated in Figs. 2–4 where it can be clearly seen that the proposed method is on par or better than most of the other techniques.

4. Conclusions

In this study, we have proposed a scheme that incorporates a feature extraction technique based on sequential evolution probabilities using amino acid dimers with varying degree of spatial separation. This technique does not ignore relations between amino acid pairs that are non-adjacent in the primary sequence and it extracts information from amino acids with varying spatial distances in the sequence to compute the features directly from PSSM, which provides more discriminatory information that may lead to better performance during classification. These features were concatenated and processed using SVM for protein fold recognition.

The proposed technique gave promising results, and the highest recorded for *n*-fold cross-validation accuracies on DD, EDD and

Please cite this article as: Saini, H., et al., Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.05.030

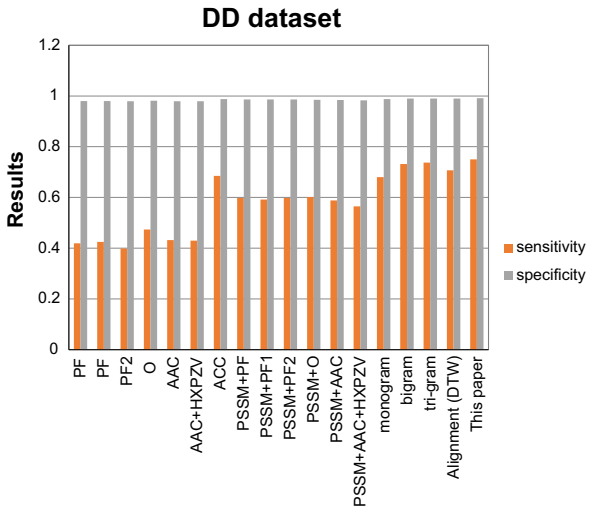

Fig. 2. Plot of sensitivity, specificity and precision values for DD dataset.

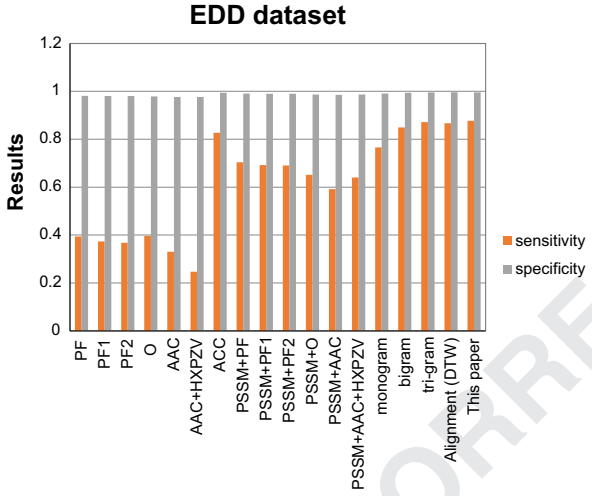

Fig. 3. Plot of sensitivity, specificity and precision values for EDD dataset.

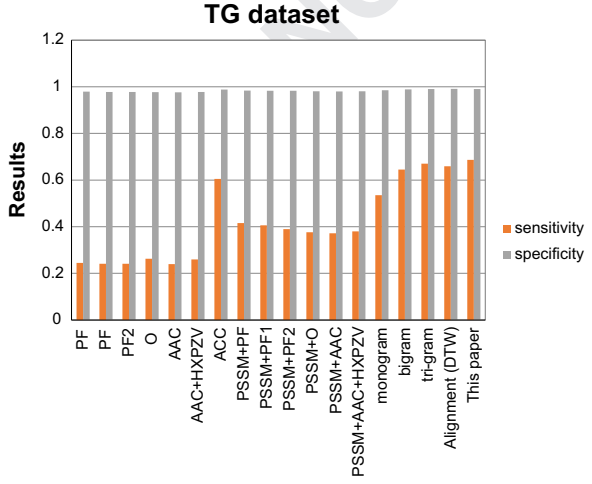

Fig. 4. Plot of sensitivity, specificity and precision values for TG dataset.

TG datasets with this scheme were 76.7%, 89.9% and 74.5%, respectively. These were amongst the highest noted while comparing with other techniques. As pointed out in (Chou and Shen, 2009) and demonstrated in a series of recent publications (Chen et al., 2014; Lin et al., 2014; Liu et al., 2015; Xu et al., 2013; Chen et al., 2012; Chen et al., 2013), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods or predictors, we shall make efforts in our future work to provide a web-server or an open source library for the techniques presented in this study.

Table A1
Individual classification accuracies of k-separated bigrams for $k=1,2,\ldots,10$ using training set.

k	DD	EDD	TG
1	65.8	82.2	63.8
2	64.2	82.5	64.3
3	65.0	81.2	61.9
4	63.7	81.2	61.5
5	63.1	79.0	59.5
6	62.2	78.6	58.1
7	60.6	78.8	58.4
8	62.2	77.3	56.9
9	61.9	76.8	56.4
10	60.8	76.6	56.1

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceptualization of the idea and the subsequent experimentation was performed by HS, GR and AS. Additionally, AS and SL provided advise and assisted in the preparation of the manuscript along with HS and GR. AD, JL and KKP extracted the required datasets and retrieved PSSM probabilities using the PSI-Blast program. SI and SM assisted the group by providing advice and financial assistance.

Acknowledgments

This project has been partially funded by the Human Genome Centre, University of Tokyo.

Appendix 1

The results shown in Table A1 below highlight that individual feature sets provide relatively good classification accuracies, indicating that there is discriminatory information present in these features that were extracted using k-separated bigrams. Another trend that is clearly visible is that the accuracies decrease steadily as the spatial distance (k) between amino acids in the sequence increases, indicating that, relatively, there is addition of noise or lack of discriminatory information when the spatial distances increase.

References

Anand, A., Pugalenthi, G., Suganthan, P.N., 2008. Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J. Theor. Biol. 253, 375–380.
Berardi, M.J., Shih, W.M., Harrison, S.C., Chou, J.J., 2011. Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. Nature 476, 109–113.
Cai, Y.-D., Zhou, G.-P., 2000. Prediction of protein structural classes by neural network. Biochimie 82, 783–785.
Cai, Y.-D., Liu, X.-J., Xu, X., Chou, K.-C., 2002. Prediction of protein structural classes by support vector machines. Comput. Chem. 26, 293–296.
Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., 2013. Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29, 960–962.

Chen, C., Shen, Z.-B., Zou, X.-Y., 2012. Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition. Protein Pept. Lett. 19, 422–429.
Chen, K., Kurgan, L., Ruan, J., 2008. Prediction of protein structural class using novel evolutionary collocation based sequence representation. J. Comput. Chem. 29, 1596–1604.
Chen, W., Lin, H., Feng, P.-M., Ding, C., Zuo, Y.-C., Chou, K.-C., 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 7, e47843.
Q5 Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucl. Acids Res. 41e68
Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2014a. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed. Res. Int. 2014, 623149.
Chen, W., Feng, P.-M., Deng, E.-Z., Lin, H., Chou, K.-C., 2014b. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem. 462, 76–83.
Chinnasamy, A., Sung, W.-K., Mittal, A., 2005. Protein structure and fold prediction using tree-augmented naive bayesian classifier. J. Bioinform. Comput. Biol. 3, 803–819.
Chmielnicki, W, 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 Struct. Funct. Bioinform. 43, 246–255.
Chou, K.-C., 2004. Structural bioinformatics and its impact to biomedical science. Curr. Med. Chem. 11, 2105–2134.
Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Chou, K.-C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218–234.
Chou, K.-C., Shen, H.-B., 2009. REVIEW : recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 01, 63–92.
Chou, K.-C., Shen, H.-B., 2010. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS One 5, e11335.
Dehzangi, A., Paliwal, K., Sharma, A., Dehzangi, O., Sattar, A., 2013. A combination of feature
Q6 extraction methods with an ensemble of different classifiers for protein structural class prediction problemIEEE/ACM Trans Comput Biol bioinformatics, 10; 564–575.
Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., Sattar, A., 2013. Enhancing protein fold prediction accuracy using evolutionary and structural featuresPattern Recognit. Bioinform.196–207
Ding, C.H.Q., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.
Ding, H., Deng, E.-Z., Yuan, L.-F., Liu, L., Lin, H., Chen, W., Chou, K.-C., 2014. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. Biomed. Res. Int. 2014, 286419.
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 Recognit. Lett. 29, 1887–1892.
Dong, Q., Zhou, S., Guan, J., 2009. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 25, 2655–2662.
Du, P., Wang, X., Xu, C., Gao, Y., 2012. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal. Biochem. 425, 117–119.
Du, P., Gu, S., Jiao, Y., 2014. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int. J. Mol. Sci. 15, 3495–3506.
Q7 Dubchak, I., Muchnik, I., Kim S.H., 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. Kalkidiki, Greece. pp. 104–107.
Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., Kim, S.H., 1999. Recognition of a protein fold in the context of the SCOP classification. Proteins Struct. Funct. Bioinform. 35, 401–407.
Ghanty, P., Pal, N.R., 2009. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans. NanoBiosci. 8, 100–110.
Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425.
Huang, J.T., Tian, J., 2006. Amino acid sequence predicts folding rate for middle size two state proteins. Proteins Struct. Funct. Bioinform. 63, 551–554.
Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Asadabadi, E.B., 2007a. Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys. Chem. 128, 87–93.
Jahandideh, S., Abdolmaleki, P., Jahandideh, M., Hayatshahi, S.H.S., 2007b. Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes. J. Theor. Biol. 244, 275–281.
Kedarisetti, K.D., Kurgan, L., Dick, S., 2006. Classifier ensembles for protein structural class prediction with varying homology. Biochem. Biophys. Res. Commun. 348, 981–988.
Klein, P., 1986. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta – Protein Struct. Mol. Enzymol. 874, 205–215.
Kong, L., Zhang, L., Lv, J., 2014. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 344, 12–18.
Kurgan, L., Chen, K., 2007. Prediction of protein structural class for the twilight zone sequences. Biochem. Biophys. Res. Commun. 357, 453–460.
Kurgan, L., Zhang, T., Zhang, H., Shen, S., Ruan, J., 2008. Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551–564.
Li, Z.-C., Zhou, X.-B., Dai, Z., Zou, X.-Y., 2009. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415–425.
Liao, B., Xiang, Q., Li, D., 2012. Incorporating secondary features into the general form of Chou's PseAAC for predicting protein structural class. Protein Pept. Lett. 19, 1133–1138.
Lin, H., Deng, E.-Z., Ding, H., Chen, W., Chou, K.-C., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo *k*-tuple nucleotide composition. Nucl. Acids Res. 42, 12961–12972.
Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., Chou, K.-C., 2014. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9, e106691.
Liu, L., Hu, X.-Z., Liu, X.-X., Wang, Y., Li, S.-B., 2012. Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions. Protein Pept. Lett. 19, 439–449.
Liu, T., Jia, C., 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol. 267, 272–275.
Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., 2012. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42, 2243–2249.
Liu, Z., Xiao, X., Qiu, W.-R., Chou, K.-C., 2015. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. 474, 69–77.
Lyons, J., Biswas, N., Sharma, A., Dehzangi, A., Paliwal, K.K., 2014. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. J. Theor. Biol. 354, 137–145.
Murzin, A.G., Brenner, S.E., Hubbar, 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. Bioinform. 39, 261–268.
OuYang, B., Xie, S., Berardi, M.J., Zhao, X., Dev, J., Yu, W., Sun, B., Chou, J.J., 2013. Unusual architecture of the p7 channel from hepatitis C virus. Nature 498, 521–525.
Paliwal, K.K., Sharma, A., Lyons, J., Dehzangi, A., 2014. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Trans. NanoBiosci. 13, 44–50.
Qin, Y.-F., Wang, C.-H., Yu, X.-Q., Zhu, J., Liu, T.-G., Zheng, X.-Q., 2012. Predicting protein structural class by incorporating patterns of over-represented *k*-mers into the general form of Chou's PseAAC. Protein Pept. Lett. 19, 388–397.
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.
Saini, H., Raicar, G., Lal, S., Dehzangi, A., Lyons, J., Paliwal, K.K., Imoto, S., Miyano, S., Sharma, A., 2014b. Genetic algorithm for an optimized weighted voting scheme incorporating *k*-separated bigram transition probabilities to improve protein fold recognition. In Proceedings of the 2014 Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE). IEEE, pp. 1–7.
Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Ananthanarayanan, R., Lyons, J., Biswas, N., Paliwal, K.K., 2014a. Protein structural class prediction via *k*-separated bigrams using position specific scoring matrix. J. Adv. Comput. Intell. Intell. Inform. 18, 474–479.
Schnell, J.R., Chou, J.J., 2008. Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451, 591–595.
Sharma, A., Paliwal, K.K., 2007. Fast principal component analysis using fixed-point algorithm. Pattern Recognit. Lett. 28, 1151–1155.
Sharma, A., Paliwal, K.K., 2010. Regularisation of eigenfeatures by extrapolation of scatter-matrix in face-recognition problem. Electron. Lett. 46, 682–683.
Sharma, A., Paliwal, K.K., 2012a. A two-stage linear discriminant analysis for face-recognition. Pattern Recognit. Lett. 33, 1157–1162.
Sharma, A., Paliwal, K.K., 2012c. A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. Pattern Recognit. 45, 2205–2213.
Sharma, A., Paliwal, K.K., Onwubolu, G.C., 2006. Class-dependent PCA, MDC and LDA: a combined classifier for pattern classification. Pattern Recognit. 39, 1215–1229.
Sharma, A., Koh, C.H., Imoto, S., Miyano, S., 2011. Strategy of finding optimal number of features on gene expression data. Electron. Lett. 47, 480–482.
Sharma, A., Imoto, S., Miyano, S., 2012d. A filter based feature selection algorithm using null space of covariance matrix for DNA microarray gene expression data. Curr. Bioinform. 7, 289–294.
Sharma, A., Paliwal, K.K., 2012b. A gene selection algorithm using bayesian Q8 classification approach. Am. J. Appl. Sci.9, 127.
Sharma, A., Imoto, S., Miyano, S., Sharma, V., 2012b. Null space based feature selection method for gene expression data. Int. J. Mach. Learn. Cybern. 3, 269–276.
Sharma, A., Imoto, S., Miyano, S., 2012c. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. 9, 754–764.
Sharma, A., Imoto, S., Miyano, S., 2012a. A between-class overlapping filter-based method for transcriptome data analysis. J. Bioinform. Comput. Biol.10, 1250010.
Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K., 2013a. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. 320, 41–46.

Please cite this article as: Saini, H., et al., Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.05.030

Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S., 2013b. Principal component analysis using QR decomposition. Int. J. Mach. Learn. Cybern. 4, 679–683.
Sharma, A., Paliwal, K.K., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., 2013c. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinform. 14, 233.
Sharma, A., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., Nakai, K., Patil, A., 2014. Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function. PLoS One 9, e89890.
Shen, H.-B., Chou, K.-C., 2006. Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.
Shen, H.-B., Chou, K.-C., 2008. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386–388.
Taguchi, Y.H., Gromiha, M.M., 2007. Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinform. 8, 404.
Xu, Y., Shao, X.-J., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2013. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1, e171.
Xu, Y., Wen, X., Wen, L.-S., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2014. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9, e105018.
Yang, T., Kecman, V., Cao, L., Zhang, C., Zhexue Huang, J., 2011. Margin-based ensemble classifier for protein fold recognition. Expert Syst. Appl. 38, 12348–12355.
Zhang, H., Zhang, T., Gao, J., Ruan, J., Shen, S., Kurgan, L., 2012. Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. Amino Acids 42, 271–283.
Zhang, L., Zhao, X., Kong, L., 2014. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 355, 105–110.

Please cite this article as: Saini, H., et al., Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.05.030