Protein Fold Recognition by Alignment of Amino Acid Residues Using Kernelized Dynamic Time Warping

James Lyons[1], Alok Sharma[2,3,*], Abdollah Dehzangi[3,4], Kuldip K. Paliwal[1],
[1]School of Engineering, Griffith University, Australia
[2]School of Engineering and Physics, University of the South Pacific, Fiji
[3]Institute for Integrated and Intelligent Systems (IIIS), Brisbane, Australia
[4]National ICT Australia (NICTA), Brisbane, Australia
*Correspondent author

Abstract

Identifying the tertiary structure of a protein is a challenging task in biological science. Protein fold recognition is an intermediate step in identifying the tertiary structure. In protein fold recognition, a protein is classified into one of its folds. The recognition of a protein fold can be done by employing feature extraction methods to extract relevant information from protein sequences and then by using a classifier to accurately recognize novel protein sequences. In the past, several feature extraction methods have been developed but with limited recognition accuracy only.

Protein sequences of varying lengths share the same fold and therefore they are very similar (in a fold) if aligned properly. To this, we develop an amino acid alignment method to extract important features from protein sequences by computing dissimilarity distances between proteins. This is done by measuring distance between two respective position specific scoring matrices of protein sequences which is used in a support vector machine framework. We demonstrated the effectiveness of the proposed method on several benchmark datasets. The method shows significant improvement in the fold recognition performance which is in the range of 4.3% to 7.6% compared to several other existing feature extraction methods.

Introduction

In biological sciences, deciphering the tertiary structures of proteins is considered to be an important and challenging task. The identification of tertiary structures provides information about protein functions which helps in understanding protein heterogeneity, protein-protein interactions and protein-peptide interactions. The computational ways of determining protein structures has gained considerable attention since it is normally very time consuming to identify protein structures by crystallography methods. Protein fold recognition is an intermediate step in the process of recognizing tertiary structure. The objective of protein fold recognition is to associate a fold to a novel protein sequence.

Protein fold recognition broadly covers feature extraction and classification tasks. The brief description of the work conducted in the past has been depicted in the Related Work Section. It has been shown in the literature that feature extraction methods using evolutionary information performs quite well in the fold recognition process (Altschul et al., 1997; Dong et al., 2009; Sharma et al., 2013). In this work, we have used this information to build a feature extraction method for protein-protein alignment. For this, we extract position specific scoring matrices (PSSMs) using PSI-BLAST and build dissimilarity matrix between two protein sequences and conduct dynamic time warping to find the alignment path. Since different proteins with varying lengths share the same fold, features extracted from aligned homologous proteins give discriminant features for protein fold recognition. In order to illustrate this, we picked 7 protein sequences and extracted their corresponding PSSMs for comparison. Out of 7 protein sequences, 4 protein sequences (Proteins A, B1, B2 and B3 in Figure 1) belong to a particular fold and the remaining 3 protein sequences (Protein C1, C2 and C3 in Figure 1) belong to different folds. We then used Protein A (see Figure 1) and found dissimilarity matrices by comparing it with all the 6 remaining protein sequences. In the first three dissimilarity matrices (i, ii and iii), PSSMs from protein sequences in the same fold are compared and the next three dissimilarity matrices (iv, v and vi), protein sequences of mutually different folds are compared. We can observed that in the first 3 figures (i, ii and iii), a diagonal path can be seen (we call an alignment path), however, in the next 3 figures, this alignment path is not clearly observed. This alignment path (which shows the dissimilarity between two proteins) can be used to distinguish between proteins of one fold with that of another fold. This is a typical example, there could be variations depending upon different proteins. Nonetheless, dissimilarity distance could be used as a measure to observe dissimilarity between proteins. From biological perspective, proteins in the same fold often have amino acid subsequences that are highly conserved. The alignment path (i.e., the dissimilarity distance) characterizes the subsequences of amino acids in these conserved regions via their PSSMs. If a certain subsequence is conserved in a fold, then each protein in that fold would have a low dissimilarity distance from that conserved region. This can help in discriminating folds that do not have the same amio acid subsequences. The details of the proposed scheme are described later. The proposed scheme provides promising results (in terms of recognition performance) when experimented on 3 benchmark datasets: Ding and Dubchak (DD) (Ding and Dubchak, 2001), Taguchi and Grohima (TG) (Taguchi and Gromiha, 2007) and extended DD (EDD) (Dong et al., 2009). The 10-fold cross-validation recognition performance on DD dataset is 74.7%, on TG dataset is

74.0% and on EDD dataset is 90.2% which is very promising when compared with other existing feature extraction methods.

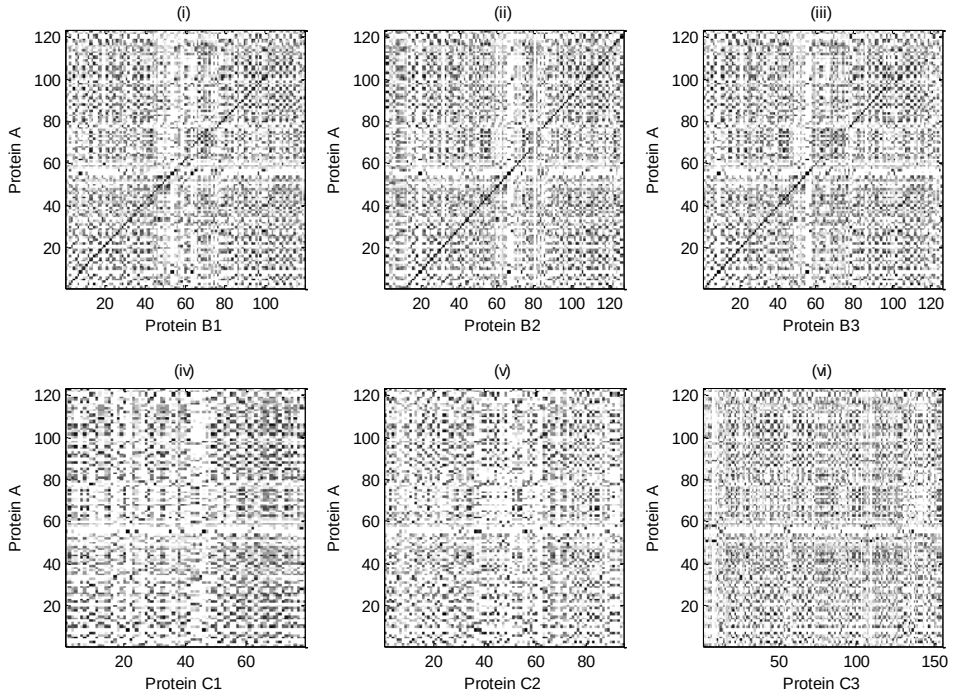

Figure 1: An illustration using dissimilarity matrix of protein sequences. The pictures above represent similarities computed between PSSMs. Dark pixels indicate corresponding rows of each PSSM are very similar. Long sequences of similar PSSM rows manifest as dark lines in the pictures. The top 3 pictures are from proteins in the same fold, the bottom three all proteins are from different folds. The contrast of these pictures has been increased for clarity in viewing.

Related work

The development of protein fold recognition research can be broadly categorized into two main tasks: feature extraction and classification. For the former task, several feature extraction techniques have been developed using structural, physicochemical and evolutionary information. 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 the 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 et al., 2012; Dehzangi et al., 2013a, 2013b and 2013c). Apart from the above mentioned 5 attributes used by Dubchak et al. (1997), features have also been extracted by incorporating other attributes of 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). These physicochemical attributes are selected in an arbitrary way and recently a systematic way of selecting physicochemical attributes was proposed by Sharma et al. (2013a; 2012). Ohlson et al., (2004) proposed a profile-profile alignment method to improve protein fold recognition. 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 concatenated in the study conducted by Yang et al., (2011), thereby, having 800 features. Chou (2001) proposed pseudo-amino acid composition (A) based features to effectively represent a protein sequence. Dong et al., (2009) have shown autocross-covariance (ACC) transformation for protein fold recognition. 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. Sharma et al. (2013b) have derived bi-gram features using evolutionary information (PSSM). For the latter task 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 (Dehzangi et al., 2009, 2010; Yang et al., 2011; Dehzangi et al., 2010, Dehzangi and Karamizadeh, 2011). Among these classifiers, SVM (or SVM-based

for ensemble strategy) classifier exhibits quite promising results (Liu et al., 2012; Kurgan et al., 2008; Ghanty and Pal, 2009).

The extraction of relevant and informative features from protein sequences is a crucial step in identifying protein folds. In order to improve protein fold recognition, we focus on carefully developing the feature extraction method. Since SVM classifier (Vapnik, 1995) provides high recognition accuracy, we use SVM classifier to compare the performance of our feature extraction method with other feature extraction methods. SVM classifiers are often employed with the Radial Basis Function (RBF) kernel. The RBF kernel (along with other common SVM kernels such as the linear and polynomial kernel) requires fixed length feature vectors. This has motivated many previous works to try and extract fixed length representations of proteins so that they can then be efficiently compared. In this work we define a kernel designed to work with variable length data. This allows us to directly compare PSSM matrices, instead of first transforming the matrix into a fixed length vector prior to comparison.

Dataset

In this study, three protein sequence datasets have been used: 1) DD-dataset (Ding and Dubchak, 2001), 2) TG-dataset (Taguchi and Gromiha, 2007) and 3) EDD-dataset (Dong et al., 2009). The DD-dataset that we have used 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 Structural Classification of Proteins (SCOP) folds which represent all major structural classes: α, β, α/β, and $\alpha+\beta$ (Ding and Dubchak, 2001). The training set and test set have been merged as a single set of data in order to perform the k-fold cross-validation process.

The TG-dataset consists of 1612 protein sequences belonging to 30 different folding types of globular proteins from SCOP. The names of the number of protein sequences in each of 30 folds have been described in Taguchi and Gromiha (2007). The sequence similarity of protein of TG datasets is no more than 25%.

The EDD-dataset consists of 3418 proteins with less than 40% sequential similarity belonging to the 27 folds that originally used in DD-dataset. We extracted the EDD-dataset from SCOP in similar manner to Dong et al., (2009) in order to study our proposed method using a larger number of samples.

Amino acid alignment method

In this section, we present the proposed feature extraction method based on the alignment of proteins. To present the overview, a flow diagram of the proposed scheme has been shown in Figure 2. The model can be subdivided into the training phase and test phase. In the training phase a set of protein sequences is used to estimate the model parameters and in the test phase, the fold of a novel protein sequence is identified. During the training of the model, PSSM matrices of protein sequences are computed by using PSI-BLAST. In the pairwise analysis step, row vectors of two PSSM matrices are used to measure pairwise distance. By comparing all the row vectors in two PSSMs we get a dissimilarity matrix. This dissimilarity matrix is then used in dynamic time warping (DTW) stage to compute dissimilarity distance between two PSSM matrices of the corresponding proteins. The obtained dissimilarity distance is then used in the kernelization stage to compute kernel distance. A protein is compared progressively with all other proteins to form a kernel matrix. This kernel matrix will then be used to train SVM parameters. Once the model parameters are estimated then the system can determine the fold of a novel protein sequence.

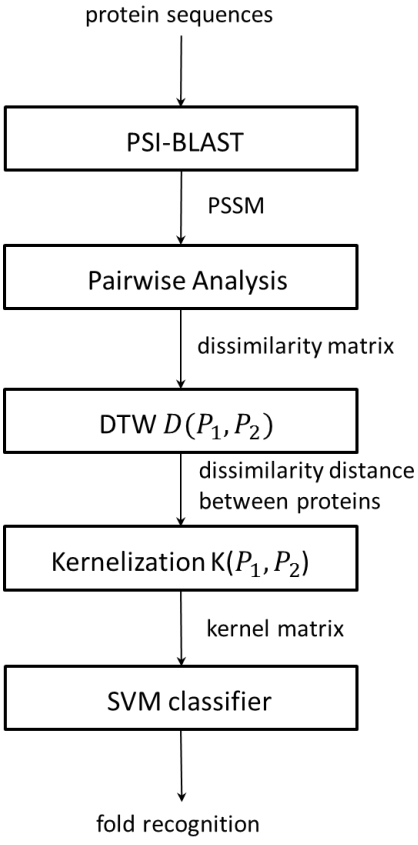

Figure 2: A flow-diagram of protein sequence classification using alignment method.

Let P and Q be the matrices representing PSSM (log probabilities) of two protein sequences of length L_1 and L_2, respectively. PSSM matrix can be interpreted as the relative probability of substitution of amino acids. The matrix P will have L_1 rows and 20 columns and the matrix Q will have L_2 rows and 20 columns. Let p_i (for $i = 1,2, \dots L_1$) and q_j (for $j = 1,2, \dots, L_2$) be the row vectors of P and Q, respectively. The dissimilarity cosine distance between p_i and q_j can be given as

$$d(p_i, q_j) = 1 - \frac{p_i q_j^{\mathrm{T}}}{\sqrt{p_i p_i^{\mathrm{T}} q_j q_j^{\mathrm{T}}}}, \text{ for } i = 1,2 \dots, L_1 \text{ and } j = 1,2, \dots, L_2 \quad (1)$$

Calculating distance d for all L_1 rows and L_2 rows would give a $L_1 \times L_2$ dissimilarity matrix S. We then employ dynamic time warping to find the minimum cost path through the dissimilarity matrix S. This would give cumulative dissimilarity matrix D. The matrix D defines the total cost of alignment between (p_1, q_1) and (p_i, q_j). Lower cost implies a better alignment, which indicates that the proteins are more similar. The computation of cumulative dissimilarity matrix D can be done in the following way

$$D_{i,j} = \min(D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}) + S_{i,j}, \text{ for } i = 1,2, \dots, L_1 \text{ and } j = 1,2, \dots, L_2 \quad (2)$$

where $D_{i,j} = []$ (empty set) for $i \leq 0$ and/or $j \leq 0$ and $S_{i,j} = d(p_i, q_j)$.

We define the distance between two PSSM matrices P and Q, as $D_{dtw}(P, Q)$. This can be expressed as $D_{dtw}(P, Q) = D_{L_1, L_2}$. The distance D_{dtw} represents dissimilarity between the aligned proteins. The kernel distance between PSSM matrices P and Q, can be represented as $K(P, Q)$, where γ is a kernel parameter (chosen by performing cross-validation on the training set). The kernel function $K(P, Q)$ is defined by $\exp(-D_{dtw}(P, Q)^2/\gamma^2)$. We then compute the kernel distance between all the pairs of proteins in the training set. This gives a kernel matrix K having n rows and n columns, where n is the number of training samples. The kernel matrix K is then further processed through the SVM classifier for parameter estimation and classification.

Support vector machine as a classifier

In this paper we used SVM (Vapnik, 1995) as a classifier. SVM is considered to be the state-of-the-art machine learning and pattern classification algorithm. It has been extensively applied in classification and regression tasks. SVM aims to find maximum

margin hyper-plane (MMH) to minimize classification error. In SVM 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). This function K is usually considered as RBF kernel, polynomial kernel or linear kernel. These kernels require fixed length feature vectors. Since the protein sequences are of varying lengths, we can't use these kernels. However, in this work we have defined a kernel function that can cater for this varying length (of proteins) problem without limiting the proteins to a fixed length vector. This would provide SVM more relevant and useful information for protein fold recognition.

In order to find a decision boundary between two folds, SVM attempts to maximize the margin between the folds, 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

$$y' = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}') + b) \quad (3)$$

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 weights and b is a bias. We use libsvm (Chang and Lin, 2011) for training and testing with our kernel function.

An illustration of alignment method using a toy problem

In order to illustrate the alignment method, let us consider a toy example of two protein sequences $P = VARA$ and $Q = VVARA$ of corresponding length $L_1 = 4$ and $L_2 = 5$, respectively. Note that we assume that the toy proteins are made of 3 amino acids A, R and V. Table 1a and Table 1b show the PSSM of these proteins.

Table 1a: PSSM of the protein P

Amino acids	A	R	V
V	1	5	6
A	4	6	-3
R	3	3	1
A	5	3	-1

Table 1b: PSSM of the protein Q

Amino acids	A	R	V
V	2	3	2
V	5	2	-3
A	5	4	6
R	0	4	1
A	2	0	-1

Let p_i (for $i = 1, \dots, 4$) and q_j (for $j = 1, \dots, 5$) are the row vectors of PSSMs of P and Q, respectively. To compute the dissimilarity distance between row 1 of Table 1a ($p_1 =$ [1,5,6]) and row 1 of Table 1b ($q_1 =$ [2,3,2]), we employ equation 1 as follows:

$$d(p_1, q_1) = 1 - \frac{p_1 q_1^{\mathrm{T}}}{\sqrt{p_1 p_1^{\mathrm{T}} q_1 q_1^{\mathrm{T}}}}$$

$d(p_1, q_1) = 1 - 0.8933 = 0.1067$ (since $p_1 q_1^{\mathrm{T}} = 29$; $p_1 p_1^{\mathrm{T}} = 62$ and $q_1 q_1^{\mathrm{T}} = 17$)

In a similar way, dissimilarity distance can be computed between all the rows of Table 1a and Table 1b. This would give similarity matrix S as follows:

$$S = \begin{bmatrix} 0.1067 & 1.0618 & 0.1171 & 0.1991 & 1.2272 \\ 0.3789 & 0.1484 & 0.6206 & 0.3479 & 0.3701 \\ 0.0541 & 0.3301 & 0.1372 & 0.2767 & 0.4870 \\ 0.3031 & 0.0677 & 0.4029 & 0.5490 & 0.1685 \end{bmatrix}$$

Dissimilarity matrix S is used in computing cumulative dissimilarity matrix D using dynamic programming (equation 2) to find the minimum cost path (alignment path) as follows:

$$D_{11} = \min(D_{01}, D_{10}, D_{00}) + S_{11}$$
$$= S_{11} = 0.1067 \text{ (since } D_{01},\ D_{10} \text{ and } D_{00} \text{ do not exist and considered as empty)}$$

In a similar way, we can compute $D_{21} = D_{11} + S_{21} = 0.4856$; $D_{12} = D_{11} + S_{12} = 1.1685$ and $D_{22} = \min(D_{12}, D_{21}, D_{11}) + S_{22} = 0.1067 + 0.1484 = 0.2552$. The computed matrix D is given as follows:

$$D = \begin{bmatrix} 0.1067 & 1.1685 & 1.2856 & 1.4847 & 2.7119 \\ 0.4856 & 0.2551 & 0.8757 & 1.2236 & 1.5937 \\ 0.5397 & 0.5852 & 0.3923 & 0.6690 & 1.1560 \\ 0.8428 & 0.6074 & 0.7952 & 0.9413 & 0.8375 \end{bmatrix}$$

By using matrix D, the distance between two proteins can be computed which is simply given by $D_{dtw}(P,Q) = D(4,5) = 0.8375$. Suppose the kernel parameter $\gamma = 10$ (evaluated by doing cross-validation on the training set) then kernel distance would be $K(P,Q) = \exp(-D_{dtw}(P,Q)^2/\gamma^2) = \exp(-0.8375^2/10\text{^}2) = 0.9930$. If $K(P,Q) = 1$ then it translates that proteins P and Q are very similar to each other. Further, if there are n training data then it will give $n \times n$ kernel matrix K which will be processed through SVM classifier for its parameter estimation.

Results and discussions

We carried out experiments on 3 benchmark datasets: DD, TG and EDD, to show the effectiveness of our proposed feature extraction method. We employ SVM classifier from libsvm (Chang and Lin, 2011) 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 hyper-plane to minimize classification error. For the SVM classifier, kernel K is used. The kernel and SVM parameters, gamma and C, are optimized using grid search. 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). However, to reduce the computational time, we adopted the k-fold cross-validation in this study as done by many investigators with SVM as the prediction engine. We use datasets to perform k-fold cross-validation for $k = 5,6,7,8,9$ and 10. For statistical stability we performed 50 times k-fold cross-validation in this paper.

The proposed feature extraction method has been compared with several other feature extraction methods and the results have been shown in Tables 2, 3 and 4. 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, AAC+HXPZV (Ding and Dubchak, 2001), ACC (Dong et al., 2009), mono-gram and bi-gram (Sharma et al., 2013b). 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 Tables 2-4,

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 2 (on DD dataset) that the highest accuracy of protein fold recognition is 74.7% which is obtained by alignment method (when $k = 9$ and $k = 10$) followed by bi-gram method which is 74.1% (when $k = 10$). Besides the enhancement achieved compared to bi-gram and mono-gram methods that we have recently proposed in our previous study, we achieved an improvement of 7% prediction accuracy compared to ACC method (which has been proposed by Dong et al., (2009) and remained unbeaten ever since). In general, the protein fold prediction accuracy by alignment method is around 0.6% to 29% higher than other methods.

Table 2: Recognition accuracy by k-fold cross validation procedure for various feature extraction techniques using SVM classifier on DD-dataset.

Feature sets	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
PF1 (Ghanty and Pal, 2009)	48.6	49.1	49.5	50.1	50.5	50.6
PF2 (Ghanty and Pal, 2009)	46.3	47.0	47.5	47.7	47.9	48.2
PF (Yang et al., 2011)	51.2	52.2	52.6	52.9	53.4	53.4
O (Taguchi and Gromihan, 2007)	49.7	50.4	50.8	50.8	51.1	51.0
AAC (Ding and Duchak, 2001)	43.6	43.9	44.2	44.8	44.6	45.1
AAC+HXPZV+ (Ding and Dubchak, 2001)	45.1	46.2	46.5	46.8	46.9	47.2
ACC (Dong et al., 2009)	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
Mono-gram (Sharma et al., 2013b)	67.7	68.4	68.6	69.1	69.4	69.6
Bi-gram (Sharma et al., 2013b)	72.6	73.1	73.7	73.7	74.1	74.1
Alignment method (this paper)	**72.6**	**73.5**	**73.8**	**74.2**	**74.7**	**74.7**

Table 3 shows accuracy on TG dataset. It can be observed from the table that the highest accuracy of protein fold recognition is by alignment method. For the first time, we have enhanced the prediction accuracy to over 70% when the sequential similarity is less than 25%. We report 74.0% (when $k = 10$) prediction accuracy for TG benchmark followed by bi-gram method which is 68.1% (Sharma et al., 2013b). In general, the accuracy is around 5.9% to 40.5% higher than other feature extraction methods.

Table 3: Recognition accuracy (in percentage) by k-fold cross validation procedure for various feature extraction techniques using SVM classifier on TG dataset.

Feature sets	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
PF1 (Ghanty and Pal, 2009)	38.1	38.4	38.6	38.7	38.8	38.8
PF2(Ghanty and Pal, 2009)	38.0	38.4	38.5	38.6	38.7	38.8
PF (Yang et al., 2011)	42.3	42.6	42.7	43.0	43.0	43.1
O (Taguchi and Gromihan, 2007)	35.8	36.1	36.2	36.1	36.3	36.3
AAC (Ding and Duchak, 2001)	31.5	31.5	31.7	31.8	31.9	32.0
AAC+HXPZV (Ding and Duchak, 2001)	35.7	36.0	36.1	36.2	36.3	36.3
ACC (Dong et al., 2009)	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
Mono-gram (Sharma et al., 2013b)	49.3	49.5	49.7	49.9	50.0	50.1
Bi-gram (Sharma et al., 2013b)	67.1	67.5	67.6	67.8	68.1	68.1
Alignment method (this paper)	**72.0**	**72.7**	**73.0**	**73.5**	**73.6**	**74.0**

Next, Table 4 depicts protein fold recognition accuracy on EDD dataset. It can be seen from the table that the highest accuracy is again obtained by alignment method. For the first time, we have enhanced the protein fold prediction accuracy to over 90% when the sequential similarity rate is less than 40%. We report 90.2% (when $k = 10$) prediction accuracy for the EDD benchmark followed by ACC which is 85.9% (Dong et al., 2009). In general, the protein fold prediction enhancement achieved by alignment method compared to previously reported results for the EDD benchmark is from 4.3% to 49.2%.

Table 4: Recognition accuracy (in percentage) by k-fold cross validation procedure for various feature extraction techniques using SVM classifier on EDD dataset.

Feature sets	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
PF1 (Ghanty and Pal, 2009)	50.2	50.5	50.5	50.7	50.8	50.8
PF2 (Ghanty and Pal, 2009)	49.3	49.5	49.7	49.8	49.8	49.9
PF (Yang et al., 2011)	54.7	55.0	55.2	55.4	55.5	55.6
O (Taguchi and Gromihan, 2007)	46.4	46.6	46.6	46.7	46.7	46.9
AAC (Ding and Duchak, 2001)	40.3	40.6	40.7	40.7	40.9	40.9
AAC+HXPZV (Ding and Duchak, 2001)	40.2	40.4	40.6	40.7	40.9	40.9
ACC (Dong et al., 2009)	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
Mono-gram (Sharma et al., 2013b)	62.7	63.0	63.3	63.3	63.4	63.6
Bi-gram (Sharma et al., 2013b)	83.6	84.0	84.1	84.3	84.3	84.5
Alignment method (this paper)	**89.4**	**89.7**	**89.9**	**90.0**	**90.1**	**90.2**

In order to study the statistical significance of the prediction accuracy enhancement

reported in this study, we conduct the paired t-test on our achieved results compared to the highest results reported in the literature. Associated probability value achieved for the paired t-test is $p = 0.03$ which confirms the statistical significance of our reported enhancement in this study compared to the state-of-the-art results found in the literature for protein fold recognition.

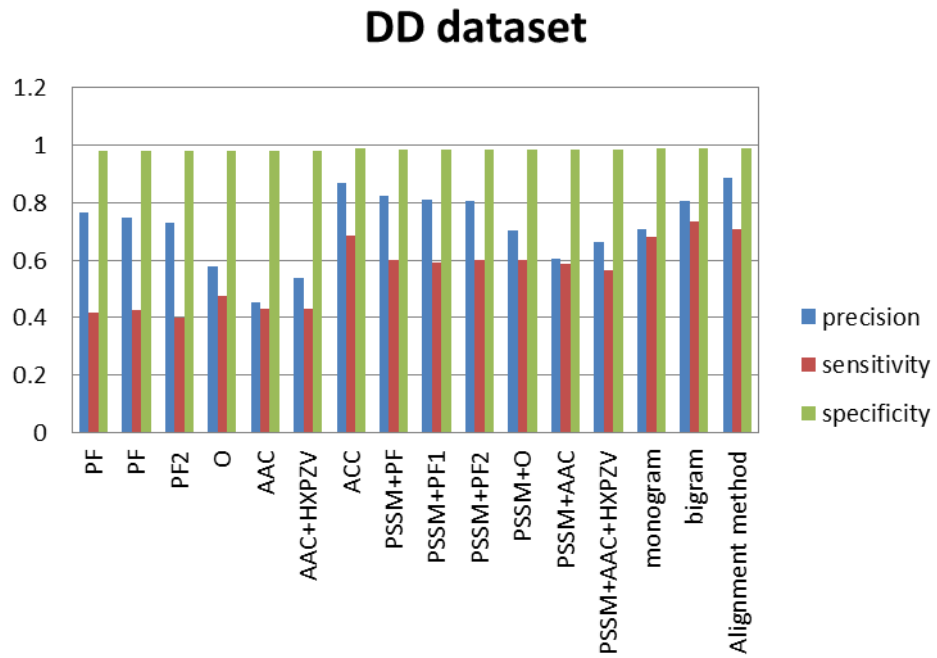

Figure 3: Precision, sensitivity and specificity of all feature sets on DD dataset.

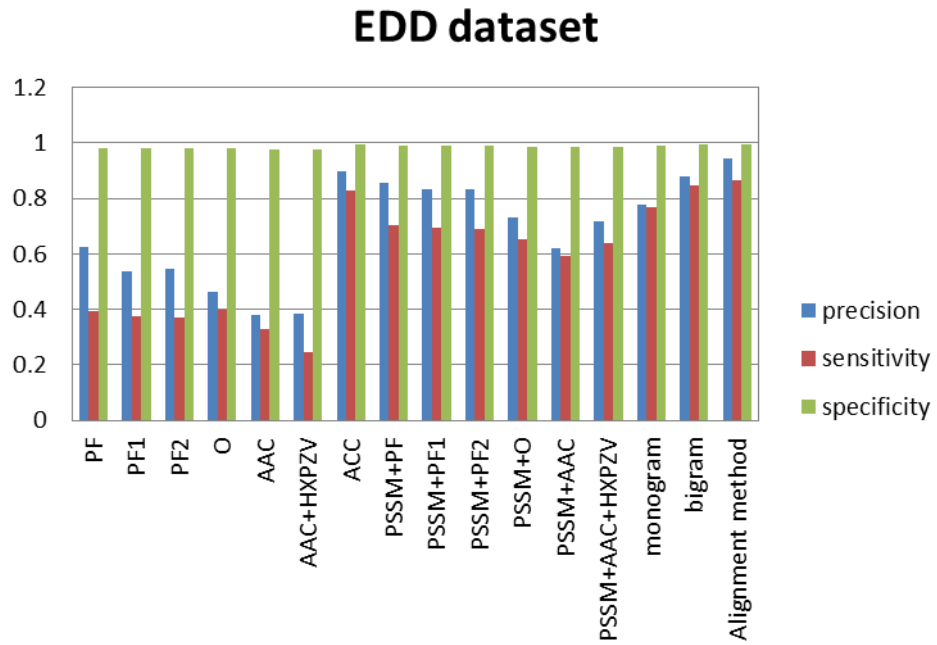

Figure 4: Precision, sensitivity and specificity of all feature sets on EDD dataset.

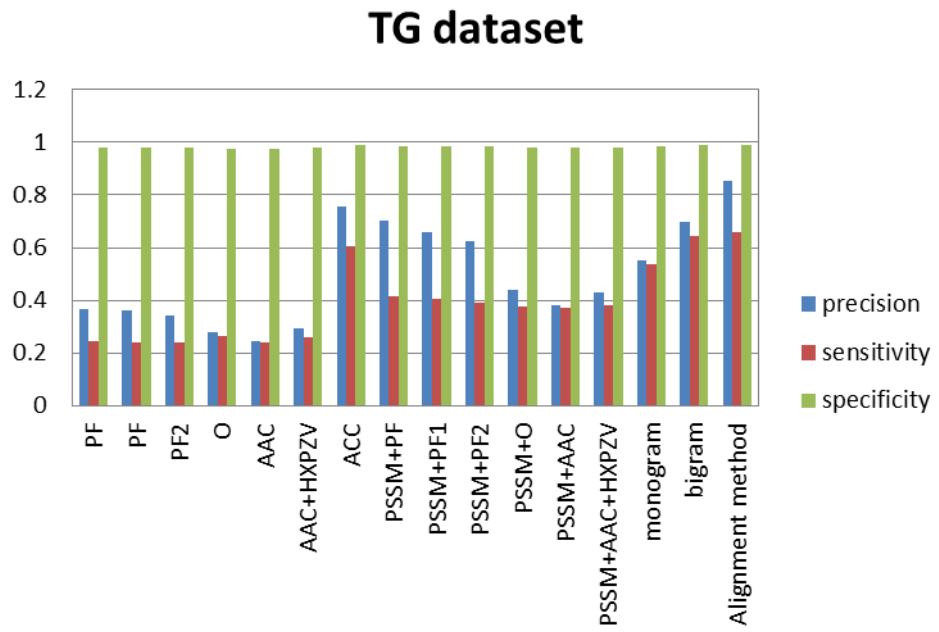

Figure 5: Precision, sensitivity and specificity of all feature sets on TG dataset.

Furthermore, we have conducted precision, sensitivity and specificity analysis of all the features used in this paper over 3 datasets. Figure 3, depicts the analysis on DD dataset, Figure 4 on EDD dataset and Figure 5 on TG dataset. It can be observed from Figures 3-5 that specificity is high for all the feature sets. However, precision and sensitivity varies. For all the datasets, precision and sensitivity are quite promising for alignment method.

Since it is very useful to have accessible codes for developing practically more useful models, we have provided Matlab based code for our method. <Link will be provided upon acceptance of this paper>

Conclusion

In this work, we developed feature extraction method using amino acid alignment scheme. The technique used PSSM log probabilities of protein sequences, to determine the distance between two proteins. This method has been compared with several other existing feature extraction methods and very promising results have been obtained. It was noted that the proposed method outperformed existing methods for three commonly used benchmarks. We have reported 74.6% prediction accuracy on DD benchmark. For the first time, we have also achieved to over 70% and 90% prediction accuracies for protein fold recognition when the sequential similar rates are less than 25% and 40%, respectively. We observed 74.0% and 90.2% prediction accuracies for TG and EDD

benchmarks, respectively. These reported results are over 5.9% and 4.3% better than the best results reported for these two benchmark datasets.

Reference

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

Bishop, C.M., Pattern recognition and machine learning, Springer Science, NY, 2006.

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

Chang, C.-C., Lin, C.-J., LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, vol. 2, issue 3, pp. 27:1-27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm

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

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

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

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

Chou, K.C., Shen, H.B., 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, 2010

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

Dehzangi A, Amnuaisuk S.P., Dehzangi O: Enhancing protein fold prediction accuracy by using ensemble of different classifiers. Australian Journal of Intelligent Information Processing Systems, vol. 26, no. 4, pp.32-40, 2010.

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

Dehzangi, A., Karamizadeh, Solving protein fold prediction problem using fusion of heterogeneous classifiers, Information an International Interdisciplinary Journal, vol. 14, no. 11, pp. 3611-3622, 2011.

Dehzangi, A., Paliwal, K.K., Sharma, A., Dehzangi, O., Sattar, A., A Combination of Feature Extraction Methods with an Ensemble of Different Classifiers for Protein Structural Class Prediction Problem. IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM, 2013a (In Press)

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A.: Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy. In: Proceeding of the Pattern Recognition in Bioinformatics. PRIB 2013, LNBI 7986, pp. 208–219, 2013b

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A.: Enhancing protein fold prediction accuracy using evolutionary and structural features. In: Proceeding of the Pattern Recognition in Bioinformatics. PRIB 2013, LNBI 7986, pp. 196–207, 2013c

Deschavanne P, Tuffery P: Enhanced protein fold recognition using a structural alphabet. Proteins: Structure, Function, and Bioinformatics, 76, pp. 129-137, 2009.

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

Ding, Y.S., Zhang, T.L., Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier, Patt. Recog. Letters, 29, pp. 1887-1892, 2008.

Dong, Q., Zhou, S., Guan, J., A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinformatics, vol. 25, no. 20, pp. 2655-2662, 2009.

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

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

Huang, J.T. and Tian, J., Amino acid sequence predicts folding rate for middle-size two-state proteins, Proteins: Structure, Function, and Bioinformatics, 63(3), pp. 551-554, 2006.

Kavousi K, Moshiri B, Sadeghi M, Araabi BN, Moosavi-Movahedi AA: A protein fold classier formed by fusing different modes of pseudo amino acid composition via PSSM. Computational Biology andChemistry, 35(1), pp. 1-9, 2011.

Kecman, V., Yang, T., Protein fold recognition with adaptive local hyper plane Algorithm, Computational Intelligence in Bioinformatics and Computational Biology, CIBCB '09. IEEE Symposium, pp. 75-78, 2009.

Klein, P., Prediction of protein structural class by discriminant analysis, BiochimBiopjysActa, 874, pp. 205-215, 1986.

Krishnaraj, Y., Reddy, C.K., Boosting methods for protein fold recognition: an empirical comparison, IEEE Int. Conf. on Bioinfor. and Biomed., pp. 393-396, 2008.

Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J., Secondary structure-based

assignment of the protein structural classes, Amino Acids, 35, pp. 551-564, 2008.

Liu T., Geng X., Zheng X., Li R., Wang J., Accurate Prediction of Protein Structural Class Using AutoCovariance Transformation of PSI-BLAST Profiles, Amino Acids, 42, pp. 2243-2249, 2012.

Najmanovich, R., Kuttner, J., Sobolev, V., Edelman, M., Side-chain flexibility in proteins upon ligand binding, Proteins: Structure, Function, and Bioinformatics, 39(3), pp. 261-268, 2000.

Ohlson, T., Wallner, B., Elofsson, A., "Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods", Proteins: Structure, Function, and Bioinformatics, 57, pp. 188-197, 2004.

Sharma, A., Paliwal, K.K., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., A Strategy to Select Suitable Physicochemical Attributes of Amino Acids for Protein Fold Recognition, BMC Bioinformatics, 2013a (accepted).

Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K., "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition", Journal of Theoretical Biology, vol. 320, no. 7, pp. 41-46, 2013b.

Sharma, A., Imoto, S., Miyano, S., "A top-r feature selection algorithm for microarray gene expression data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, 2012, pp. 754-764.

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

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

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

Vapnik, V.N., The nature of statistical learning theory, Springer-Verlag, New York, 1995.

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

Yang, T. Kecman, V., Cao, L., Zhang, C., Huang, J.Z., Margin-based ensemble classifier for protein fold recognition, Expert Systems with Applications, 38, pp. 12348-12355, 2011.

Ying, Y., Huang, K., Campbell, C., Enhanced protein fold recognition through a novel data integration approach, BMC Bioinformatics, vol. 10, no. 1, 267, 2009.

Valavanis, I.K., Spyrou, G.M., Nikita, K.S., A comparative study of multi-classification methods for protein fold recognition, Int. J. Comput.Intelligence in Bioinformatics and Systems Biology, 1(3), pp. 332-346, 2010.

Zhang, H., Zhang, T., Gao, J., Ruan, J., Shen, S., Kurgan, L.A., Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent

accessibility, Amino Acids, pp. 1-13, 2010.

Zhang, T.L., Ding, Y.S., Chou, K.C., Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern, Theoretical Biology, 250, pp. 186-193, 2008.