Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids

Article *in* Journal of Theoretical Biology · May 2016

Impact Factor: 2.12 · DOI: 10.1016/j.jtbi.2016.05.002

5 authors, including:

Harsh Saini

University of the South Pacific

5 PUBLICATIONS **10** CITATIONS

SEE PROFILE

Iman (Abdollah) Dehzangi

University of Iowa

40 PUBLICATIONS **311** CITATIONS

SEE PROFILE

Alok Sharma

Jubilant Life Sciences

123 PUBLICATIONS **992** CITATIONS

SEE PROFILE

Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids

Gaurav Raicar, Harsh Saini, Abdollah Dehzangi, Sunil Lal, Alok Sharma

www.elsevier.com/locate/yjtbi

PII: S0022-5193(16)30074-1
DOI: http://dx.doi.org/10.1016/j.jtbi.2016.05.002
Reference: YJTBI8652

To appear in: *Journal of Theoretical Biology*

Received date: 21 January 2016
Revised date: 20 April 2016
Accepted date: 2 May 2016

Cite this article as: Gaurav Raicar, Harsh Saini, Abdollah Dehzangi, Sunil La and Alok Sharma, Improving protein fold recognition and structural clas prediction accuracies using physicochemical properties of amino acids, *Journa of Theoretical Biology,* http://dx.doi.org/10.1016/j.jtbi.2016.05.002

Title of paper:

Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids

Authors:

1. Mr. Gaurav Raicar (corresponding author)

Affiliation: The University of the South Pacific, Fiji Islands

Email: raicar_g@usp.ac.fj

2. Mr. Harsh Saini

Affiliation: The University of the South Pacific, Fiji Islands

Email: saini_h@usp.ac.fj

3. Dr. Abdollah Dehzangi

Affiliation: University of Iowa, USA

Email: i.dehzangi@gmail.com

4. Dr. Sunil Lal

Affiliation: Massey University, New Zealand

Email: s.lal@massey.ac.nz

5. Dr. Alok Sharma

Affiliations:

a) The University of the South Pacific, Fiji Islands

b) IIIS, Griffith University, Australia

c) RIKEN, Japan

Email: sharma_al@usp.ac.fj

Abstract:

Predicting the three-dimensional (3-D) structure of a protein is an important task in the field of bioinformatics and biological sciences. However, directly predicting the 3-D structure from the primary structure is hard to achieve. Therefore, predicting the fold or structural class of a protein sequence is generally used as an intermediate step in determining the protein's 3-D structure. For protein fold recognition (PFR) and structural class prediction (SCP), two steps are required – feature extraction step and classification step. Feature extraction techniques generally utilize syntactical-based information, evolutionary-based information and physicochemical-based information to extract features. In this study, we explore the importance of utilizing the physicochemical properties of amino acids for improving PFR and SCP accuracies. For this, we propose a Forward Consecutive Search (FCS) scheme which aims to strategically select physicochemical attributes that will supplement the existing feature extraction techniques for PFR and SCP. An exhaustive search is conducted on all the existing 544 physicochemical attributes using the proposed FCS scheme and a subset of physicochemical attributes is identified. Features extracted from these selected attributes are then combined with existing syntactical-based and evolutionary-based features, to show an improvement in the recognition and prediction performance on benchmark datasets.

Keywords:

Protein fold recognition, structural class prediction, physicochemical properties, syntactical-based features, evolutionary-based features, forward consecutive search scheme

Introduction

In the field of bioinformatics and biological sciences, predicting the three-dimensional (3-D) structure of a protein plays a crucial role. The functions of protein, being closely linked to its structure enable us to further understand the cellular functions, protein-protein interactions

and aids the development of new drug designs and therapies (Chmielnicki and Stapor, 2012). The multitude of protein sequences generated due to large-scale sequencing projects are significantly higher than the known 3-D protein structure. Computational techniques have to be employed to determine the structure of a protein quickly and efficiently.

Directly predicting the protein 3-D structure from its sequence is hard to achieve. However, classifying protein sequences to their fold or structural class is a transitional stage in determining the 3-D structure of a protein. In order to determine the fold or structural class of a protein sequence, two steps are required: 1) feature extraction step and 2) classification step. In feature extraction step, informative features are extracted from primary protein sequences. These features are further used in the classification step for protein fold recognition (PFR) and structural class prediction (SCP). If the extracted features are well discriminative, it can help improving the recognition and prediction rate. This makes feature extraction a crucial step in the overall procedure (Dehzangi et al., 2013a, 2013b, 2013c, 2013d, 2014a, 2014c; Deschavanne and Tuffery, 2009; Dong et al., 2009; Kavousi et al., 2011; Lyons et al., 2014, 2015, 2016; Paliwal et al., 2014b; Sharma et al., 2013a, 2014; Saini et al., 2014, 2015).

A lot of research has been done in the domain of protein SCP (Chou and Zhang, 1994; Chou, 1995; Bahar et al., 1997; Zhou, 1998; Chou and Maggiora, 1998; Zhou and Assa-Munt, 2001; Heffernan et al., 2015a, 2015b). One of the important progresses made in this domain was a study conducted by Chou and Cai (2004). They proposed a scheme whereby the feature vector of a protein sample was represented by its functional domain composition to formulate the predictor. The validation was made on a very stringent benchmark dataset which covers the following 7 classes: (i) all-alpha, (ii) all-beta, (iii) alpha/beta, (iv) alpha+beta, (v) multi-domain, (vi) small protein, and (vii) peptide. The cutoff threshold was 20%, meaning that none of proteins included in the benchmark dataset has greater than 20% pairwise sequence identity to any other in a same subset. For such an extremely stringent benchmark dataset, the overall jackknife success rate by the "Functional Domain Composition" method was over 90%. The pseudo amino acid composition approach has also been widely used by many investigators (Chen et al., 2006b, 2012; Sahu and Panda, 2010; Zhang et al., 2014; Qin, 2012) for predicting protein structural classes.

In the literature, many feature extraction techniques have been developed and used for PFR and SCP. Features are generally extracted by utilizing syntactical-based, evolutionary-based and physicochemical-based information. Features which are dependent on physicochemical attributes can reveal global properties of proteins (Bulashevska and Eils, 2006; Chinnasamy

et al., 2005). These features are able to maintain high discriminatory information even when the sequence similarity is low (Dubchak et al., 1997; Pal and Chakraborty, 2003). Therefore, physicochemical-based features could be a viable option when compared to syntactical and evolutionary based features, since the latter would perform weakly on datasets with low sequence similarity. This phenomenon is commonly known as the twilight zone (Kurgan and Homaeian, 2005; Mizianty and Kurgan, 2009).

Many researchers have proposed features based on syntactical, evolutionary and physicochemical information. Here we cover some of the important work. Dubchak et al. (1997) suggested syntactical-based and physicochemical-based features. The physicochemical-based features were extracted from five attributes – hydrophobicity (H), predicted secondary structure based on normalized frequency of α-helix (X), polarity (P), polarizability (Z) and van der Waals volume (V). Features extracted from these attributes have been used extensively in PFR and SCP problems (Ding and Dubchak, 2001; Krishnaraj and Reddy, 2008). Other physicochemical attributes that have been used, include flexibility (Najmanovich et al., 2000), bulkiness (Huang and Tian, 2006) and solvent accessibility (Zhang et al., 2012). Dehzangi and Phon-Amnuaisuk (2011) explored four new physicochemical attributes in addition to the five attributes used by Ding and Dubchak (2001). Sharma et al. (2013b) proposed a strategic selection scheme to identify suitable physicochemical attributes out of a subset of 30 attributes. The selected attributes showed an improvement in recognition performance. Dehzangi et al. (2014b) explored the impact of 55 different physicochemical attributes for PFR.

On the other hand, Taguchi and Gromiha (2007) have argued that only syntactical-based features should be considered, since features extracted from physicochemical attributes have no significant information. This contradiction depicts that further research is required in order to fully explore the potential of physicochemical attributes (Sharma et al., 2013b). In this study, we explore the impact of strategically selecting physicochemical attributes to supplement the existing feature extraction techniques for PFR and SCP. For this, we propose a Forward Consecutive Search (FCS) scheme which is based on a greedy search algorithm (Guyon and Elisseeff, 2003; Cormen et al., 1990, Sharma et al., 2012a). The FCS scheme will be used to thoroughly explore all the 544 physicochemical attributes (Kawashima et al., 2008) (a full list of attributes can be found at link – http://www.genome.jp/aaindex/) and identify a subset of suitable physicochemical attributes that will supplement the existing feature extraction techniques for PFR and SCP.

This scheme is used on the Ding and Dubchak (DD) dataset (2001), Taguchi and Gromiha (TG) dataset (2007) and extended Ding and Dubchak (EDD) dataset (Dong et al., 2009). For each syntactical-based and evolutionary-based feature, a subset of the best physicochemical attributes are selected. An improvement in PFR and SCP is noted after appending the physicochemical-based features corresponding to these selected attributes. The improvements in all three datasets using 10-fold cross-validation ranged from 0.5% - 28.3% for PFR and 0.5% - 18.6% for SCP.

Related Work

Apart from features extracted from physicochemical attributes, syntactical-based and evolutionary-based features have also been widely used for PFR and SCP. Taguchi and Gromiha (2007) have proposed features that are based on the amino acid occurrence. To rely more on the order of the amino acids in the protein sequence, Chou (2001) proposed Pseudo amino acid composition (PseAAC) based features. Similarly, Huang et al. (2003) used bigram features based on the order of the amino acid in the protein sequence. Later on, based on a similar concept, Ghanty and Pal (2009) employed pairwise frequency of amino acids that were separated by one residue (PF1) and pairwise frequency of adjacent residues (PF2). These pairwise frequency features contained 400 features each. In a study conducted by Yang et al. (2011), these pairwise frequency features were concatenated to produce 800 features. Shamim et al. (2007) proposed features that are extracted from the structural information of amino acid residues and pairs. If after concatenation of features, the dimensionality of the features is too high, then this dimensionality can be controlled by selecting a subset of important features only (Sharma et al., 2006, 2011, 2012b, 2012c, 2012d, 2013c; Sharma and Paliwal, 2007, 2010, 2012a, 2012b, 2012c, 2015a and 2015b). Liu et al. (2012) and Kurgan et al. (2008) have shown autocorrelation features for protein sequences.

Evolutionary-based features have also gained popularity, e.g. feature based on Position Specific Scoring Matrix (PSSM). PSSM is a representation of a protein sequence which defines the probability of any given amino acid occurring at a particular position in the sequence. Sharma et al. (2013a) employed monogram and bigram features extracted from the PSSM directly, thus solving the problem of having zero components in the feature vector as compared to the bigram feature vector suggested by Ghanty and Pal (2009). Similarly, Saini et al. (2015) proposed features that consist of probabilistic expressions of amino acid dimers that are spatially varied. Efforts have also been made to extract features such as monogram and bigram occurrence on the consensus protein sequence instead of the raw protein sequence (Paliwal et al., 2014a). Lyons et al. (2014) compared their proposed

feature extraction technique with features such as AAC + HXPZV and PSSM + AAC + HXPZV, where HXPZV were attributes used in a study by Dubchak et al. (1997).

After feature extraction step, the classification step is applied to predict the folds or structural classes of the protein sequences using the extracted features. Many classifiers have been developed and used for PFR and SCP. Classifiers such as K-Nearest Neighbor (KNN) (Ding and Zhang, 2008, Nanni, 2006), Bayesian classifiers (Chinnasamy et al., 2005), Artificial Neural Network (ANN) (Cai and Zhou, 2000, Bologna and Appel, 2002), Ensemble classifiers (Shen and Chou, 2006; Chmielnicki and Stąpor, 2011), Support Vector Machine (SVM) (Chen et al., 2006a, Zhou et al., 2008) and hierarchical classification (Lin et al., 2013, Sharma et al., 2016) have been implemented and used most commonly.

As demonstrated by a series of recent publications (Jia et al., 2016a, 2016b, 2016c; Liu, B. et al., 2016a, 2016b; Liu, Z. et al., 2016; Chen et al., 2016) in compliance with Chou's 5-step rule (Chou, 2011), to establish a really useful sequence-based statistical predictor for a biological system, we should follow the following five guidelines: (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; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we describe how to deal with these steps one-by-one.

Dataset

Three datasets were employed in this study are DD dataset (Ding and Dubchak, 2001), TG dataset (Taguchi and Gromiha, 2007) and EDD dataset (Dong et al., 2009). The benchmarked DD dataset consists of 311 protein sequences in the training set where the sequence identity between any given pair of protein sequences is no more than 35% for aligned subsequences longer than 80 residues. The test set consists of 383 protein sequences where the sequence identity is less than 40%. Both the training and test sets belong to 27 SCOP folds which represent all the major structural classes: $\alpha,\ \beta,\ \alpha+\beta,\ \alpha/\beta$ (Murzin et al., 1995).

The EDD dataset consists of 3418 protein sequences which have less than 40% sequence similarity and belong to the same 27 SCOP folds that were used in the DD dataset. The dataset is extracted from 1.75 SCOP in a similar fashion to Dong et al. (2009).

The TG dataset consists of 1612 protein sequences which belong to 30 different folding types of globular proteins (Taguchi and Gromiha, 2007) and have less than 25% sequence similarity between them.

Since there are no pre-defined training set and test set for the EDD and TG datasets, the latter datasets were partitioned into training and test sets at approximately 3:2 ratios, respectively. After partitioning, the EDD training set contains 2082 protein sequences and test set contains 1336 protein sequences while the TG training set contains 1010 protein sequences and the test set contains 602 protein sequences. The distribution of the protein folding classes in the training and test set were kept proportionally equal. The training set was used to find physicochemical attributes.

Features

In this study, we aim to show that PFR and SCP performance can be improved by utilizing information present in the physicochemical properties of amino acids when used in conjunction with syntactical-based and evolutionary-based features. Although Taguchi and Gromiha (2007) have argued that features extracted from physicochemical attributes contain no significant information, we have shown in this study that features extracted from selected physicochemical attributes have led to significant improvements in PFR and SCP when used with existing syntactical-based and evolutionary-based features found in the literature. The following features are explored in this study and are discussed below.

Syntactical-based and evolutionary-based features

1. *Occurrence (O),* is the frequency of amino acids (there are 20 unique amino acids) in a protein sequence, thus producing 20 features (Taguchi and Gromiha, 2007).
2. *Pairwise frequency (PF1)* of amino acids separated by one residue, is the frequency of pairs of amino acids in a protein sequence, thus producing 400 features (Ghanty and Pal, 2009).
3. *Bigram* feature represents the transitional probabilities from one amino acid to another and is based on PSSM, producing 400 features (Sharma et al., 2013a).
4. *Separated dimers* consists of the probabilistic expressions of amino acid dimers that have spatial separations from $k = 1, 2, .., K$, where K denotes the upper bound of k, and produces 400× k features (Saini et al., 2015).

A prefix of *PSSM+* before a syntactical/evolutionary-based feature name (O, PF1) indicates that the feature was extracted from consensus protein sequences rather than the raw protein sequences. The consensus protein sequences are derived by replacing each amino acid from the raw protein sequence with the amino acid having the highest probability in the PSSM.

Physicochemical-based features

In addition to the above syntactical-based and evolutionary-based features, we also extract a set of physicochemical-based features. However, extracting physicochemical-based features directly from attributes has some drawbacks. As mentioned in the introduction, while physicochemical-based features are able to maintain high discriminatory information based on amino acid residues, they do not incorporate information regarding the positioning of amino acids.

Therefore, in order to solve this problem, we first find the probabilistic expression of the residues of the physicochemical attributes by combining them with PSSM probabilities. This way we can incorporate information based on amino acid positions while extracting physicochemical-based features. Therefore, evolutionary-based information is joined with physicochemical-based information. This would provide more discriminatory information which would help in improving PFR and SCP.

The PSSM probabilities P of a protein sequence is a matrix of size $L \times 20$, where L is the length of the protein sequence and the residues R for a physicochemical attribute j is a vector of size 20×1. Thus, the probabilistic expression f of the residues of a physicochemical attribute j can be calculated by finding the product of P and R as shown by equation (1) below:

$$f = PR \tag{1}$$

It should be noted that the order of the amino acids (e.g. A R N…) in matrix P and vector R must be the same. After this step, we extract the physicochemical-based features by using an autocorrelation of the probabilistic residue values (f) of the protein sequences. This can be illustrated mathematically by equation (2) below:

$$A_i = \frac{1}{L}\sum_{m=1}^{L-i}(f_m - \mu)\,(f_{m+i} - \mu) \quad (2)$$

Where f_m is the probabilistic residue value of the m^{th} amino acid in a protein sequence and μ is the average of L probabilistic residues. In this study, we use $i = 1, 2, 3, \ldots, 20$, thus producing 20 autocorrelation features for each protein sequence corresponding to a given physicochemical attribute j.

Methodology

In order to show improvements in PFR and SCP by utilizing physicochemical-based features, we first establish a baseline by noting the n-fold cross-validation accuracy of the syntactical-based and evolutionary-based features (O, PF1, PSSM+O, PSSM+PF1, Bigram and Separated dimers), on all three datasets - DD, TG and EDD.

After establishing the baseline, we added the physicochemical-based features successively to the syntactical-based and evolutionary-based features by using a simple scheme that is based on an exhaustive greedy search algorithm. The aim of this scheme is to identify a subset of physicochemical attributes and explore the performance of features extracted from these selected attributes in combination with the syntactical-based and evolutionary-based features. This scheme will be referred to as the *Forward Consecutive Search* scheme and is illustrated in Figure 1.

Forward Consecutive Search Scheme

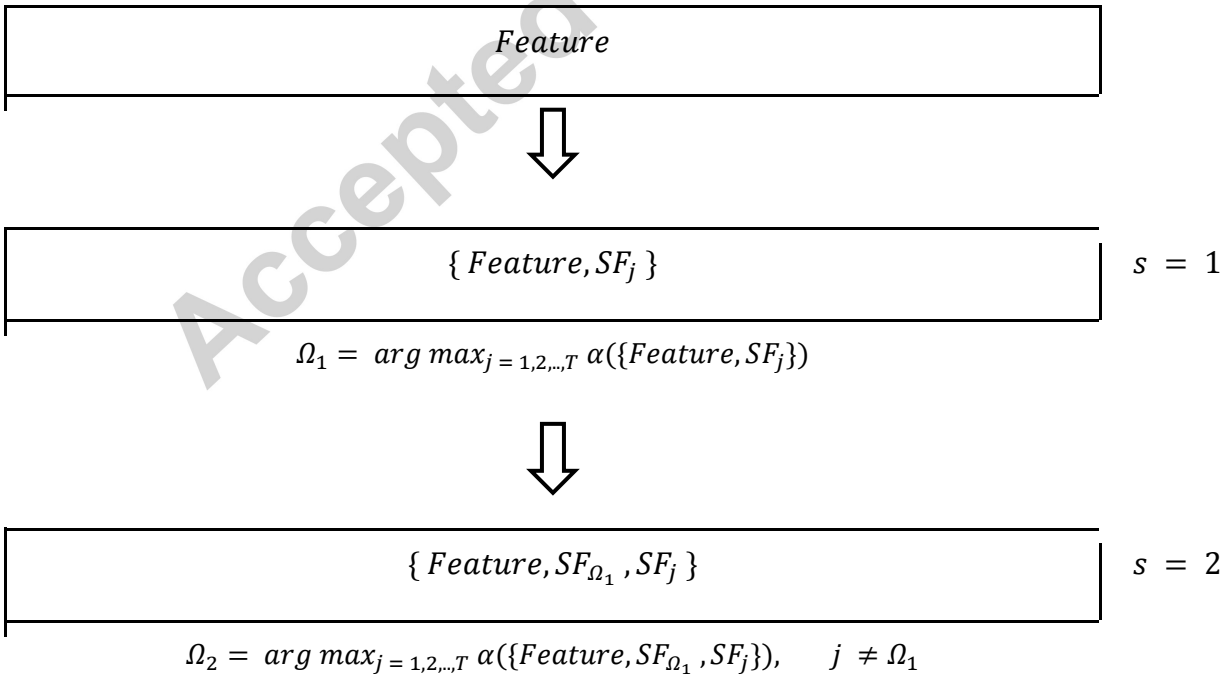
In Figure 1, any syntactical/evolutionary-based feature is the input denoted as *Feature* in the figure, *Successive Feature* (SF_j) represents the 20 autocorrelation features ($A_1, A_2, A_3, \ldots, A_{20}$) extracted from a physicochemical attribute j using equation (2), α represents the n-fold cross-validation accuracy on a combination of features {*Feature*, SF}, T represents the maximum number of physicochemical attributes used, S represents the total number of levels used in the scheme and Ω_s represents the physicochemical attribute selected at a given level s. The FCS scheme was employed only on the training sets of the

datasets. Thus, the test samples were kept isolated while the search was being conducted to identify the subset of physicochemical attributes.

In the FCS scheme, a physicochemical-based feature is taken at a time in combination with a syntactical/evolutionary-based feature and the average classification accuracy is computed on the combination of features $\{ Feature,\ SF_j \}$ using n-fold cross-validation process on the training set of the data. The attribute j corresponding to the physicochemical-based feature (SF_j) that exhibits the highest classification accuracy in combination with $Feature$ is retained and progresses to the next level; i.e., $\Omega_1 = \ arg\ max_{j\ =\ 1,2,..,T}\ \alpha(\{Feature,\ SF_j\})$.

In the next level, another physicochemical-based feature is taken at a time (where $j \neq \Omega_1$) in combination with the previous set of features $\{ Feature, SF_{\Omega_1}, SF_j \}$ and the average classification accuracy is computed. The attribute j corresponding to the physicochemical-based feature that exhibits the highest classification accuracy in combination with the previous set of features is retained and progresses to the next level. This process continues until all the physicochemical attributes have been ranked. The number of attributes ranked is however dependent on the number of levels being used in the scheme.

Figure 1 - Forward Consecutive Search (FCS) Scheme



$$\{\, Feature, SF_{\Omega_1}, SF_{\Omega_2}, \ldots, SF_j \,\} \qquad s = S$$

$$\Omega_s = \; arg\, max_{j\,=\,1,2,..,T}\; \alpha(\{Feature, SF_{\Omega_1}, SF_{\Omega_2}, \ldots, SF_j\}), \quad j \neq \Omega_1, \Omega_2, .., \Omega_{s-1}$$

Results

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 test, and jackknife test (Chou and Zhang, 1995). 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 as elaborated in (Chou, 2011) and demonstrated by equations 28-30 therein. Accordingly, the jackknife test has been increasingly recognized and widely used by investigators to examine the quality of various predictors (Chou and Cai, 2005; Shen and Chou, 2007; Nanni et al., 2014; Mondal and Pai, 2014; Ali and Hayat, 2015; Kumar et al., 2015; Dehzangi et al., 2015; Chen, 2015). However, to reduce the computational time, we adopted the 10-fold cross-validation in this study as done by many investigators.

In this study, we have used LibSVM (Chang and Lin, 2011) version 3.17 with *Radial Basis Function* as the kernel function. The C parameter was set to 1000 and all other parameters were left to its default values. In the literature, only a limited number of physicochemical attributes were explored for PFR and SCP. In this study, all the 544 physicochemical attributes ($T = 544$) are considered. The FCS scheme was run for five levels ($S = 5$), and for each level, 10-fold cross-validation was used to identify the best physicochemical attribute for that level. At the end of the five levels, a subset of the five best attributes (Ω_1, Ω_2, Ω_3, Ω_4, Ω_5) were identified for each syntactical/evolutionary-based feature.

To show the improvements in PFR and SCP performance, 10-fold cross-validation accuracies of all the syntactical-based and evolutionary-based features explored in this study without adding the physicochemical-based features are noted as a baseline. These accuracies are then compared against the 10-fold cross-validation of the combined features $\{Feature, SF_{\Omega_1}, SF_{\Omega_2}, SF_{\Omega_3}, SF_{\Omega_4}, SF_{\Omega_5}\}$.

The improvements in PFR and SCP using the combined features on all three datasets, using 10-fold cross-validation are highlighted in Tables 1 - 3. For PFR, the improvements are as follows: 0.6% - 14.6% on the DD dataset, 1% - 15% on the TG dataset and 0.5% - 28.3% on the EDD dataset. For SCP, the improvements are noted as follows: 1.1% - 13% on the DD dataset, 1.6% - 17.7% on the TG dataset and 0.5% - 18.6% on the EDD dataset.

Discussion

From Tables 1 – 3, it is evident that physicochemical-based features together with syntactical-based or evolutionary-based features improves the performance. For completeness, the increment in accuracy when adding physicochemical-based features to the syntactical-based features and evolutionary-based features over the five levels used in this study has been provided in Tables 4 – 6. Though, the improvement of adding physicochemical-based features to evolutionary-based features is less compared to the improvement with syntactical-based features, improvement is still observed when a subset of attributes was systematically selected in both the cases.

We have also provided box plots which show the accuracies over twenty runs for syntactical-based and evolutionary-based features and the improved accuracies after adding physicochemical-based features to the syntactical-based features and evolutionary-based features. Figures 2 - 7 illustrates the box plots for PFR and Figures 8 – 13 illustrates the box plots for SCP respectively.

Furthermore, we have also conducted paired t-test with 5% significance level to show the statistical significance of improvements seen in this study. We compared the results (in terms of 10-fold cross-validation accuracies) of the structural-based features and evolutionary-based features for PFR and SCP with the results of the structural-based features and evolutionary-based features with physicochemical-based features added to them for PFR and SCP (the degree of freedom is 5). The paired t-test results for features PF1, PSSM + PF1, O, PSSM + O, Bigram and Separated Dimers are 0.001, 0.002, 0.0003, 0.0002, 0.01 and 0.001 respectively. These results show that the improvements in accuracies after adding physiochemical-based features is significant.

The FCS scheme used in this study, requires $j(j+1)/2$ search combinations, where $j = 544$. Each level in the scheme produces 20 dimensional features, thus for s levels, $20 \times s$ dimensional features are produced. Since we consider all the 544 physicochemical attributes in this study, the computational resources and time needed to evaluate 10-fold

cross-validation accuracy on the physicochemical-based features extracted in each level increases dramatically as the number of levels s increases (complexity: $O(s!)$). Due to this limitation, in this study we have restricted to only five levels in the FCS scheme.

After evaluating the results in this study, it can be said that information present in physicochemical attributes can play an important role in improving PFR and SCP. The physicochemical attributes selected, mostly varied for different syntactical-based and evolutionary-based features across the three datasets. It was seen that certain physicochemical attributes were selected multiple times on each dataset as well as across multiple datasets. This shows that these physicochemical attributes hold significant information compared to other physicochemical attributes but need to be used with other physicochemical attributes to be effective. However, the performance can be further increased by selecting other physicochemical properties or combination of physicochemical properties using different techniques or schemes (Guyon and Elisseeff, 2003; Cormen et al., 1990) and by also using different classifiers. Table 7 summarizes the rank of attributes based on its frequency of selection over all the three datasets. In addition to this, the attributes that were selected in this study is summarized in the Appendix Section.

As demonstrated in a series of recent publications (Jia et al., 2016a, 2016b, 2016c; Liu B. et al., 2016a, 2016b; Liu Z. et al., 2016; Chen et al., 2016) in developing new prediction methods, user-friendly and publicly accessible web-servers will significantly enhance their impacts (Chou, 2015), and we shall make efforts in our future work to provide a web-server for the prediction method presented in this study.

Conclusion

In this study, we aim to improve PFR and SCP accuracies by utilizing information present in the physicochemical properties of amino acids. For this, we have employed a Forward Consecutive Search (FCS) scheme to select the most suitable physicochemical attributes for improving PFR and SCP. The FCS scheme utilizes all the 544 physicochemical attributes and a subset of the five best physicochemical attributes is selected for each syntactical/evolutionary-based feature.

It was shown that features extracted from these carefully selected physicochemical attributes led to an improvement in PFR and SCP performance when used in combination with the syntactical-based and evolutionary-based features. The improvements using 10-fold cross-validation for PFR, were: 0.6% - 14.6% on the DD dataset, 1% - 15% on the TG dataset and

0.5% - 28.3% on the EDD dataset. For SCP, the improvements were: 1.1% - 13% on the DD dataset, 1.6% - 17.7% on the TG dataset and 0.5% - 18.6% on the EDD dataset.

Table 1 - DD Dataset n-fold cross validation

	Feature	Baseline Accuracy (n = 10)	Improved Accuracy (n = 10)	Rank
PFR	**PF1**	50.6%	62.3%	537, 339, 199, 317, 466
	PSSM + PF1	66.4%	69%	314, 453, 351, 469, 1
	O	51%	65.6%	12, 535, 314, 70, 1
	PSSM + O	64.9%	70.6%	537, 179, 399, 440, 1
	Bigram	74.1%	74.7%	463, 394, 151, 205, 471
	Separated dimers (K = 7)	76%	77.1%	463, 536, 16, 1, 203
SCP	**PF1**	71.8%	79.1%	179, 216, 84, 466, 340
	PSSM + PF1	81.8%	83.7%	239, 461, 442, 1, 340
	O	67.8%	80.8%	12, 537, 179, 346, 1
	PSSM + O	77.1%	82.9%	537, 345, 70, 472, 1
	Bigram	83.3%	84.4%	463, 114, 308, 1, 2
	Separated dimers (K = 7)	86.4%	87.5%	84, 536, 114, 394, 350

***** n-fold cross-validation was carried out 100 times for statistical stability
***** Improved Accuracy refers to the n-fold cross-validation accuracy of the combination of features {*Feature*, SF}

Table 2 - TG Dataset n-fold cross validation

	Feature	Baseline Accuracy (n = 10)	Improved Accuracy (n = 10)	Rank
PFR	**PF1**	38.8%	50.4%	532, 341, 199, 461, 340
	PSSM + PF1	52.7%	59%	180, 343, 465, 463, 440
	O	36.3%	51.3%	535, 199, 349, 490, 491
	PSSM + O	46.7%	57.3%	512, 348, 461, 1, 2
	Bigram	68.1%	70.5%	494, 222, 205, 147, 81
	Separated dimers (K = 3)	73.5%	74.5%	151, 347, 460, 471, 1
SCP	**PF1**	69.9%	80.3%	209, 314, 346, 151, 443
	PSSM + PF1	77.2%	84.7%	209, 355, 442, 346, 205
	O	63.6%	81.3%	537, 209, 351, 442, 199
	PSSM + O	73.4%	84.3%	199, 348, 343, 442, 463
	Bigram	81.5%	86.8%	494, 351, 469, 217, 244
	Separated dimers (K = 3)	87.7%	89.3%	211, 63, 483, 3, 488

***** n-fold cross-validation was carried out 100 times for statistical stability

***** Improved Accuracy refers to the n-fold cross-validation accuracy of the combination of features {*Feature*, SF}

Table 3 - EDD Dataset n-fold cross validation

	Feature	Baseline Accuracy (n = 10)	Improved Accuracy (n = 10)	Rank
PFR	**PF1**	50.8%	75.1%	512, 389, 341, 221, 399
	PSSM + PF1	75.2%	82%	239, 348, 355, 340, 442
	O	46.9%	75.2%	535, 177, 178, 312, 206
	PSSM + O	68.5%	76.1%	211, 494, 466, 115, 1
	Bigram	84.5%	87.2%	111, 394, 151, 460, 1
	Separated dimers (K = 4)	89.7%	90.2%	534, 197, 462, 347, 89
SCP	**PF1**	71%	87.5%	206, 535, 355, 312, 179
	PSSM + PF1	86.1%	91.3%	206, 355, 348, 469, 191
	O	66.5%	85.1%	184, 312, 177, 399, 341
	PSSM + O	80.7%	91.3%	206, 239, 339, 84, 355
	Bigram	89.3%	90.6%	147, 365, 417, 131, 122
	Separated dimers (K = 4)	94%	94.5%	532, 211, 356, 115, 70

***** n-fold cross-validation was carried out 100 times for statistical stability

***** Improved Accuracy refers to the n-fold cross-validation accuracy of the combination of features {*Feature*, SF}

Table 4 - Increment Accuracy for features in DD dataset

	Feature	Baseline Accuracy (n = 10)	Accuracy after Level 1 (n = 10)	Accuracy after Level 2 (n = 10)	Accuracy after Level 3 (n = 10)	Accuracy after Level 4 (n = 10)	Accuracy after Level 5 (n = 10)
PFR	**PF1**	50.6%	58.7%	58.9%	60.6%	61.6%	62.3%
	PSSM + PF1	66.4%	67%	67.1%	67.5%	67.8%	69%
	O	51%	61.5%	65.1%	65.5%	65.5%	65.6%
	PSSM + O	64.9%	68.3%	69.5%	70%	70%	70.6%
	Bigram	74.1%	74.3%	74.2%	74.4%	74.4%	74.7%
	Separated dimers (K = 7)	76%	76.6%	76.5%	76.5%	76.9%	77.1%
SCP	**PF1**	71.8%	76.3%	78.3%	78.5%	78.9%	79.1%
	PSSM + PF1	81.8%	82.3%	82.3%	82.5%	82.7%	83.7%
	O	67.8%	76.2%	78.8%	79.6%	79.9%	80.8%
	PSSM + O	77.1%	82.5%	83.1%	82.8%	82.9%	82.9%
	Bigram	83.3%	83.8%	84.1%	84.1%	84.1%	84.4%
	Separated dimers (K = 7)	86.4%	86.6%	87.1%	87.4%	87.3%	87.5%

Table 5 - Increment Accuracy for features in TG dataset

	Feature	Baseline Accuracy (n = 10)	Accuracy after Level 1 (n = 10)	Accuracy after Level 2 (n = 10)	Accuracy after Level 3 (n = 10)	Accuracy after Level 4 (n = 10)	Accuracy after Level 5 (n = 10)
PFR	**PF1**	38.8%	46.2%	47.3%	49%	49.8%	50.4%
	PSSM + PF1	52.7%	56.1%	57.1%	58.4%	58.6%	59%
	O	36.3%	49.4%	50.4%	51%	51%	51.3%
	PSSM + O	46.7%	55.4%	56.7%	57%	57%	57.3%
	Bigram	68.1%	69.8%	69.9%	70%	69.8%	70.5%
	Separated dimers (K = 3)	73.5%	74.7%	74.8%	74.8%	74.7%	74.5%
SCP	**PF1**	69.9%	78%	78.9%	79.6%	79.7%	80.3%
	PSSM + PF1	77.2%	82.9%	84.1%	84.5%	84.6%	84.7%
	O	63.6%	76.8%	80.5%	81.1%	81.7%	81.3%
	PSSM + O	73.4%	81.3%	82.7%	83.4%	84%	84.3%
	Bigram	81.5%	85.2%	85.7%	86.2%	86.5%	86.8%
	Separated dimers (K = 3)	87.7%	89%	89%	89.1%	89.1%	89.3%

Table 6 - Increment Accuracy for features in EDD dataset

	Feature	Baseline Accuracy (n = 10)	Accuracy after Level 1 (n = 10)	Accuracy after Level 2 (n = 10)	Accuracy after Level 3 (n = 10)	Accuracy after Level 4 (n = 10)	Accuracy after Level 5 (n = 10)
PFR	**PF1**	50.8%	65.2%	71%	73.1%	74.5%	75.1%
	PSSM + PF1	75.2%	80.1%	80.5%	81.8%	81.9%	82%
	O	46.9%	65.4%	70.5%	72.9%	74.8%	75.2%
	PSSM + O	68.5%	73.9%	75.4%	76%	76.1%	76.1%
	Bigram	84.5%	87.1%	87.2%	86.9%	87.2%	87.2%
	Separated dimers (K = 4)	89.7%	89.8%	89.9%	89.9%	89.8%	90.2%
SCP	**PF1**	71%	82.3%	85%	85.8%	87.3%	87.5%
	PSSM + PF1	86.1%	89.7%	90.4%	90.9%	91%	91.3%
	O	66.5%	83.2%	83.7%	84.7%	84.8%	85.1%
	PSSM + O	80.7%	88.1%	89.6%	90.6%	90.8%	91.3%
	Bigram	89.3%	90.1%	90.2%	90.3	90.3%	90.6%
	Separated dimers (K = 4)	94%	94.2%	94.3%	94.3%	94.4%	94.5%

Table 7 – Rank of attributes based on its frequency counts over all the Datasets

Attribute	Total Frequency
1	11
199, 355, 442, 463, 537	5
151, 206, 340, 348, 535	4

70, 84, 179, 209, 211, 239, 312, 314, 341, 351, 394, 399, 461, 469, 494	3
2, 12, 114, 115, 147, 177, 205, 339, 343, 346, 347, 440, 460, 466, 471, 491, 512, 532, 536	2
3, 16, 63, 81, 89, 111, 122, 131, 178, 180, 184, 191, 197, 203, 216, 217, 221, 222, 244, 308, 317, 345, 349, 350, 356, 365, 389, 417, 453, 462, 465, 472, 483, 488, 490, 534	1

Figure 2 - Boxplot for PF1 Feature for PFR

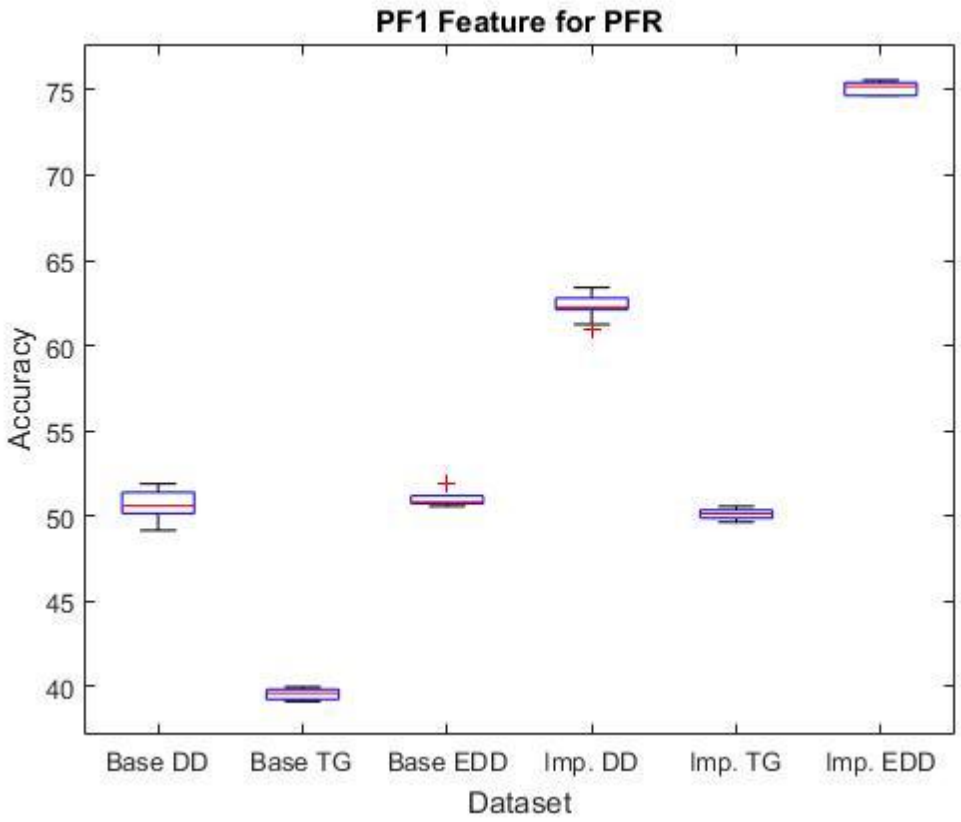


Figure 3 - Boxplot for PSSM + PF1 Feature for PFR

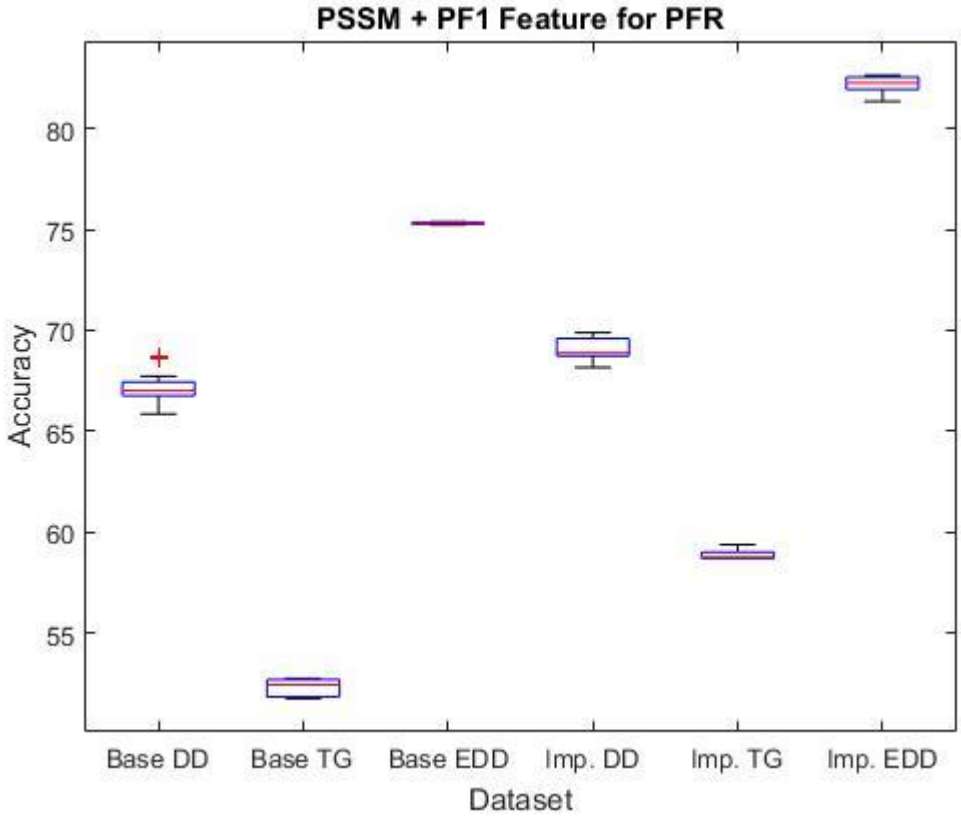


Figure 4 - Boxplot for O Feature for PFR

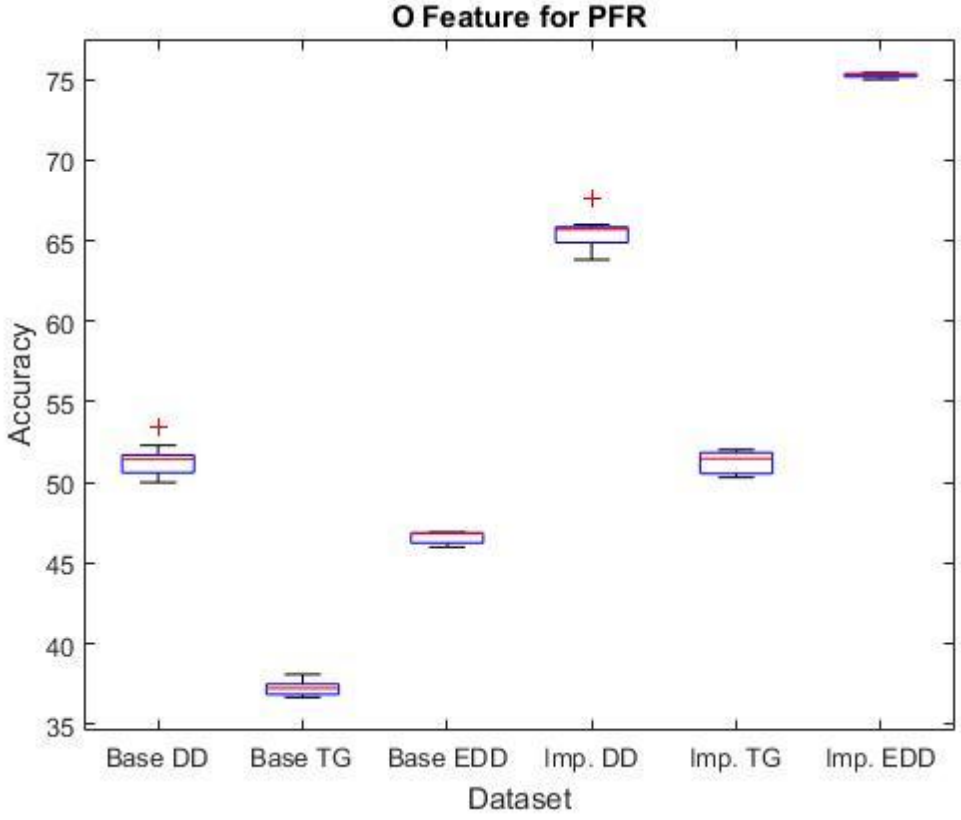


Figure 5 - Boxplot for PSSM + O Feature for PFR

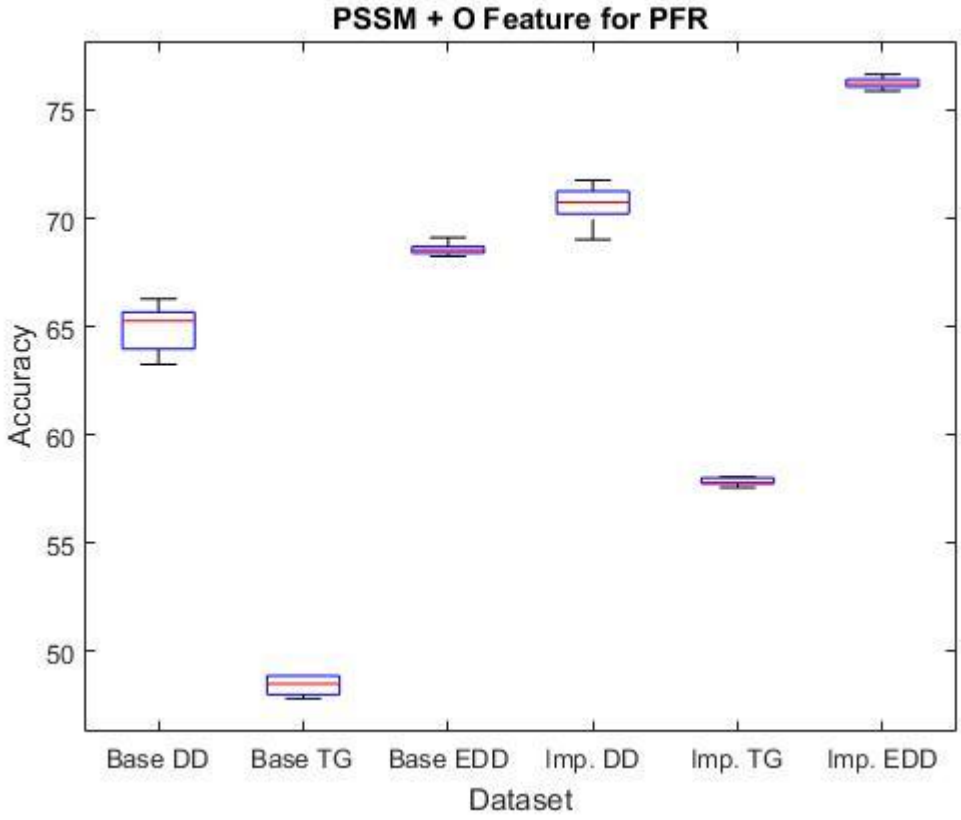


Figure 6 - Boxplot for Bigram Feature for PFR

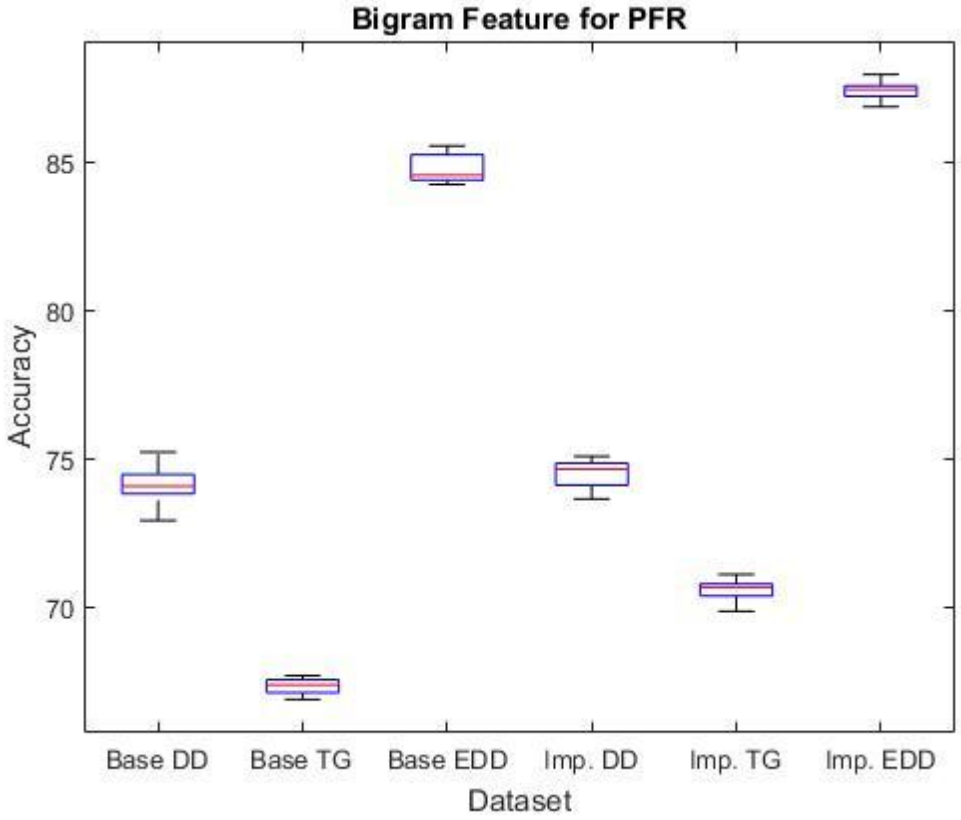


Figure 7 - Boxplot for Separated Dimers Feature for PFR

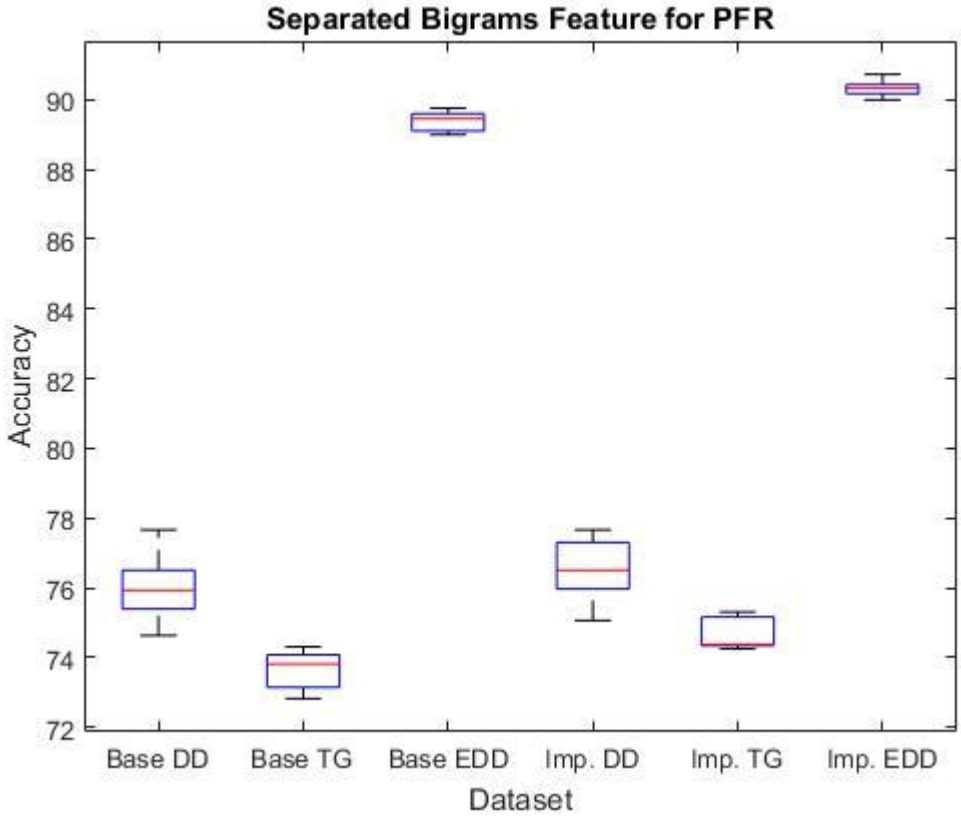


Figure 8 - Boxplot for PF1 Feature for SCP

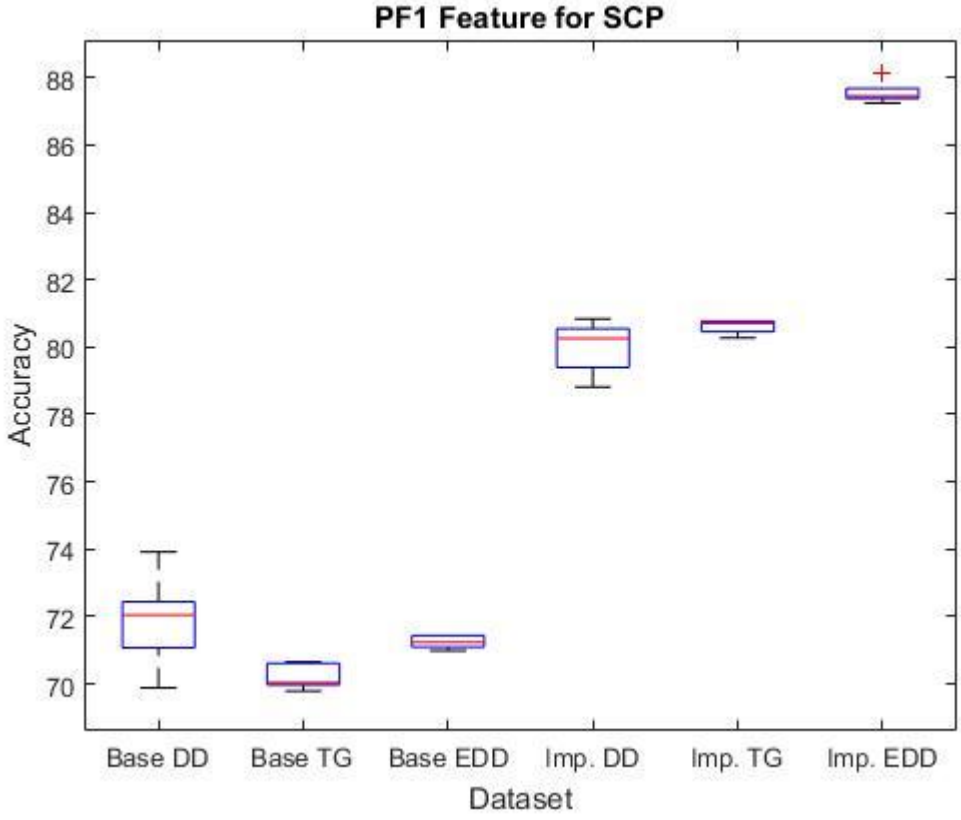


Figure 9 - Boxplot for PSSM + PF1 Feature for SCP

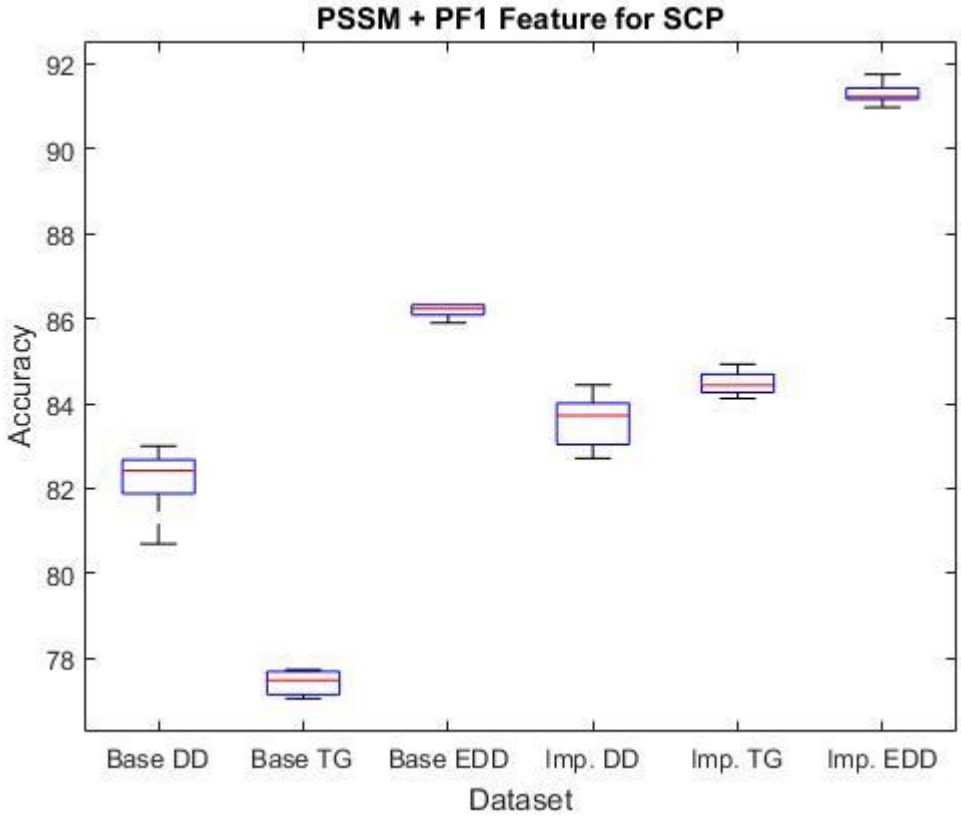


Figure 10 - Boxplot for O Feature for SCP

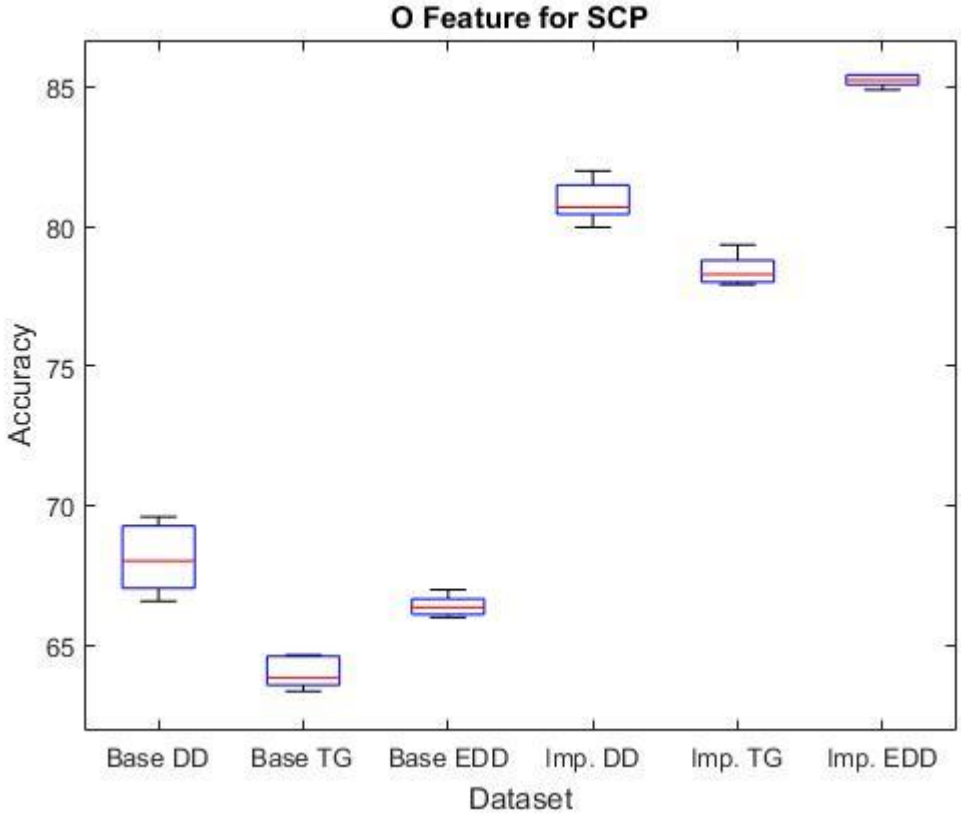


Figure 11 - Boxplot for PSSM + O Feature for SCP

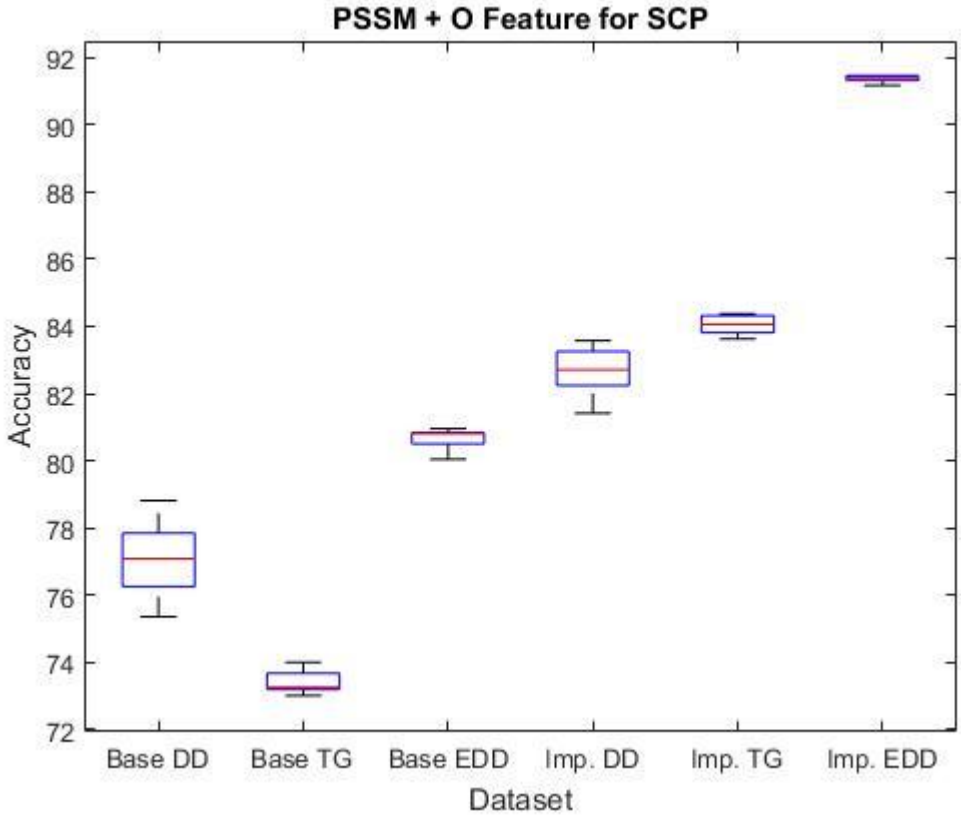


Figure 12 - Boxplot for Bigram Feature for SCP

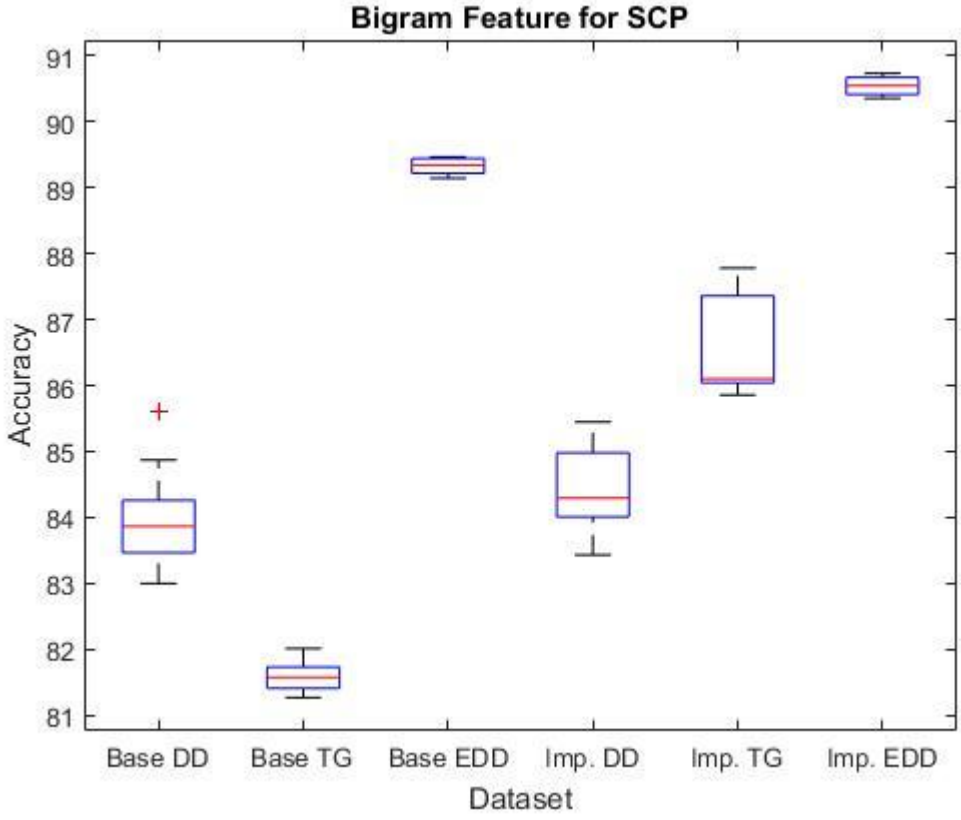
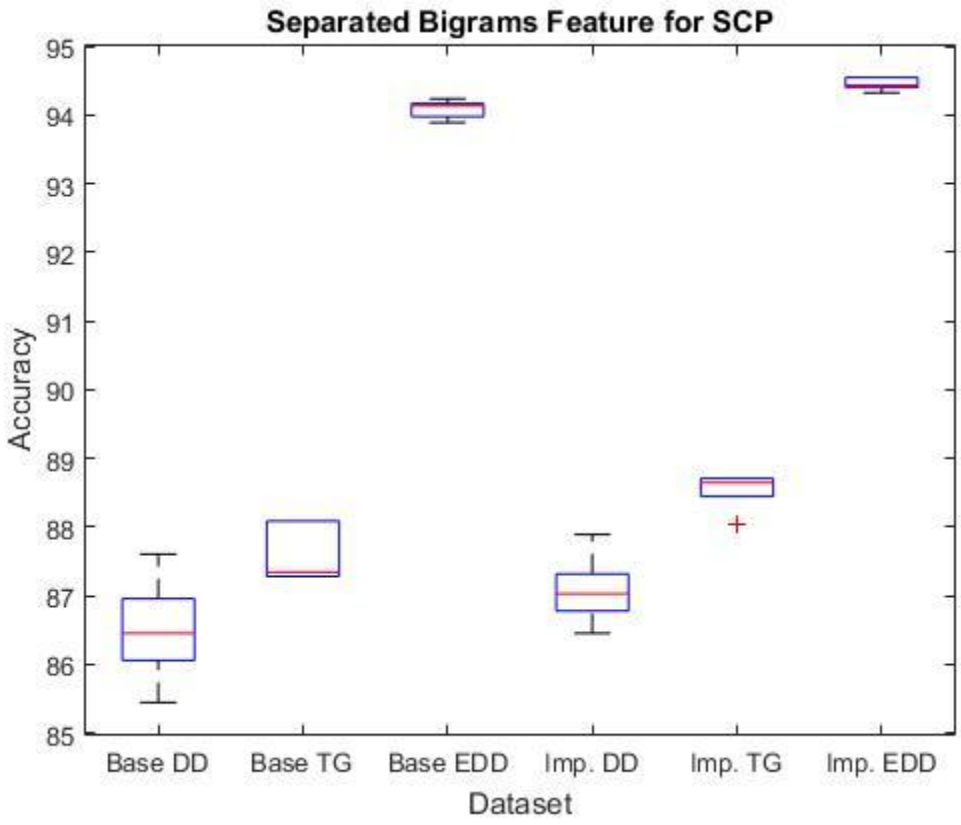


Figure 13 - Boxplot for Separated Bigrams Feature for SCP



Appendix - Physicochemical attributes selected in this study

(The full list of attributes can be found at link – http://www.genome.jp/aaindex/)

Number	Attribute (Reference)
1	alpha-CH chemical shifts (Andersen et al., 1992)
2	Hydrophobicity index (Argos et al., 1982)
3	Signal sequence helical potential (Argos et al., 1982)
12	Retention coefficient in TFA (Browne et al., 1982)
16	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
63	Size (Dawson, 1972)
70	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)

81	STERIMOL length of the side chain (Fauchere et al., 1988)
84	N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)
89	Negative charge (Fauchere et al., 1988)
111	Polarity (Grantham, 1974)
114	Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)
115	Hydrophilicity value (Hopp-Woods, 1981)
122	Normalized relative frequency of bend R (Isogai et al., 1980)
131	Transfer free energy (Janin, 1979)
147	Side chain interaction parameter (Krigbaum-Rubin, 1971)
151	Hydropathy index (Kyte-Doolittle, 1982)
177	Refractivity (McMeekin et al., 1964)
178	Retention coefficient in HPLC, pH7.4 (Meek, 1980)
179	Retention coefficient in HPLC, pH2.1 (Meek, 1980)
180	Retention coefficient in NaClO4 (Meek-Rossetti, 1981)
184	Average side chain orientation angle (Meirovitch et al., 1980)
191	AA composition of mt-proteins (Nakashima et al., 1990)
197	AA composition of membrane proteins (Nakashima et al., 1990)
199	Transmembrane regions of non-mt-proteins (Nakashima et al., 1990)
203	AA composition of CYT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)

205 AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)

206 AA composition of MEM of single-spanning proteins (Nakashima-Nishikawa, 1992)

209 AA composition of MEM of multi-spanning proteins (Nakashima-Nishikawa, 1992)

211 14 A contact number (Nishikawa-Ooi, 1986)

216 Average non-bonded energy per residue (Oobatake-Ooi, 1977)

217 Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977)

221 Optimized average non-bonded energy per atom (Oobatake et al., 1985)

222 Optimized side chain interaction parameter (Oobatake et al., 1985)

239 HPLC parameter (Parker et al., 1986)

244 Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al., 1980)

308 Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982)

312 Value of theta(i) (Rackovsky-Scheraga, 1982)

314 Transfer free energy from chx to wat (Radzicka-Wolfenden, 1988)

317 Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)

339 Information measure for alpha-helix (Robson-Suzuki

340 Information measure for N-terminal helix (Robson-Suzuki, 1976)

341 Information measure for middle helix (Robson-Suzuki, 1976)

343 Information measure for extended (Robson-Suzuki, 1976)

345 Information measure for extended without H-bond (Robson-Suzuki, 1976)

346 Information measure for turn (Robson-Suzuki, 1976)

347 Information measure for N-terminal turn (Robson-Suzuki, 1976)

348 Information measure for middle turn (Robson-Suzuki, 1976)

349 Information measure for C-terminal turn (Robson-Suzuki, 1976)

350 Information measure for coil (Robson-Suzuki, 1976)

351 Information measure for loop (Robson-Suzuki, 1976)

355 Side chain hydropathy, uncorrected for solvation (Roseman, 1988)

356 Side chain hydropathy, corrected for solvation (Roseman, 1988)

365 Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)

389 Hydration potential (Wolfenden et al., 1981)

394 Unfolding Gibbs energy in water, pH9.0 (Yutani et al., 1987)

399 Bulkiness (Zimmerman et al., 1968)

417 Normalized positional residue frequency at helix termini C1 (Aurora-Rose, 1998)

440 Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000)

442 Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al., 2000)

443 Distribution of amino acid residues in the alpha-helices in mesophilic proteins (Kumar et al., 2000)

453 Averaged turn propensities in a transmembrane helix (Monne et al., 1999)

460 Composition of amino acids in nuclear proteins (percent) (Cedano et al., 1997)

Surface composition of amino acids in intracellular proteins of thermophiles (percent)
461 (Fukuchi-Nishikawa, 2001)

Surface composition of amino acids in intracellular proteins of mesophiles (percent)
462 (Fukuchi-Nishikawa, 2001)

Surface composition of amino acids in extracellular proteins of mesophiles (percent)
463 (Fukuchi-Nishikawa, 2001)

Interior composition of amino acids in intracellular proteins of thermophiles (percent)
465 (Fukuchi-Nishikawa, 2001)

Interior composition of amino acids in intracellular proteins of mesophiles (percent)
466 (Fukuchi-Nishikawa, 2001)

Entire chain composition of amino acids in intracellular proteins of thermophiles
469 (percent) (Fukuchi-Nishikawa, 2001)

Entire chain composition of amino acids in extracellular proteins of mesophiles
471 (percent) (Fukuchi-Nishikawa, 2001)

Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-
472 Nishikawa, 2001)

483 Amphiphilicity index (Mitaku et al., 2002)

Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O (Wilce et al.
488 1995)

Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al.
490 1995)

491 Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O (Wilce

et al. 1995)

494 Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)

512 Buriability (Zhou-Zhou, 2004)

532 PRIFT index (Cornette et al., 1987)

534 ALTFT index (Cornette et al., 1987)

535 ALTLS index (Cornette et al., 1987)

536 TOTFT index (Cornette et al., 1987)

537 TOTLS index (Cornette et al., 1987)

References

1. Ali, F., & Hayat, M. (2015). Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition. Journal of theoretical biology, 384, 78-83.

2. Bahar, I., Atilgan, A. R., Jernigan, R. L., & Erman, B. (1997). Understanding the recognition of protein structural classes by amino acid composition. Proteins Structure Function and Genetics, 29(2), 172-185.

3. Bologna, G., & Appel, R. D. (2002, November). A comparison study on protein fold recognition. In *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on* (Vol. 5, pp. 2492-2496). IEEE.

4. Bulashevska, A., & Eils, R. (2006). Predicting protein subcellular locations using hierarchical

ensemble of Bayesian classifiers based on Markov chains.*Bmc Bioinformatics*, *7*(1), 298.

5. Cai, Y. D., & Zhou, G. P. (2000). Prediction of protein structural classes by neural network. *Biochimie*, *82*(8), 783-785.

6. Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

7. Chen, C., Tian, Y. X., Zou, X. Y., Cai, P. X., & Mo, J. Y. (2006a). Using pseudo-amino acid composition and support vector machine to predict protein structural class. *Journal of Theoretical Biology*, *243*(3), 444-448.

8. Chen, C., Zhou, X., Tian, Y., Zou, X., & Cai, P. (2006b). Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Analytical biochemistry, 357(1), 116-121.

9. 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 and peptide letters, 19(4), 422-429.

10. Chen, W., Ding, H., Feng, P., Lin, H., & Chou, K. C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget.

11. Chen, W., Feng, P., Ding, H., Lin, H., & Chou, K. C. (2015). Using deformation energy to analyze nucleosome positioning in genomes. Genomics.

12. Chinnasamy, A., Sung, W. K., & Mittal, A. (2005). Protein structure and fold prediction using tree-augmented naive Bayesian classifier. *Journal of Bioinformatics and Computational Biology*, *3*(04), 803-819.

13. Chmielnicki, W., & Stąpor, K. (2012). A hybrid discriminative/generative approach to protein fold recognition. *Neurocomputing*, *75*(1), 194-198.

14. Chmielnicki, W., & Stąpor, K. (2011). A combined SVM-RDA classifier for protein fold recognition. Bio-Algorithms and Med-Systems, 7.

15. Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, *43*(3), 246-255.

16. Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology, 273(1), 236-247.

17. Chou, K. C., & Zhang, C. T. (1995). Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology, 30(4), 275

18. Chou, K. C., & Cai, Y. D. (2005). Prediction of membrane protein types by incorporating amphipathic effects. Journal of chemical information and modeling, 45(2), 407-413.-349.

19. Chou, K. C. (2015). Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry, 11(3), 218-234.

20. Chou, K. C., & Zhang, C. T. (1994). Predicting protein folding types by distance functions that make allowances for amino acid interactions. Journal of Biological Chemistry, 269(35), 22014-22020.

21. Chou, K. C. (1995). A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins: Structure, Function, and Bioinformatics, 21(4), 319-344.

22. Chou, K. C., & Maggiora, G. M. (1998). Domain structural class prediction. Protein Engineering, 11(7), 523-538.

23. Chou, K. C., & Cai, Y. D. (2004). Predicting protein structural class by functional domain composition. Biochemical and biophysical research communications, 321(4), 1007-1009.

24. Cormen, T. H., & Leiserson, C. E. and R. L. Rivest (1990), Introduction to Algorithms.

25. Craven, M., Mural, R. J., Hauser, L. J., & Uberbacher, E. C. (1995, December). Predicting protein folding classes without overly relying on homology. In *ISMB* (Vol. 3, pp. 98-106).

26. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., & Sattar, A. (2014a). Proposing a highly accurate protein structural class predictor using segmentation-based features. *BMC genomics*, *15*(Suppl 1), S2.

27. Dehzangi, A., Sharma, A., Lyons, J., Paliwal, K. K., & Sattar, A. (2014b). A mixture of physicochemical and evolutionary–based feature extraction approaches for protein fold recognition. *International Journal of Data Mining and Bioinformatics*, *11*(1), 115-138.

28. Dehzangi, A., Paliwal, K., Sharma, A., Dehzangi, O., & Sattar, A. (2013a). A combination of

feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, *10*(3), 564-575.

29. Dehzangi, A., & Phon-Amnuaisuk, S. (2011). Fold prediction problem: The application of new physical and physicochemical-based features. *Protein and Peptide Letters*, *18*(2), 174-185.

30. Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., & Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC. *Journal of theoretical biology*, 364, 284-294.

31. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., & Sattar, A. (2014c). A segmentation-based method to extract structural and evolutionary features for protein fold recognition. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 11(3), 510-519.

32. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., & Sattar, A. (2013b). Enhancing protein fold prediction accuracy using evolutionary and structural features. In *Pattern Recognition in Bioinformatics* (pp. 196-207). Springer Berlin Heidelberg.

33. Dehzangi, A., Paliwal, K., Sharma, A., Lyons, J., & Sattar, A. (2013c). Protein fold recognition using an overlapping segmentation approach and a mixture of feature extraction models. In AI 2013: *Advances in Artificial Intelligence* (pp. 32-43). Springer International Publishing.

34. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., & Sattar, A. (2013d). Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy. In *Pattern Recognition in Bioinformatics* (pp. 208-219). Springer Berlin Heidelberg.

35. Deschavanne, P., & Tufféry, P. (2009). Enhanced protein fold recognition using a structural alphabet. *Proteins: Structure, Function, and Bioinformatics*, *76*(1), 129-137.

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

37. 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 Letters*, *29*(13), 1887-1892.

38. Dong, Q., Zhou, S., & Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, *25*(20), 2655-2662.

39. Dubchak, I., Muchnik, I. B., & Kim, S. H. (1997, June). Protein folding class predictor for SCOP: approach based on global descriptors. In *Ismb* (pp. 104-107).

40. Ghanty, P., & Pal, N. R. (2009). Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. *NanoBioscience, IEEE Transactions on*, *8*(1), 100-110.

41. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.

42. Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., ... & Zhou, Y. (2015a). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports*, 5.

43. Heffernan, R., Dehzangi, A., Lyons, J., Paliwal, K., Sharma, A., Wang, J., ... & Yang, Y. (2015b). Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics*, btv665.

44. Huang, C. D., Lin, C. T., & Pal, N. R. (2003). Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification. *NanoBioscience, IEEE Transactions on*, *2*(4), 221-232.

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

46. Jia, J., Liu, Z., Xiao, X., Liu, B., & Chou, K. C. (2016a). iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules, 21(1), 95.

47. Jia, J., Liu, Z., Xiao, X., Liu, B., & Chou, K. C. (2016b). pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of Theoretical Biology.

48. Jia, J., Liu, Z., Xiao, X., Liu, B., & Chou, K. C. (2016c). iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components

and optimizing imbalanced training dataset. Analytical biochemistry.

49. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic acids research*, *36*(suppl 1), D202-D205.

50. Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B. N., & Moosavi-Movahedi, A. A. (2011). A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM. *Computational biology and chemistry*, *35*(1), 1-9.

51. Krishnaraj, Y., & Reddy, C. K. (2008, November). Boosting methods for protein fold recognition: an empirical comparison. In *Bioinformatics and Biomedicine, 2008. BIBM'08. IEEE International Conference on* (pp. 393-396). IEEE.

52. Kumar, R., Srivastava, A., Kumari, B., & Kumar, M. (2015). Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. Journal of theoretical biology, 365, 96-103.

53. Kurgan, L., & Homaeian, L. (2005). Prediction of secondary protein structure content from primary sequence alone–a feature selection based approach. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 334-345). Springer Berlin Heidelberg.

54. Kurgan, L. A., Zhang, T., Zhang, H., Shen, S., & Ruan, J. (2008). Secondary structure-based assignment of the protein structural classes. *Amino Acids*, *35*(3), 551-564.

55. Lin, C., Zou, Y., Qin, J., Liu, X., Jiang, Y., Ke, C., & Zou, Q. (2013). Hierarchical classification of protein folds using a novel ensemble classifier. *PloS one*, 8(2), e56499.

56. Liu, B., Fang, L., Liu, F., Wang, X., & Chou, K. C. (2016a). iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. *Journal of Biomolecular Structure and Dynamics*, 1-13.

57. Liu, B., Fang, L., Long, R., Lan, X., & Chou, K. C. (2016b). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, btv604.

58. 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*(6), 2243-2249.

59. Liu, Z., Xiao, X., Yu, D. J., Jia, J., Qiu, W. R., & Chou, K. C. (2016). pRNAm-PC: Predicting N 6-methyladenosine sites in RNA sequences via physical–chemical properties. *Analytical biochemistry*.

60. Lyons, J., Dehzangi, A., Heffernan, R., Yang, Y., Zhou, Y., Sharma, A., & Paliwal, K. (2015). Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles from Hidden Markov Models. *IEEE Transactions on NanoBioscience*.

61. 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. *Journal of theoretical biology*, *354*, 137-145.

62. Lyons, J., Paliwal, K. K., Dehzangi, A., Hefferman, R., Tsunoda, T., & Sharma, A. (2016). Protein fold recognition using HMM-HMM alignment and Dynamic Programming. *Journal of Theoretical Biology*.

63. Mondal, S., & Pai, P. P. (2014). Chou′ s pseudo amino acid composition improves sequence-based antifreeze protein prediction. Journal of theoretical biology, 356, 30-35.

64. Mizianty, M. J., & Kurgan, L. (2009). Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC bioinformatics*, *10*(1), 414.

65. 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. *Journal of molecular biology*, *247*(4), 536-540.

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

67. Nanni, L. (2006). Ensemble of classifiers for protein fold recognition. *Neurocomputing*, *69*(7), 850-853.

68. Nanni, L., Brahnam, S., & Lumini, A. (2014). Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. Journal of theoretical biology, 360, 109-116.

69. Pal, N. R., & Chakraborty, D. (2003). Some new features for protein fold prediction. In *Artificial*

Neural Networks and Neural Information Processing—ICANN/ICONIP 2003 (pp. 1176-1183). Springer Berlin Heidelberg.

70. Paliwal, K. K., Sharma, A., Lyons, J., & Dehzangi, A. (2014a). A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. *NanoBioscience, IEEE Transactions on*, *13*(1), 44-50.

71. Paliwal, K. K., Sharma, A., Lyons, J., & Dehzangi, A. (2014b) Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information. *BMC bioinformatics*, 15(Suppl 16), S12.

72. 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 and peptide letters, 19(4), 388-397.

73. Sahu, S. S., & Panda, G. (2010). A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational biology and chemistry, 34(5), 320-327.

74. Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Lyons, J., ... & Miyano, S. (2015). Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. *Journal of theoretical biology*, *380*, 291-298.

75. Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Rajeshkannan, A., ... & Paliwal, K. K. (2014). Protein structural class prediction via k-separated bigrams using position specific scoring matrix. *J. Adv. Comput. Intell. Intell. Informatics*, *8*(4).

76. 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.

77. Sharma, A., Imoto, S., & Miyano, S. (2012a). A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *9*(3), 754-764.

78. 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. *Journal of theoretical biology*, *320*, 41-46.

79. 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*(2), e89890.

80. Sharma, A., Paliwal, K. K., Dehzangi, A., Lyons, J., Imoto, S., & Miyano, S. (2013b). A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. *BMC bioinformatics*, *14*(1), 233.

81. Sharma, A., Paliwal, K. K., & Onwubolu, G. C. (2006). Class-dependent PCA, MDC and LDA: A combined classifier for pattern classification. *Pattern Recognition*, 39(7), 1215-1229.

82. Sharma, A., Koh, C. H., Imoto, S., & Miyano, S. (2011). Strategy of finding optimal number of features on gene expression data. *Electronics letters*, 47(8), 480-482.

83. Sharma, A., Imoto, S., Miyano, S., & Sharma, V. (2012b). Null space based feature selection method for gene expression data. *International Journal of Machine Learning and Cybernetics*, 3(4), 269-276.

84. Sharma, A., Imoto, S., & Miyano, S. (2012c). A between-class overlapping filter-based method for transcriptome data analysis. *Journal of bioinformatics and computational biology*, 10(05), 1250010.

85. 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. *Current Bioinformatics*, 7(3), 289-294.

86. Sharma, A., Paliwal, K. K., Imoto, S., & Miyano, S. (2013c). Principal component analysis using QR decomposition. *International Journal of Machine Learning and Cybernetics*, 4(6), 679-683.

87. Sharma, A., & Paliwal, K. K. (2007). Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, 28(10), 1151-1155.

88. Sharma, A., & Paliwal, K. K. (2010). Regularisation of eigenfeatures by extrapolation of scatter-matrix in face-recognition problem. *Electronics Letters*, 46(10), 1.

89. Sharma, A., & Paliwal, K. K. (2012a). A two-stage linear discriminant analysis for face-recognition.

Pattern Recognition Letters, 33(9), 1157-1162.

90. Sharma, A., & Paliwal, K. K. (2012b). A gene selection algorithm using Bayesian classification approach. *American Journal of Applied Sciences*,9(1), 127-131.

91. 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 Recognition*, 45(6), 2205-2213.

92. Sharma, A., & Paliwal, K. K. (2015). A deterministic approach to regularized linear discriminant analysis. *Neurocomputing*, 151, 207-214.

93. Sharma, A., Boroevich, K., Shigemizu, D., Kamatani, Y., Kubo, M., & Tsunoda, T. (2016). Hierarchical Maximum Likelihood Clustering Approach, IEEE Transactions on Biomedical Engineering, DOI: 10.1109/TBME.2016.2542212.

94. Shen, H. B., & Chou, K. C. (2006). Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, 22(14), 1717-1722.

95. Shen, H. B., & Chou, K. C. (2007). Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers, 85(3), 233-240.

96. Taguchi, Y. H., & Gromiha, M. M. (2007). Application of amino acid occurrence for discriminating different folding types of globular proteins. *BMC bioinformatics*, *8*(1), 404.

97. Yang, T., Kecman, V., Cao, L., Zhang, C., & Huang, J. Z. (2011). Margin-based ensemble classifier for protein fold recognition. *Expert Systems with Applications*, *38*(10), 12348-12355.

98. 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*(1), 271-283.

99. 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. Journal of theoretical biology, 355, 105-110.

100. Zhou, G. P. (1998). An intriguing controversy over protein structural class prediction. Journal of

protein chemistry, 17(8), 729-738.

101. Zhou, G. P., & Assa-Munt, N. (2001). Some insights into protein structural class prediction. Proteins: Structure, Function, and Bioinformatics, 44(1), 57-59.

102. Zhou, X. B., Chen, C., Li, Z. C., & Zou, X. Y. (2008). Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. *Amino Acids*, *35*(2), 383-388.

103. Zimmerman, J. M., Eliezer, N., & Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology*, *21*(2), 170-201.

Highlights

- A Forward Consecutive Search (FCS) scheme is proposed
- Physicochemical attributes are strategically selected
- Physicochemical-based features supplement existing feature extraction techniques
- Improvements in prediction accuracies after utilizing physicochemical information