EvoStruct-Sub: An Accurate Gram-positive Protein Subcellular Localization Predictor Using Evolutionary and Structural Features

Md. Raihan Uddin, Alok Sharma, Dewan Md Farid,
Md. Mahmudur Rahman, Abdollah Dehzangi, Swakkhar Shatabda

PII: S0022-5193(18)30055-9
DOI: 10.1016/j.jtbi.2018.02.002
Reference: YJTBI 9345

To appear in: *Journal of Theoretical Biology*

Received date: 30 December 2017
Revised date: 18 January 2018
Accepted date: 3 February 2018

Please cite this article as: Md. Raihan Uddin, Alok Sharma, Dewan Md Farid, Md. Mahmudur Rahman, Abdollah Dehzangi, Swakkhar Shatabda, EvoStruct-Sub: An Accurate Gram-positive Protein Subcellular Localization Predictor Using Evolutionary and Structural Features, *Journal of Theoretical Biology* (2018), doi: 10.1016/j.jtbi.2018.02.002

Highlights

- Structural Features for Classification
- Evolutionary Features for Classification
- Effective Prediction
- Feature Selection

EvoStruct-Sub: An Accurate Gram-positive Protein Subcellular Localization Predictor Using Evolutionary and Structural Features

Md. Raihan Uddin[1], Alok Sharma[2,3,5], Dewan Md Farid[1], Md. Mahmudur Rahman[4], Abdollah Dehzangi[4,*] and Swakkhar Shatabda[1,*]

[1] *Department of Computer Science and Engineering, United International University, Bangladesh*
[2] *School of Engineering and Physics, University of the South Pacific, Fiji*
[3] *Institute for Integrated and Intelligent Systems, Griffith University, Australia*
[4] *Department of Computer Science, Morgan State University, United States*
[5] *RIKEN Center for Integrative Medical Sciences, Yokohama, Japan*
[*] *Corresponding authors*

Abstract

Determining subcellular localization of proteins is considered as an important step towards understanding their functions. Previous studies have mainly focused solely on Gene Ontology (GO) as the main feature to tackle this problem. However, it was shown that features extracted based on GO is hard to be used for new proteins with unknown GO. At the same time, evolutionary information extracted from Position Specific Scoring Matrix (PSSM) have been shown as another effective features to tackle this problem. Despite tremendous advancement using these sources for feature extraction, this problem still remains unsolved. In this study we propose EvoStruct-Sub which employs predicted structural information in conjunction with evolutionary information extracted directly from the protein sequence to tackle this problem. To do this we use several different feature extraction method that have been shown promising in subcellular localization as well as similar studies to extract effective local and global discriminatory information. We then use Support Vector Machine (SVM) as our classification technique to build EvoStruct-Sub. As a result, we are able to enhance Gram-positive subcellular localization prediction accuracies by up to 5.6% better than previous studies including the studies that used GO for

feature extraction.

Keywords: Proteins subcellular localization, Evolutionary-based Features, Structural-based Features, Classification, Support Vector Machine, Feature Selection

1. INTRODUCTION

The functioning of a protein depends on its location in the cell. In fact, it just functions properly in one or a few locations in the cell. Knowing those locations can provide important information about functioning of the proteins and how they interact with other micro-molecules. Therefore, determining protein subcellular localization is considered as an important step towards understanding its functioning [1, 2].

Of all proteins, bacterial proteins are the most important proteins to determine their functions because of their biological aspects which are both harmful and useful [3]. Some bacteria can cause a wide range of diseases while some others play the role of catalyst in biological interactions. Some bacteria are also widely used to produce antibiotics. Bacteria are categorized as a kind of prokaryotic microorganism that can be divided in two groups, Gram-positive and Gram-negative [4]. Gram-positive bacteria are those that are stained dark blue or violet by Gram-staining while Gram-negative bacteria cannot retain the stain, instead taking up the counter-stain and appearing red or pink [3].

During the past two decades and since the introduction of bacterial proteins subcellular localization, a wide range of machine learning methods with many different combination and types of features have been proposed to tackle this problem [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 24, 25, 26]. For example, PSORT (predictor) used sequence features based on sorting signal [27], SubLoc (predictor) uses SVM with AAC to obtain higher accuracy [28], and TargetP (predictor) uses ANN and N-terminal sequence to predict subcellular locations [29]. In addition Pierleoni et al. used N-terminal, AAC and alignment profile to predict the subcellular localization [30]. Similarly, Tamura

and Akutsu used alignment of block sequence [31] and Chang et al. developed and used gapped-dipeptide and probabilistic latent semantic analysis method for prediction of Gram-negative bacterial protein [32]. Lee et al. Despite all the efforts have been made so far, the protein subcellular localization prediction problem for bacterial proteins have remained unsolved.

As pointed in a recent review [33], in the last decade or so, a number of web-servers were also developed for predicting the subcellular localization of proteins with both single site and multiple sites based on their sequences information alone. They can be roughly classified into two series [34]. One is the PLoc series and the other is iLoc series. The PLoc series contains the six web-servers [4], [35], [36], [37], [38], [39] to deal with eukaryotic, human, plant, Gram-positive, Gram-negative, and virus proteins, while the iLoc series contains the seven web-servers [40], [41], [42] , [43], [3], [40], [44] to deal with eukaryotic, human, plant, animal, Gram-positive, Gram-negative, and virus proteins, respectively.

In addition, Huang and Yuan analyzed series of classifiers for subcellular localization, but these were limited to single location site. For multi label prediction, Gpos-mplock and Gneg-mplock (predictor) are proposed [36], [38] to predict protein localization in Gram-positive and Gram-negative bacteria; and Plant-mploc (predictor) is developed [37] which uses top down strategy to predict single or multiple protein localization in plant protein. Virus-mploc (predictor) [39] was developed with fusion of classifiers and features of functional domain and gene ontology to predict virus proteins. To increase the quality of prediction, three revised version of the prediction systems were developed: iloc-Gpos (predictor) [3], iloc-plant (predictor) [42], iloc-virus (predictor) [44]. Huang and Yuan used AAC, evolution information and PseACC with backward propagation (BP) and radial basis function (RBF) neural network to predict both single and multi-site subcellular proteins.

Many of those studies that have mentioned earlier relied on Gene Ontology as their feature to tackle this problem [45, 46, 47, 48, 49]. Despite promising results achieved using GO, it is hard or even for some cases impossible to use these features with new proteins with unknown GO. Therefore, there is an

emphasis on proposing methods that rely on features directly extracted from protein sequence without using any other extra information or meta data that are available for just known proteins.

Early studies have focused on using features that are extracted from the occurrence of amino acids from the protein sequence [50, 45, 51]. Later studies try to incorporate physicochemical based features to enhance the prediction performance [52]. However, the protein subcellular localization prediction problem remained limited using these sources of features.

More recent studies have started using evolutionary information to tackle this problem [2, 53, 54, 5, 7, 55, 6]. Using these features they demonstrated significant enhancement and even achieved comparable performance compared to use of GO as input feature. In fact, application of evolutionary based features have demonstrated its superiority over using occurrence or physicochemical based features in many similar studies found in the literature [56, 57, 58, 59, 60, 61]. However, further enhancement have remained out of reach relying on these features. In addition, many of those studies tried to address Gram-positive and Gram-negative subcellular localization at the same time. Despite lots of similarities, still they have their own differences in nature based on their biological properties. Therefore, similar to all the other subcellular-localization prediction problems, a well designed method that is tailored for that specific task (either Gram-positive or Gram-negative) has a better chance to achieve more promising results.

To develop a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications [20, 21, 22, 23, 24, 33, 12, 62, 63, 64, 65, 66], one should observe the Chou's 5-step rule [67]; i.e., making the following five steps very clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to properly perform cross-validation tests to objectively

evaluate the anticipated accuracy of the predictor; (v) how to establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by- one.

To address these problems, here we propose EvoStruct-Sub which in addition to evolutionary information uses predicted structural information to specifically predict Gram-positive subcellular localization. To do this, we first extract evolutionary information from Position Specific Scoring Matrix (PSSM) [68] and predicted secondary structure using SPDER 2.0 [69, 70]. We then extract global and local discriminatory information using segmentation technique for our classification task [71, 72]. We finally use Support Vector Machine (SVM) to our extracted features to build EvoStruct-Sub. By applying EvoStruct-Sub to Gram-positive subcellular localization, we achieve up to 95.4% prediction accuracy for this task. In addition, we achieve to over 90.0% prediction accuracy for this problem when we use our method for multi-label samples. These results are over 5.0% better than previously reported results found in the literature [73, 5, 6].

2. MATERIALS AND METHODS

In this section, we describe the materials and methods required to develop EvoStruct-Sub.

2.1. Benchmark Dataset

In this research we use a dataset which have been used widely in literature for Gram-positive subcellular localization [74], [75], [76], [77], [78]. The benchmark that we use in this study was introduced in [74], [75], [76], [77]. This benchmark contains total 523 protein samples which belongs to four Gram-positive subcellular localizations. Among this 523 samples there are total 519 different protein sample. Among 519 proteins there are total 515 protein samples which belongs to only one or single location while the rest 4 protein samples belongs to two locations. Thus Gram-positive bacterial protein benchmark contains total 523 (515 + 4 * 2) protein samples. The name and the number of

proteins in these four locations are shown in Table 1. This dataset is available at: http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi.

Table 1: Details of Gram-positive bacterial proteins' dataset

No.	subcellular location	Total protein samples
1	Cell membrane	174
2	Cell wall	18
3	Cytoplasm	208
4	Extracellular	123
Total number of locative proteins		523
Total number of different proteins		519

2.2. Feature Extraction

We extract evolutionary information from Position Specific Scoring Matrix (PSSM) [68] and structural information from the SPIDER 2.0 [70, 69].

PSSM which is produced as the output of PSIBLAST is an scoring matrix that provide substitution probabilities of a given amino acid with other amino acids based on its specific position in a protein[68]. PSSM is a $L \times 20$ matrix, where L is the length of the input protein. Here we use PSIBLAST with three iteration and its cut off value (E) set to 0.001 to produce PSSM .

SPIDER 2.0 (Scoring Protein Interaction Decoys using Exposed Residues) is an accurate method that predict different aspects of local structure such as secondary structure, torsion angle, and Accessible Surface Area (ASA), simultaneously [70, 69]. As an output it produces a $L \times 8$ matrix that include three columns of the probability of contribution of amino acids to each of the secondary structure elements (α-helix, β-strand, coil), one column for ASA, and four columns for the torsion angles (ϕ, ψ, θ, τ) [79, 80]. For the rest of this paper, we will refer to this matrix as SPD3 for simplicity. SPD3 has been recently used in many different fields and demonstrated promising results [81, 82, 83, 84].

Here we have extracted a wide range of features based on different concepts that have been investigate in the literature both for PSSM and SPD3. As a

result the combination of 6 feature groups attained the best result for our task. This might be due to the consistency of these sets of features with each other. However, further investigations in future can potentially provide better understanding of available discriminatory information in these feature groups and consequently provide further prediction enhancement. These 6 feature groups are explained in detail in the following sections.

2.2.1. PSSM-AAO Feature

PSSM-AAO is amino acid occurrence based on PSSM. This feature group is directly extracted from PSSM matrix. It aims at capturing global discriminatory information regarding the substitution probabilities of the amino acids with respect to their positions in the protein sequence [85, 86, 87]. This feature is extracted by summation of the substitution score of a given amino acid with all the amino acids along the protein sequence. The equation for this feature is given below:

$$PSSM - AAO_j = \sum_{i=1}^{L} N_{i,j} \;\; (j = 1, ..., 20) \tag{1}$$

Here N is the corresponding matrix, L is the protein length and j is the respective column. The dimensionality of this feature vector is 20. Algorithm for extracting composition feature is shown in Algorithm 1.

2.2.2. PSSM-SD Feature

This method is specifically proposed to add more local discriminatory information about how the amino acids, based on their substitution probabilities (extracted from PSSM), are distributed along the protein sequence [88]. This method is explained in detail in [71] [5].

Algorithm for extracting PSSM-SD feature is shown in Algorithm 2.

As shown in [5] using $F_p = 25$ gives the best result for this method. As a result, in our final experiment we have adopted $F_p = 25$ which produces 80 $((100 \div 25) \times 20 = 80)$ features.

Algorithm 1: PSSM-AAO Feature Extraction

$N \leftarrow$ PSSM Matrix;
$L \leftarrow$ Length of the Protein;
$C \leftarrow$ Number of matrix column;
$V \leftarrow$ Empty array of size C;
for $j = 0;\ j < C;\ j = j + 1$ **do**
    $sum \leftarrow 0$;
    for $i = 0;\ i < L;\ i = i + 1$ **do**
        $sum = sum + N_{i,j}$;
    end
    $V_j = sum$;
end

2.2.3. PSSM-SAC Feature

This feature was introduced in [5, 89, 86]. It was shown that information about the interaction of neighboring amino acids along the protein sequence can play an important role in providing significant local discriminatory information and enhancing protein subcellular localization prediction accuracy [6, 87]. To extract this information, the concept of auto covariance has been used for different segments of proteins. This is done to enforce local discriminatory information extracted from PSSM. We use the similar approach that was adopted and explained in [5]. We also use the distance factor of 10 as it was also shown in this study as the most effective parameter to extract features for protein subcellular localization.

2.2.4. Auto Covariance of Predicted Secondary Structure

A correlation factor coupling adjacent residues along the protein sequence is known as Auto covariance (AC) [59, 90, 72]. It is also known as a kind of variant

Algorithm 2: PSSM-SD Feature Extraction

```
N ← PSSM Matrix;
L ← Length of the Protein;
C ← Number of matrix column;
F_p ← Desired value of F_p, e.g 5, 10, 25;
V ← Empty array of size (100 ÷ F_p) × C;
k ←  0 ;
for j = 0; j < C; j = j + 1 do
    T_j ← Sum of jth column;
    partialSum ← 0;
    i ← 0;
    for tp = F_p; tp <= 50; tp = tp + F_p do
        while partialSum <= tp × (T_j ÷ 100) do
            partialSum = partialSum + N_{i,j};
            i = i + 1;
        end
        V_k = i;
        k = k + 1;
    end
    partialSum ← 0;
    i ← L;
    index ← 0;
    for tp = F_p; tp <= 50; tp = tp + F_p do
        while partialSum <= tp × (T_j ÷ 100) do
            partialSum = partialSum + N_{i,j};
            i = i − 1;
            index = index + 1;
        end
        V_k = index;
        k = k + 1;
    end
end
```

of auto cross covariance. It is a very powerful statistical tool which is used to analyze sequences of vectors [91]. The Auto Covariance transformation has been

widely applied in various fields of bioinformatics [92], [93], [94], [95], [96], [97]. Auto Covariance variables are able to avoid producing too many variants. The equation for this feature is given below:

$$AutoCovariance_{k,j} = \frac{1}{L}\sum_{i=1}^{L-k} N_{i,j}N_{i+k,j} \;\; (j = 1, ..., 20 \; and \; k = 1...DF) \quad (2)$$

where DF is the distance factor. Different values have been tested to find out the effective value of DF which gives the highest accuracy rate of prediction. In this research we have tested total 15 values for DF (DF = 1,2,3,4,.....,12,13,14,15) and took only one value which is DF = 10. We have observed that DF = 10 gives the highest accuracy rate for this task. So, the effective value of DF is used as 10 for the employed benchmark. The dimensionality of this feature vector will be $(Number\ of\ columns) \times DF$. Since we are using this method to extract features based on the predicted secondary structure which consists of three columns in SPD3, we will have 30 features in total. Algorithm for extracting auto covariance feature is shown at Algorithm 3.

2.2.5. Composition of Torsion Angles

This feature is extracted from the Spider SPD3. Torsion angles are shown as effectives components to capture continuous information based on the secondary structure of proteins [79, 80]. To calculate this feature we have taken columns corresponding to torsion angles one at the time, summed up all the values and finally divided them by L. The equation for this method is given below:

$$Composition_j = \frac{1}{L}\sum_{i=1}^{L} N_{i,j} \quad (3)$$

Here N is the corresponding matrix, L is the protein length and j is the respective column. The dimensionality of this feature vector will be $(Number\ of\ columns)$ which is four for our case corresponding to four torsion angles. Algorithm for extracting this feature is shown in Algorithm 4.

Algorithm 3: Auto Covariance Feature Extraction

$DF \leftarrow 10$;
$P \leftarrow$ Matrix from which feature will be extracted;
$L \leftarrow$ Length of the Protein;
$V \leftarrow$ Empty array of size $L\times$ (Number of matrix column);
$C \leftarrow$ Number of matrix column;
for $k = 0;\ k < DF;\ k = k + 1$ **do**
  for $j = 0;\ j < C;\ j = j + 1$ **do**
    $sum \leftarrow 0$;
    for $i = 0;\ i < L - k;\ i = i + 1$ **do**
      $sum = sum + P_{i,j}P_{i+k,j}$;
    end
    $V_{k,j} = \frac{sum}{L}$;
  end
end

2.2.6. One-Lead Bi-gram of ASA

We extract this feature based on the Bi-gram concept that have been previously used in [87, 60, 6]. Accessible Surface Area (ASA) can provide important information on the locality of neighboring amino acids in the proteins 3D structure [98]. We adopt this method to extract one-lead Bi-gram for the ASA. Algorithm for extracting one-lead Bi-gram feature is shown in Algorithm 5.

The equation for this feature is given below:

$$OneLeadBigram_{k,l} = \frac{1}{L}\sum_{i=1}^{L-2} N_{i,k}N_{i+2,l} \tag{4}$$

The dimensionality of this feature vector will be:

$(Number\ of\ columns) \times (Number\ of\ columns)$.

With the explosive growth of biological sequences in the post-genomic era,

Algorithm 4: Composition Feature Extraction

$N \leftarrow$ Matrix from which feature will be extracted;
$L \leftarrow$ Length of the Protein;
$C \leftarrow$ Number of matrix column;
$V \leftarrow$ Empty array of size C;
for $j = 0;\ j < C;\ j = j + 1$ **do**
    $sum \leftarrow 0$;
    for $i = 0;\ i < L;\ i = i + 1$ **do**
        $sum = sum + N_{i,j}$;
    end
    $V_j = \frac{sum}{L}$;
end

one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a comprehensive review [33]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition or PseAAC [45] was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics [99]. Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder', 'propy', and 'PseAAC-General', were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC [67], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode, "Gene Ontology" mode, and "Sequential Evolution" or "PSSM" mode [67].

Algorithm 5: One-Lead Bi-gram Feature Extraction

$N \leftarrow$ Matrix from which feature will be extracted;
$L \leftarrow$ Length of the Protein;
$C \leftarrow$ Number of matrix column;
$V \leftarrow$ Empty array of size $C \times C$;
for $k = 0;\ k < C;\ k = k + 1$ **do**
    for $l = 0;\ l < C;\ l = l + 1$ **do**
        $sum \leftarrow 0$;
        for $i = 0;\ i < L - 2;\ i = i + 1$ **do**
            $sum = sum + N_{i,k} N_{i+2,l}$;
        end
        $V_{k,l} = \dfrac{sum}{L}$;
    end
end

Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC [100] was developed for generating various feature vectors for DNA/RNA sequences that have proved to be very useful as well [66, 67, 101, 57, 99, 100, 102]. Particularly, recently a very powerful web-server called 'Pse-in-One' [101] and its updated version 'Pse-in-One2.0' [57] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to users' need or defined by users' own. In the current study, we are to use the six features extracted from the PSSM and SPIDER to formulate the protein sequences for predicting their subcellular localization.

2.3. Support Vector Machine

SVM is considered to be one of the best pattern recognition techniques [103]. It is also widely used in Bioinformatics and has outperformed other classifiers and obtained promising results for protein subcellular localization. It aims to

reduce the prediction error rate by finding the hyperplane that produces the largest margin based on the concept of support vector theory. It transforms the input data to higher dimensions using the kernel function to be able to find support vectors (for non linear cases). The classification of some known points in input space x_i is y_i which is defined to be either -1 or $+1$. If x' is a point in input space with unknown classification then:

$$y' = \sin(\sum_{i=1}^{n} a_i y_i K(x_i, x') + b) \tag{5}$$

where y' is the predicted class of point x'. The function $K()$ is the kernel function, n is the number of support vectors and a_i are adjustable weights and b is the bias. In this study, the SVM classifier is implemented with the LIBSVM toolbox using the Radial Basis Function (RBF) as its kernel [104]. RBF kernel is adopted in our experiments due to its better performance than other kernels functions (e.g. polynomial kernel, linear kernel, and sigmoid). RBF kernel is defined as follows:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

where γ is the regularization parameter, x_i and x_j are input feature vectors. In this study, the γ in addition to the cost parameter C (also called the soft margin parameter) are optimized using grid search algorithm which is also implemented in the LIBSVM package. Despite its simplicity, grid search has been shown to be an effective method to optimize these parameters. We tuned those parameter using grid search implemented in LIBSVM. As a result we used Cost parameter $(C) = 3000$, and $\gamma = 0.005$.

3. VALIDATION METHOD

For our experiment we have adopted two types of validation method namely, 10-fold cross validation and jackknife (also known as leave-one-out) cross validation.

10-Fold Cross Validation: in 10-fold cross-validation, the original sample set is randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimation. The advantage of this method is that all samples are used for both training and validation.

Jackknife Test: in this method the original sample is randomly partitioned into n subsamples where n is the total number of samples in the dataset. Of the n subsamples a single subsample (means exactly one sample from the sample dataset) is retained as the validation data for testing the model, and the remaining $n-1$ subsamples are used as training data. The cross-validation process is then repeated n times, with each of the n subsamples used exactly once as the validation data. The n results from the folds can then be combined to produce a single estimation. The advantage of this method is that all observations are used for both training and validation and each observation is used for validation exactly once. It can help to build a more general and robust method. The main disadvantage of jackknife test is it takes more time to complete a full training and testing process.

4. EVALUATION MEASUREMENT

Here we use Sensitivity, Specificity, Matthew's Correlation Coefficient (MCC) and accuracy to provide more information about the statistical significance of our achieved results. Sensitivity, specificity, MCC and accuracy are statistical measures of the performance of a binary classification test, also known in statistics as classification function. These have been widely used in the literature for

protein subcellular localization [48, 105, 41]. **Sensitivity** (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified. **Specificity** (also called the true negative rate) measures the proportion of negatives that are correctly identified. The value of sensitivity and specificity varies between 0 and 1. Having specificity, and sensitivity equal to 1 represents a fully accurate model while 0 represents a fully inaccurate. On the other hand, **MCC** measures the prediction quality of the model. MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The **accuracy** refers to the total correctly classified instances over the number of samples present in the dataset. The equation for calculating sensitivity, specificity, MCC and accuracy are given below:

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{7}$$

$$MCC = \frac{(TN \times TP) - (TN \times FP)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100 \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100 \tag{9}$$

where TP is the number of correctly identified (true positive) samples, FN is the number of incorrectly rejected samples (false negative), TN is the number of correctly rejected (true negative) samples, and FP is the number of incorrectly accepted samples (false positive).

This set of metrics is valid only for the single-label systems (in which each protein only belongs to one and only class). For the multi-label systems (in which a protein might belong to several classes), whose existence has become more frequent in system biology [25, 12] and system medicine [106, 107] and

biomedicine [108], a completely different set of metrics as defined in [109] is needed. This set of metrics widely used in the literature [110, 111, 112, 113, 64, 114, 115, 63, 116, 117, 65, 57] . Also, the rigorous metrics used in here to examine the quality of a new predictor for a multi-label system was taken from [109].

5. RESULTS AND DISCUSSION

In this section, we present the results of the experiments that were carried in this study. All the methods were implemented in Python. Each of the experiments were carried 10 times and only the average is reported as results. The general architecture of our method is shown in Figure 1.

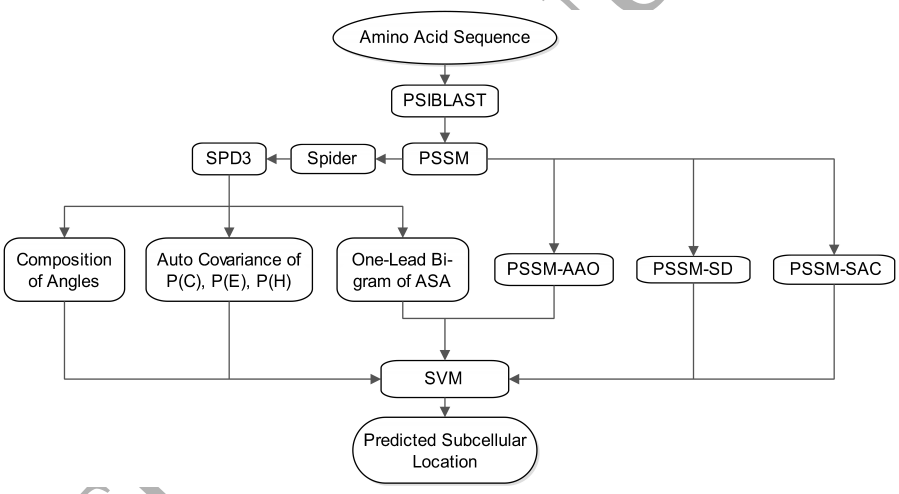

Figure 1: The general architecture of EvoStruct-Sub

5.1. Single Label Classification

In this paper we first calculate the single label classification results. For single label classification we calculate two types of accuracy, one is overall accuracy and another one is average accuracy. To calculate overall accuracy we

Table 2: Comparison of the results achieved for single label classification

	Gram-positive benchmark accuracy		
	Overall		Average
	10-Fold test	Jackknife test	Jackknife test
Huang and Yuan [74]	83.7	-	-
Pacharawongsakda et al., [78]	-	-	-
Dehzangi et al., [118]	83.6	-	-
Dehzangi et al., [5]	87.7	88.2	-
Sharmaet al., [6]	84.3	85	89.8
This Paper	-	**91.01**	**95.4**

use **sensitivity** and to calculate average accuracy we use **average accuracy**. Average accuracy is computed as follows:

$$Average\ Accuracy = \frac{1}{n}\sum_{j=1}^{n} accuracy_j$$

where n is the number of classes in the dataset. A comparison of single label classification result is shown in Table 2.

As it is shown in here, EvoStruct-Sub achieves to over 90% for overall and 95.0% average prediction accuracy. In overall, EvoStruct-Sub achieves 91.01% prediction accuracy which is 6.29% better than the best result reported in the literature for this task.

5.2. Multi Label Classification

Since our employed benchmark contains multi labeled proteins, besides single label classification, we also perform multi-label classification. For calculating multi-label classification result, we use overall locative accuracy and overall absolute accuracy. The overall locative accuracy and overall absolute accuracy are defined as follows:

$$Locative\ Accuracy = \frac{1}{N_{dif}}\sum_{i=1}^{N_{dif}} Z_i \tag{10}$$

Table 3: Comparison of the results achieved for multi label classification

	Gram-positive benchmark	
	Locative Accuracy	Absolute Accuracy
Sharma et al., [6]	84.8	85.16
This Paper	91.71	90.94

$$Absolute\ Accuracy = \frac{1}{N_{dif}} \sum_{i=1}^{N_{dif}} C_i \tag{11}$$

where N_{loc} is the number of locative proteins, $Z_i = 1$ if at least one subcellular locations of the $i-th$ protein are correctly predicted and 0 otherwise, $C_i = 1$ if all the subcellular locations of query protein are exactly predicted and 0 otherwise. Therefore the overall absolute accuracy is striker than overall locative accuracy [34]. The results achieved for EvoStruct-Sub compared to the best result reported in the literature [6] is shown in Table 3.

As shown in this table, EvoStruct-Sub achieves to over 90% prediction accuracy for locative and absolute methods. EvoStruct-Sub achieves 91.71% and 90.4% prediction accuracies which are over 6.91% and 5.78% better than those reported in [6], respectively.

5.3. Investigating the Impact of Proposed Features on the Achieved Results

Here we investigate the impact of each individual feature group that we proposed in this study in two steps. We first combine features extracted from the PSSM one by one and record the results. We then combine features extracted from SPD3 one by one and again record the results. We finally add the features extracted from SPD3 one by one to the features extracted from PSSM. In this way, we can investigate the impact of our extracted features based on their sources and how they impact on the prediction performance. This comparison is shown in Table 4. As it is shown in Table 4, in general, features extracted

Table 4: Investigating the impact of features extracted from PSSM, SPD3, and their combinations

	Average		
	Sensitivity	Specificity	MCC
PSSM-AAO	0.62	0.91	0.36
PSSM-AAO, PSSM-SD	0.75	0.94	0.60
PSSM-AAO, PSSM-SD, PSSM-SAC	0.81	0.95	0.62
Auto-Covariance of P(C) P(E) P(H)	0.48	0.86	0.16
Auto-Covariance of P(C) P(E) P(H), Composition-Angles	0.48	0.86	0.10
Auto-Covariance of P(C) P(E) P(H), Composition-Angles, One-Lead Bi-gram ASA	0.49	0.87	0.12
PSSM-AAO, PSSM-SD, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H)	0.81	0.96	0.61
PSSM-AAO, PSSM-SD, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H), Composition-Angles	0.81	0.96	0.60
PSSM-AAO, PSSM-SD, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H), Composition-Angles, One-Lead Bi-gram ASA	0.82	0.96	0.64

from PSSM provide better performance than SPD3. However, the best results achieve by adding the SPD3-based features to PSSM-based features. It highlights the incremental impact of structural features extracted from SPD3 on the achieved results and on enhancing the protein subcellular localization prediction performance.

We then investigate the impact of each of our proposed feature groups individually on the achieved results. To do this, we exclude each of feature group from the combination of features one at the time. In other words, we exclude each one of our feature groups which leave us with the combination of 5 remaining feature groups. The result for this experiments are demonstrated in Table 5. As it is shown in Table 5, we still can achieve very good results using the combination of 5 feature groups. However, none of those combinations achieve to the results of using all 6 feature groups at the time. In other words, incorporation of all 6 proposed feature groups is vital to enhance protein subcellular localization prediction problem.

Table 5: Investigating the impact of each individual feature group on our achieved results.

	Average		
	Sensitivity	Specificity	MCC
PSSM-AAO, PSSM-SD, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H), Composition-Angles	0.81	0.96	0.60
PSSM-AAO, PSSM-SD, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H), One-Lead Bi-gram ASA	0.82	0.96	0.59
PSSM-AAO, PSSM-SD, PSSM-SAC, Composition-Angles, One-Lead Bi-gram ASA	0.81	0.96	0.69
PSSM-AAO, PSSM-SD, Auto-Covariance of P(C) P(E) P(H), Composition-Angles, One-Lead Bi-gram ASA	0.73	0.95	0.46
PSSM-AAO, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H), Composition-Angles, One-Lead Bi-gram ASA	0.63	0.92	0.27
PSSM-SD, PSSM-SAC, Auto-Covariance of P(C) P(E) P(H), Composition-Angles, One-Lead Bi-gram ASA	0.76	0.95	0.54

6. CONCLUSION

In this study, we have proposed EvoStruct-Sub for predicting Gram-positive bacterial protein subcellular localization. To build EvoStruct-Sub we have extracted a wide range of features from PSSM and SPD3 and among them selected 6 features with total feature vector size of 235. We then used SVM to our extracted features for the classification task. As our benchmark is multi-label dataset, so we have reported both single label and multi label prediction accuracies. For single label classification our reported result is 91.01% and for multi label classification our reported result for locative accuracy is 91.71% and for absolute accuracy is 90.94%. These results in all cases are up to 6% better than previously reported results in the literature for Gram-positive subcellular localization. These enhancements highlight the effectiveness of EvoStruct-Sub to explore the potential information embedded in the PSSM and SPD3 for Gram-positive subcellular localization prediction problem.

As pointed out in [119] and demonstrated in a series of recent publications (see, e.g., [116, 117]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have

increasing impacts on medical science, driving medicinal chemistry into an unprecedented revolution [99], we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper.

7. ACKNOWLEDGEMENTS

We would like to thank Professor Kuo-Chen Chou for sharing Gram-positive and Gram-negative protein subcellular localizations benchmarks which are introduced in **Cell-PLoc 2.0** package.

References

[1] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2004) 10–19.

[2] K.-C. Chou, H.-B. Shen, Recent progress in protein subcellular location prediction, Analytical biochemistry 370 (1) (2007) 1–16.

[3] Z.-C. Wu, X. Xiao, K.-C. Chou, iloc-gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins, Protein and peptide letters 19 (1) (2012) 4–14.

[4] H.-B. Shen, K.-C. Chou, Gpos-mploc: a top-down approach to improve the quality of predicting subcellular localization of gram-positive bacterial proteins, Protein and peptide letters 16 (12) (2009) 1478–1484.

[5] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into chou s general pseaac, Journal of theoretical biology 364 (2015) 284–294.

[6] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda, A. Sharma, Predict gram-positive and gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into

chou's general pseaac, IEEE Transactions on NanoBioscience 14 (8) (2015) 915–926.

[7] H. Saini, G. Raicar, A. Dehzangi, S. Lal, A. Sharma, Subcellular localization for gram positive and gram negative bacterial proteins using linear interpolation smoothing model, Journal of theoretical biology 386 (2015) 25–33.

[8] S. Shatabda, S. Saha, A. Sharma, A. Dehzangi, iphloc-es: Identification of bacteriophage protein locations using evolutionary and structural features, Journal of theoretical biology 435 (2017) 229–237.

[9] A. Sharma, K. K. Paliwal, S. Imoto, S. Miyano, A feature selection method using improved regularized linear discriminant analysis, Machine vision and applications 25 (3) (2014) 775–786.

[10] A. Sharma, A. Dehzangi, J. Lyons, S. Imoto, S. Miyano, K. Nakai, A. Patil, Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function, PloS one 9 (2) (2014) e89890.

[11] X.-X. Chen, H. Tang, W.-C. Li, H. Wu, W. Chen, H. Ding, H. Lin, Identification of bacterial cell wall lyases via pseudo amino acid composition, BioMed Research International 2016.

[12] X. Cheng, X. Xiao, K.-C. Chou, ploc-meuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key go information into general pseaac, Genomics.

[13] G.-L. Fan, Q.-Z. Li, Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of chous pseudo amino acid composition, Journal of Theoretical Biology 304 (2012) 88–95.

[14] S. Mei, Multi-kernel transfer learning based on chou's pseaac formulation for protein submitochondria localization, Journal of Theoretical Biology 293 (2012) 121–130.

[15] S. Mei, Predicting plant protein subcellular multi-localization by chou's pseaac formulation based multi-label homolog knowledge transfer learning, Journal of theoretical biology 310 (2012) 80–87.

[16] S. Wan, M.-W. Mak, S.-Y. Kung, Goasvm: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of chou's pseudo-amino acid composition, Journal of Theoretical Biology 323 (2013) 40–48.

[17] Y.-S. Jiao, P.-F. Du, Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into chou's general pseudo-amino acid compositions, Journal of theoretical biology 416 (2017) 81–87.

[18] B. Yu, S. Li, W.-Y. Qiu, C. Chen, R.-X. Chen, L. Wang, M.-H. Wang, Y. Zhang, Accurate prediction of subcellular location of apoptosis proteins combining chous pseaac and psepssm based on wavelet denoising, Oncotarget 8 (64) (2017) 107640.

[19] S. Zhang, X. Duan, Prediction of protein subcellular localization with oversampling approach and chou's general pseaac, Journal of theoretical biology 437 (2018) 239–250.

[20] X. Cheng, X. Xiao, K.-C. Chou, ploc-mplant: predict subcellular localization of multi-location plant proteins by incorporating the optimal go information into general pseaac, Molecular BioSystems 13 (9) (2017) 1722–1727.

[21] X. Cheng, X. Xiao, K.-C. Chou, ploc-mvirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal go information into general pseaac, Gene 628 (2017) 315–321.

[22] X. Cheng, X. Xiao, K.-C. Chou, ploc-mgneg: Predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general pseaac, Genomics.

[23] X. Cheng, X. Xiao, K.-C. Chou, ploc-mhum: predict subcellular localization of multi-location human proteins via general pseaac to winnow out the crucial go information, Bioinformatics 1 (2017) 9.

[24] X. Cheng, S.-G. Zhao, W.-Z. Lin, X. Xiao, K.-C. Chou, ploc-manimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (22) (2017) 3524–3531.

[25] X. Xiao, X. Cheng, S. Su, Q. Mao, K.-C. Chou, ploc-mgpos: incorporate key gene ontology information into general pseaac for predicting subcellular localization of gram-positive bacterial proteins, Natural Science 9 (09) (2017) 330.

[26] X. Cheng, X. Xiao, K.-C. Chou, ploc-meuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key go information into general pseaac, Genomics.

[27] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. Adams-Collier, K. Nakai, Wolf psort: protein localization predictor, Nucleic acids research 35 (suppl_2) (2007) W585–W587.

[28] H. Chen, N. Huang, Z. Sun, Subloc: a server/client suite for protein subcellular location based on soap, Bioinformatics.

[29] O. Emanuelsson, H. Nielsen, S. Brunak, G. Von Heijne, Predicting subcellular localization of proteins based on their n-terminal amino acid sequence, Journal of molecular biology 300 (4) (2000) 1005–1016.

[30] A. Pierleoni, P. L. Martelli, P. Fariselli, R. Casadio, Bacello: a balanced subcellular localization predictor, Bioinformatics 22 (14) (2006) e408–e416.

[31] T. Tamura, T. Akutsu, Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition, BMC bioinformatics 8 (1) (2007) 466.

[32] J.-M. Chang, E. C.-Y. Su, A. Lo, H.-S. Chiu, T.-Y. Sung, W.-L. Hsu, Psldoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis, Proteins: Structure, Function, and Bioinformatics 72 (2) (2008) 693–710.

[33] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Medicinal chemistry 11 (3) (2015) 218–234.

[34] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Medicinal chemistry 11 (3) (2015) 218–234.

[35] K.-C. Chou, H.-B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mploc 2.0, PLoS One 5 (4) (2010) e9931.

[36] H.-B. Shen, K.-C. Chou, A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mploc 2.0, Analytical biochemistry 394 (2) (2009) 269–274.

[37] K.-C. Chou, H.-B. Shen, Plant-mploc: a top-down strategy to augment the power for predicting plant protein subcellular localization, PloS one 5 (6) (2010) e11335.

[38] H.-B. Shen, K.-C. Chou, Gneg-mploc: a top-down strategy to enhance the quality of predicting subcellular localization of gram-negative bacterial proteins, Journal of theoretical biology 264 (2) (2010) 326–333.

[39] H.-B. Shen, K.-C. Chou, Virus-mploc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites, Journal of Biomolecular Structure and Dynamics 28 (2) (2010) 175–186.

[40] X. Xiao, Z.-C. Wu, K.-C. Chou, A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites, PloS one 6 (6) (2011) e20592.

[41] K.-C. Chou, Z.-C. Wu, X. Xiao, iloc-hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Molecular Biosystems 8 (2) (2012) 629–641.

[42] Z.-C. Wu, X. Xiao, K.-C. Chou, iloc-plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites, Molecular BioSystems 7 (12) (2011) 3287–3297.

[43] W.-Z. Lin, J.-A. Fang, X. Xiao, K.-C. Chou, iloc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Molecular BioSystems 9 (4) (2013) 634–644.

[44] X. Xiao, Z.-C. Wu, K.-C. Chou, iloc-virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, Journal of Theoretical Biology 284 (1) (2011) 42–51.

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

[46] X. Wang, H. Li, Q. Zhang, R. Wang, Predicting subcellular localization of apoptosis proteins combining go features of homologous proteins and distance weighted knn classifier, BioMed research international 2016.

[47] Y. Xu, X. Wang, Y. Wang, Y. Tian, X. Shao, L.-Y. Wu, N. Deng, Prediction of posttranslational modification sites from amino acid sequences with kernel methods, Journal of theoretical biology 344 (2014) 78–87.

[48] W. Chen, H. Lin, Recent advances in identification of rna modifications, Non-Coding RNA 3 (1) (2016) 1.

[49] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, K.-C. Chou, itis-psetnc: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Analytical biochemistry 462 (2014) 76–83.

[50] L. Nanni, A. Lumini, Genetic programming for creating chous pseudo amino acid based features for submitochondria localization, Amino acids 34 (4) (2008) 653–660.

[51] E. Altman, K. Young, J. Garrett, R. Altman, R. Young, Subcellular localization of lethal lysis proteins of bacteriophages lambda and phix174., Journal of virology 53 (3) (1985) 1008–1011.

[52] I. Dubchak, I. B. Muchnik, S.-H. Kim, Protein folding class predictor for scop: approach based on global descriptors., in: Ismb, 1997, pp. 104–107.

[53] Z.-C. Li, X.-B. Zhou, Z. Dai, X.-Y. Zou, Prediction of protein structural classes by chous pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis, Amino acids 37 (2) (2009) 415.

[54] Y. Liu, W. Gu, W. Zhang, J. Wang, Predict and analyze protein glycation sites with the mrmr and ifs methods, BioMed research international 2015.

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

[56] M. Rahimi, M. R. Bakhtiarizadeh, A. Mohammadi-Sangcheshmeh, Oogenesis_pred: A sequence-based method for predicting oogenesis proteins by six different modes of chou's pseudo amino acid composition, Journal of theoretical biology 414 (2017) 128–136.

[57] B. Liu, H. Wu, K.-C. Chou, Pse-in-one 2.0: An improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences, Natural Science 9 (04) (2017) 67.

[58] A. Dehzangi, S. Karamizadeh, Solving protein fold prediction problem using fusion of heterogeneous classifiers, International Information Institute (Tokyo). Information 14 (11) (2011) 3611.

[59] A. Dehzangi, K. Paliwal, A. Sharma, O. Dehzangi, A. Sattar, 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 10 (3) (2013) 564–575.

[60] K. K. Paliwal, A. Sharma, J. Lyons, A. Dehzangi, Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information, BMC bioinformatics 15 (S16) (2014) S12.

[61] A. Dehzangi, S. Phon-Amnuaisuk, O. Dehzangi, Enhancing protein fold prediction accuracy by using ensemble of different classifiers, Australian Journal of Intelligent Information Processing Systems 26 (4) (2010) 32–40.

[62] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, J.-H. Jia, K.-C. Chou, ikcr-pseens: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics.

[63] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, K.-C. Chou, irna-2methyl: Identify rna 2'-o-methylation sites by incorporating sequence-coupled effects into general pseknc and ensemble classifier, Medicinal Chemistry 13 (8) (2017) 734–743.

[64] B. Liu, F. Yang, K.-C. Chou, 2l-pirna: A two-layer ensemble classifier for identifying piwi-interacting rnas and their function, Molecular Therapy-Nucleic Acids 7 (2017) 267–277.

[65] B. Liu, F. Yang, D.-S. Huang, K.-C. Chou, ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc, Bioinformatics 34 (1) (2017) 33–40.

[66] A. Ehsan, K. Mahmood, Y. Khan, S. Khan, K. Chou, A novel modeling in mathematical biology for classification of signal peptides, Scientific Reportsdoi:10.1038/s41598-018-19491-y.

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

[68] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic acids research 25 (17) (1997) 3389–3402.

[69] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific reports 5 (2015) 11476.

[70] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Zhou, Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks, Prediction of Protein Secondary Structure (2017) 55–63.

[71] A. Dehzangi, A. Sharma, J. Lyons, K. K. Paliwal, A. Sattar, A mixture of physicochemical and evolutionary–based feature extraction approaches for protein fold recognition, International journal of data mining and bioinformatics 11 (1) (2014) 115–138.

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

[73] G. Huang, Y. Zhou, Y. Zhang, B.-Q. Li, N. Zhang, Y.-D. Cai, Prediction of carbamylated lysine sites based on the one-class k-nearest neighbor method, Molecular BioSystems 9 (11) (2013) 2729–2740.

[74] C. Huang, J. Yuan, Using radial basis function on the general form of chou's pseudo amino acid composition and pssm to predict subcellular lo-

cations of proteins with both single and multiple sites, Biosystems 113 (1) (2013) 50–57.

[75] K.-C. Chou, H.-B. Shen, et al., Cell-ploc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science 2 (10) (2010) 1090.

[76] K.-C. Chou, H.-B. Shen, Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nature protocols 3 (2) (2008) 153–162.

[77] H.-B. Shen, K.-C. Chou, Gpos-ploc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins, Protein Engineering Design and Selection 20 (1) (2007) 39–46.

[78] E. Pacharawongsakda, T. Theeramunkong, Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of chou's pseaac, IEEE transactions on nanobioscience 12 (4) (2013) 311–320.

[79] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, Y. Yang, Predicting backbone cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network, Journal of computational chemistry 35 (28) (2014) 2040–2046.

[80] R. Heffernan, A. Dehzangi, J. Lyons, K. Paliwal, A. Sharma, J. Wang, A. Sattar, Y. Zhou, Y. Yang, Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins, Bioinformatics (2015) btv665.

[81] G. Taherzadeh, Y. Yang, T. Zhang, A. W.-C. Liew, Y. Zhou, Sequence-based prediction of protein–peptide binding sites using support vector machine, Journal of computational chemistry.

[82] G. Taherzadeh, Y. Zhou, A. W.-C. Liew, Y. Yang, Sequence-based prediction of protein–carbohydrate binding sites using support vector machines, Journal of chemical information and modeling 56 (10) (2016) 2115–2122.

[83] Y. López, A. Dehzangi, S. P. Lal, G. Taherzadeh, J. Michaelson, A. Sattar, T. Tsunoda, A. Sharma, Sucstruct: Prediction of succinylated lysine residues by using structural properties of amino acids, Analytical Biochemistry.

[84] S. Y. Chowdhury, S. Shatabda, A. Dehzangi, Idnaprot-es: Identification of dna-binding proteins using evolutionary and structural features, Scientific Reports 7 (2017) 14938.

[85] Y. Taguchi, M. M. Gromiha, Application of amino acid occurrence for discriminating different folding types of globular proteins, BMC bioinformatics 8 (1) (2007) 404.

[86] A. Dehzangi, K. K. Paliwal, A. Sharma, J. G. Lyons, A. Sattar, Protein fold recognition using an overlapping segmentation approach and a mixture of feature extraction models., in: Australasian Conference on Artificial Intelligence, Springer, 2013, pp. 32–43.

[87] A. Sharma, J. Lyons, A. Dehzangi, K. K. Paliwal, A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition, Journal of theoretical biology 320 (2013) 41–46.

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

[89] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, A. Sattar, Enhancing protein fold prediction accuracy using evolutionary and structural features, in: IAPR International Conference on Pattern Recognition in Bioinformatics, Springer, 2013, pp. 196–207.

[90] A. Dehzangi, A. Sattar, Ensemble of diversely trained support vector machines for protein fold recognition., in: ACIIDS (1), 2013, pp. 335–344.

[91] S. Wold, J. Jonsson, M. Sjörström, M. Sandberg, S. Rännar, Dna and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures, Analytica Chimica Acta 277 (2) (1993) 239–253.

[92] Y. Guo, M. Li, M. Lu, Z. Wen, Z. Huang, Predicting g-protein coupled receptors–g-protein coupling specificity based on autocross-covariance transform, Proteins: structure, function, and bioinformatics 65 (1) (2006) 55–60.

[93] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences, Nucleic acids research 36 (9) (2008) 3025–3030.

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

[95] J. Wu, M.-L. Li, L.-Z. Yu, C. Wang, An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition, The protein journal 29 (1) (2010) 62–67.

[96] Y.-h. Zeng, Y.-z. Guo, R.-q. Xiao, L. Yang, L.-z. Yu, M.-l. Li, Using the augmented chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, Journal of theoretical biology 259 (2) (2009) 366–372.

[97] T. Liu, X. Zheng, C. Wang, J. Wang, Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation, Protein and peptide letters 17 (10) (2010) 1263–1269.

[98] G. Taherzadeh, Y. Zhou, A. W.-C. Liew, Y. Yang, Structure-based prediction of protein–peptide binding regions using random forest, Bioinformatics.

[99] K.-C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Current topics in medicinal chemistry 17 (21) (2017) 2337–2358.

[100] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, K.-C. Chou, Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition, Analytical biochemistry 456 (2014) 53–60.

[101] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.-C. Chou, Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences, Nucleic acids research 43 (W1) (2015) W65–W71.

[102] W. Chen, H. Lin, K.-C. Chou, Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic sequences, Molecular BioSystems 11 (10) (2015) 2620–2634.

[103] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.

[104] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 27.

[105] X. Cheng, X. Xiao, K.-C. Chou, ploc-meuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key go information into general pseaac, Genomicsdoi:doi:10.1016/j.ygeno.2017.08.

[106] X. Cheng, S.-G. Zhao, X. Xiao, K.-C. Chou, iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2016) 341–346.

[107] X. Cheng, S.-G. Zhao, X. Xiao, K.-C. Chou, iatc-mhyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (35) (2017) 58494.

[108] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, K.-C. Chou, iptm-mlys: identifying multiple lysine ptm sites and their different types, Bioinformatics 32 (20) (2016) 3116–3123.

[109] K.-C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Molecular Biosystems 9 (6) (2013) 1092–1100.

[110] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, isno-aapair: incorporating amino acid pairwise coupling into pseaac for predicting cysteine s-nitrosylation sites in proteins, PeerJ 1 (2013) e171.

[111] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, irspot-psednc: identify recombination spots with pseudo dinucleotide composition, Nucleic acids research 41 (6) (2013) e68–e68.

[112] W. Chen, H. Tang, J. Ye, H. Lin, K.-C. Chou, irna-pseu: Identifying rna pseudouridine sites, Molecular TherapyNucleic Acids 5 (7) (2016) e332.

[113] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.-C. Chou, irna-psecoll: Identifying the occurrence sites of different rna modifications by incorporating collective effects of nucleotides into pseknc, Molecular Therapy-Nucleic Acids 7 (2017) 155–163.

[114] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.-C. Chou, irna-ai: identifying the adenosine to inosine editing sites in rna sequences, Oncotarget 8 (3) (2017) 4208.

[115] B. Liu, S. Wang, R. Long, K.-C. Chou, irspot-el: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2016) 35–41.

[116] Y. Xu, Z. Wang, C. Li, K.-C. Chou, ipreny-pseaac: identify c-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into pseaac, Medicinal Chemistry 13 (6) (2017) 544–551.

[117] L.-M. Liu, Y. Xu, K.-C. Chou, ipgk-pseaac: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general pseaac, Medicinal Chemistry 13 (6) (2017) 552–559.

[118] A. Dehzangi, S. Sohrabi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features, BMC bioinformatics 16 (4) (2015) S1.

[119] K.-C. Chou, H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes, Natural Science 1 (02) (2009) 63.