



ELSEVIER

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC

Abdollah Dehzangi^{a,b,*}, Rhys Heffernan^c, Alok Sharma^{a,d}, James Lyons^c,
Kuldip Paliwal^c, Abdul Sattar^{a,b}

^a Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia

^b National ICT Australia (NICTA), Brisbane, Australia

^c School of Engineering, Griffith University, Brisbane, Australia

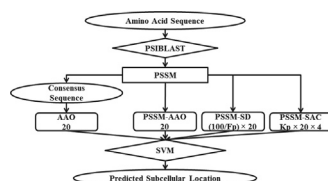
^d School of Engineering and Physics, University of the South Pacific, Fiji

HIGHLIGHTS

- Protein subcellular localization is defined as predicting the functioning location of a given protein in the cell.
- It is considered an important step towards protein function prediction and drug design.
- Propose segmentation distribution and segmented auto-covariance feature extraction methods to explore local evolutionary-based information.
- Use consensus sequence-based and semi-occurrence to extract global evolutionary-based information.
- Use Support Vector Machine (SVM) as the classification technique.
- Enhance Gram-positive and Gram-negative subcellular localization prediction accuracies by up to 6.4%.

GRAPHICAL ABSTRACT

AAO: Consensus Sequence-based Occurrence; PSSM-AAO: Evolutionary-based occurrence (Semi-occurrence); PSSM-SD: Segmented Distribution (Fp: distribution factor); PSSM-SAC: Segmented Auto-covariance (Kp: distance factor).



AAO: Consensus Sequence-based Occurrence
PSSM-AAO: Evolutionary-based occurrence (Semi-occurrence)
PSSM-SD: Segmented Distribution (Fp: distribution factor)
PSSM-SAC: Segmented Auto-covariance (Kp: distance factor)

ARTICLE INFO

Article history:

Received 15 May 2014

Received in revised form

11 August 2014

Accepted 17 September 2014

Available online 28 September 2014

Keywords:

Evolutionary-based features
Segmented autocorrelation
Segmented distribution
Support Vector Machine (SVM)

ABSTRACT

Protein subcellular localization is defined as predicting the functioning location of a given protein in the cell. It is considered an important step towards protein function prediction and drug design. Recent studies have shown that relying on Gene Ontology (GO) for feature extraction can improve protein subcellular localization prediction performance. However, relying solely on GO, this problem remains unsolved. At the same time, the impact of other sources of features especially evolutionary-based features has not been explored adequately for this task. In this study, we aim to extract discriminative evolutionary features to tackle this problem. To do this, we propose two segmentation based feature extraction methods to explore potential local evolutionary-based information for Gram-positive and Gram-negative subcellular localizations. We will show that by applying a Support Vector Machine (SVM) classifier to our extracted features, we are able to enhance Gram-positive and Gram-negative subcellular

* Corresponding author at: P.O. 4111, Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane, Australia.

Tel.: +61 7 3735 5381; fax: +61 7 3735 4066.

E-mail address: a.dehzangi@griffith.edu.au (A. Dehzangi).

localization prediction accuracies by up to 6.4% better than previous studies including the studies that used GO for feature extraction.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Most proteins can only function in one specific place in the cell (e.g. nucleus, membrane) while some other proteins can function in several places in the cell (Chou, 2013). For a given protein, to function properly, it needs to be in one or a few specific locations in the cell as it malfunctions in all other places (Emanuelsson et al., 2000). Therefore, protein subcellular localization prediction is considered an important step towards protein function prediction and consequently, drug design (Emanuelsson et al., 2000; Chou and Elrod, 1999; Chou and Shen, 2007a). Among all kind of proteins, bacterial proteins are among the most important proteins to determine their functions due to the wide range of both harmful and useful roles they play in biological interactions (Xiao et al., 2011a). Bacteria are categorized as a kind of prokaryotic micro-organism that can be divided into two groups, namely: Gram-positive and Gram-negative. 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 (Gardy and Brinkman, 2006).

The exponential increase in the number of sequenced proteins and the much slower rate of determining protein structure using experimental approaches highlights the demands for a fast computational approach to address this problem as an alternative to experimental methods (Chou and Shen, 2007a). Among the computational approaches, pattern recognition-based methods have attained the most promising results. Since the introduction of the protein subcellular localization prediction over two decades ago, a wide range of pattern recognition-based approaches have been proposed to solve this problem (Nakai and Kanehisa, 1991, 1992; Nakashima and Nishikawa, 1994). The performance of a pattern recognition technique to address protein subcellular localization prediction problem depends on the classification technique as well as features being used (Mei, 2012; Li et al., 2012; Hu et al., 2012; Sharma et al., 2013a; Tantoso and Li, 2008; Fan and Li, 2012).

To tackle this problem, a wide range of classification techniques have been implemented and used (Chou and Shen, 2007a, 2006a; Chou et al., 2010; Chou and Shen, 2008; Wan et al., 2013; Chou and Cai, 2002; Yu et al., 2013; Nanni et al., 2013a, 2013b; Shen and Chou, 2007; Huang and Yuan, 2013). Among these classifiers, *Support Vector Machine (SVM)* (Wan et al., 2013; Yu et al., 2013; Pierleoni et al., 2011; Du and Yu, 2013; Matsuda et al., 2005) or *K-Nearest Neighbor (KNN)* based classifiers (Chou et al., 2010, 2012; Shen and Chou, 2010a; Chen et al., 2013a, 2012) have attained the most promising results. However, the most significant enhancements in protein subcellular localization prediction have been achieved by improving the feature extraction techniques rather than improving on the classification techniques being used (Tantoso and Li, 2008).

In general, extracted features for this problem can be categorized into sequential, physicochemical, structural, evolutionary, and Gene Ontology-based (GO) features (Ashburner et al., 2000). Early studies have relied on sequential-based features which are extracted from the alphabetic sequence of proteins such as occurrence of amino acids along the protein sequence (occurrence feature group), or percentage of the occurrence of the amino acids along the protein sequence (composition feature group) (Nakai and Kanehisa, 1992; Nakashima and Nishikawa, 1994;

Matsuda et al., 2005; Zhang et al., 2013). To add more information about the physical (e.g. size of the amino acids (Nanni et al., 2013b)), and chemical (e.g. hydrophobicity (Pierleoni et al., 2011)) properties of the amino acids as well as information about the structure of the proteins (predicted secondary structure of the proteins (Hu et al., 2012; Li et al., 2012; Tahir et al., 2014)), physicochemical and structural information has been added to the sequential-based features. Adding these features obtained better results than relying solely on sequential-based features. However, with these features, protein subcellular localization prediction accuracy has remained limited (Nanni et al., 2013b; Li et al., 2012).

Later studies have used evolutionary-based features as more informative features to address protein subcellular localization. This information has been mainly extracted from the substitution probabilities of the amino acids along the protein sequence from *Position Specific Scoring Matrix (PSSM)* (Chou and Shen, 2007b; Nanni et al., 2012; Sun et al., 2010). However, relying on PSSM for feature extraction, protein subcellular localization still has room for improvement. The main reason is that previous studies failed to capture local discriminatory information embedded in PSSM properly. They have mainly tried to extract this local information using the protein sequence as a single block which has failed to achieve this goal (Nanni et al., 2013b, 2013a; Dehzangi et al., 2014a).

The most significant enhancement for protein subcellular localization prediction accuracy has been achieved by using *Gene Ontology (GO)* for feature extraction (Chou et al., 2010; Wan et al., 2013; Pacharawongsakda and Theeramunkong, 2013; Mei, 2012; Chou et al., 2011; Xiao et al., 2011a; Lin et al., 2013; Wu et al., 2011). The GO describes the properties of genes in organisms. The GO database was initially established to represent molecular function, biological process and cellular components of proteins. Despite its importance, GO has three main drawbacks. First, extracting GO for proteins produces a large number of features (over 18,000) which needs further process and filtering to extract discriminatory features (Pacharawongsakda and Theeramunkong, 2013; Mei, 2012). Second, the GO information for new proteins is unavailable and many studies use homology-based approaches to extract GO for these proteins which introduces inaccuracy (Wan et al., 2013; Mei, 2012). Finally, GO provides information regarding the functioning of the proteins which includes prior knowledge and is not considered as an ab-initio feature source (Huang and Yuan, 2013; Tang et al., 2013). Hence, GO needs further investigation and filtering to be used as a reliable source for feature extraction purposes.

As realized by a series recent publications (Chen et al., 2013a, b; Fan et al., 2014; Guo et al., 2014; Liu et al., 2014; Qiu et al., 2014b, 2014a; Ding et al., 2014; Xu et al., 2014) and called for by Chou in Chou (2011), to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps one-by-one."

In this study, we aim to explore potential discriminatory information embedded in PSSM to tackle protein subcellular localization. To capture this information, we propose two segmentation-based feature extraction techniques, namely: segmented distribution method to capture local density and distribution information and segmented auto covariance method to capture local information about the interaction and correlation of neighboring amino acids. These feature groups have attained promising results in similar studies (Dehzangi et al., 2014a, 2013b, 2013c, 2014b). Furthermore, we employ the occurrence feature group to capture global information from evolutionary information embedded in PSSM and to complement local information extracted using the segmented distribution and the segmented auto covariance features (Taguchi and Gromiha, 2007). In this manner, we propose a new reliable method that captures the potential discriminatory information of the evolutionary-based features for protein subcellular localization.

We also employ an SVM classifier which is considered among the best classification techniques used for this problem and has obtained consistent results using segmentation-based features. By applying SVM to our extracted features, we achieve 88.2% and 80.0% prediction accuracies for Gram-positive and Gram-negative protein subcellular localizations, respectively. These results are 3.2% and 6.4% better than previously reported results for Gram-positive and Gram-negative protein subcellular localization which have also used GO for feature extraction (Huang and Yuan, 2013; Pacharawongsakda and Theeramunkong, 2013).

2. Benchmarks

In this study, we use two benchmarks that have been widely used in the literature for Gram-positive and Gram-negative subcellular localizations. For the Gram-positive subcellular localization, we use the benchmark that was proposed in Chou et al. (2010), Chou and Shen (2008), Shen and Chou (2007). This benchmark consists of 519 different proteins belonging to four Gram-positive subcellular locations. Among these 519 proteins, 515 belong to one location while four belong to two locations. Hence, there are 523 ($515 + 4 \times 2$) samples in total. The name of these four locations and the number of proteins in each location are shown at Table 1. This benchmark is available at: <http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi>.

For the Gram-negative we also use the benchmark that was introduced in Chou et al. (2010), Chou and Shen (2008), and Chou and Shen (2006b). This benchmark consists of 1392 different proteins belonging to eight Gram-negative subcellular locations. Among these proteins 1328 belong to one location and 64 to two locations. Therefore, there are 1456 ($1328 + 64 \times 2$) total samples in this benchmark. The name of these eight locations and the number of proteins in each location is provided in Table 2. This benchmark is available at: <http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/>.

To classify the multi-location proteins, we use the same method as used in Huang and Yuan (2013) and Pacharawongsakda and Theeramunkong (2013). We copy the multilabel proteins as several single-label samples based on the number of their labels in our benchmarks. For example a protein with two labels has been used as

Table 1
The name and the number of proteins in each location in Gram-positive bacteria.

No.	Subcellular location	No. of proteins
1	Cell membrane	174
2	Cell wall	18
3	Cytoplasm	208
4	Extracellular	123

Table 2
The name and the number of proteins in each location in Gram-negative bacteria.

No.	Subcellular location	No. of proteins
1	Cell inner membrane	557
2	Cell outer membrane	124
3	Cytoplasm	410
4	Extracellular	133
5	Fimbrium	32
6	Flagellum	12
7	Nucleoid	8
8	Periplasm	180

two single label samples. Then, we conduct the classification task. In this way, by adding extra error, we calculate the worst case as we might not be able to predict the extra label of a single protein that has several labels. Calculating a protein's location in this way, we guarantee that we consider the worst case for our prediction task.

3. Feature extraction method

The main aim of this study is to explore local and global discriminatory information embedded in PSSM for protein subcellular localization. To do this, we will extract a feature group from the transformed protein sequence using evolutionary information in PSSM, namely: consensus sequence-based occurrence (AAO). We will also extract three more feature groups directly from PSSM, namely: semi-occurrence (PSSM-AAO), segmented distribution (PSSM-SD), and segmented auto covariance (PSSM-SAC). Therefore, we first need to produce PSSM.

We produce PSSM for our employed benchmarks as the output (using three iterations to produce PSSM) of the PSI-BLAST algorithm using NCBI's non redundant (NR) database and cut off value (E) of 0.001 (Altschul et al., 1997). PSSM provides the substitution probability of a given amino acid based on its position in a protein sequence with all 20 amino acids. It consists of an $L \times 20$ matrix (where L is the length of protein sequence and 20 columns represent 20 amino acids) which includes the log-odds of the substitution probabilities of the amino acids (Chou and Shen, 2007b; Xu et al., 2013a). In this study, we have used the log-odds values to extract our features. It was shown in the literature that using these numbers produce similar output as using the probability values (Dehzangi et al., 2014a, 2014b). In the following subsections these four feature extraction methods will be explained in detail.

3.1. Consensus sequence-based occurrence (AAO)

This feature group is extracted to provide global information regarding the occurrence of the amino acids along the protein sequence with respect to evolutionary information captured in PSSM. To incorporate evolutionary-based information into the original protein sequence, we transform it using PSSM in the following manner. In the protein consensus sequence, amino acids along the original protein sequence (O_1, O_2, \dots, O_L) are replaced with the corresponding amino acids with the maximum substitution probabilities in PSSM (C_1, C_2, \dots, C_L). We first calculate the index of the amino acid with the highest substitution probability (based on its position in the protein sequence) as follows:

$$I_i = \operatorname{argmax}\{P_{ij} : 1 \leq j \leq 20\}, \quad 1 \leq i \leq L, \quad (1)$$

where P_{ij} is the substitution probability of the amino acid at location i with the j th amino acid in PSSM. Then, we replace the amino acid at i th location of original protein sequence by the I_i th amino acid to form the consensus sequence. After producing the

protein consensus sequence, we produce the occurrence feature group as a 20 dimensional feature vector consisting of the number of occurrence of each amino acid along the protein consensus sequence. Despite its simplicity, the occurrence feature group has shown its effectiveness in maintaining global discriminatory information with respect to the length of a protein sequence (Dehzangi et al., 2013a; Taguchi and Gromiha, 2007; Dehzangi et al., 2014b).

3.2. Semi occurrence (PSSM-AAO)

This feature group is directly extracted from PSSM. It aims at capturing global discriminatory information regarding the substitution probabilities of the amino acids with respect to their positions in the protein sequence. PSSM-AAO is produced by summation of the substitution score of a given amino acid with all the amino acids along the protein sequence which is calculated as follows:

$$PSSM - AAO_j = \sum_{i=1}^l P_{ij}, \quad (j = 1, \dots, 20). \quad (2)$$

The main advantage of using the occurrence feature group over using the composition feature group (in both AAO and PSSM-AAO feature groups) is that it captures information regarding the length of the proteins whereas this information is disregarded in the composition feature group (Taguchi and Gromiha, 2007; Dehzangi et al., 2013a). To extract the occurrence features, we calculate the summation of the total substitution score for a given amino acids while in composition based features, we divide this number with the length of the amino acids. In fact, in composition method, we normalize the occurrence of the amino acids, for a given protein, based on its length. We do not do this normalization which maintain the general total occurrence of the amino acids based on their substitution scores. Since we do not normalize by dividing it by length, it becomes implicitly a part of that feature (Taguchi and Gromiha, 2007; Dehzangi et al., 2013a; Sharma et al., 2013b; Dehzangi et al., 2013d).

3.3. Segmented distribution (PSSM-SD)

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 (Dehzangi and Phon-Amnuaisuk, 2011). We propose this segmentation method in the manner where segments of a protein sequence are of unequal lengths and each segment is represented by a distribution feature which is computed as follows. First, for the j th column in PSSM, we calculate the total substitution probability $T_j = \sum_{i=1}^l P_{ij}$. Then, starting from the first row of PSSM, we calculate the partial sum S_1 of the substitution probabilities of the first i amino acids until reaching $F_p\%$ of the total sum $S_1 = \sum_{i=1}^{I_j^1} P_{ij}$. Using the distribution factor $F_p\%$, we calculate the I_j^1 , where I_j^1 corresponds to the number of the amino acids such that the summation of their substitution probabilities is less than or equal to the $F_p\%$ of T_j . Similarly, we continue to calculate the partial sum of the first i amino acids (starting from the first row of PSSM) until reaching $n \times F_p\% = 50\%$ of the total sum ($S_n = \sum_{i=1}^{I_j^n} P_{ij}$) and calculate the I_j^n corresponding to the number of amino acids such that the summation of their substitution probabilities is less than or equal to 50% of T_j . Therefore, starting from the first row of PSSM, we extract n features ($I_j^1, I_j^2, \dots, I_j^n$) corresponding to the number of segments until reaching 50% of T_j .

We repeat the same process beginning from the last row of PSSM for the j th column. We calculate the partial sum of the

substitution probabilities of the first i amino acids until reaching $n \times F_p\% = 50\%$ of the total sum which is $S_{n+1} = \sum_{i=1}^{I_j^{n+1}} P_{ij}$ until reaching $S_{2n} = \sum_{i=1}^{I_j^{2n}} P_{ij}$, respectively and calculate I_j^{n+1} until reaching I_j^{2n} . I_j^{n+1} and I_j^{2n} correspond to the number of amino acids such that the summation of their substitution probabilities are less than or equal to $F_p\%$ and $n \times F_p\% = 50\%$ of T_j , respectively (starting from the last row of PSSM). Therefore, starting from the last row of the PSSM, we extract n features ($I_j^{n+1}, I_j^{n+2}, \dots, I_j^{2n}$) corresponding to the number of segments until reaching 50% of T_j . In this manner we extract $2n$ segmented distribution features for each column in PSSM. The method used to calculate PSSM-SD feature group from the j th column of PSSM is shown in Fig. 1. We repeat the same process for all 20 columns corresponding to 20 amino acids in PSSM.

In this study, we adopt three values for F_p (5, 10, and 25) to investigate the effectiveness of the number of segments on the achieved results and find the suitable number of segments to explore local discriminatory information embedded in PSSM. We have used other choices for F_p but these three remains the best representatives of all the choices. To maintain the generality and simplicity of the segmentation method, we avoid a very specific segmentation method as it might not be applicable for all cases. For PSSM-SD feature group, using $F_p = 5$, we divide the protein sequence into 20 segments ($n=10$ from each side) and extract 400 features in total in this feature group ($20 \times 20=400$). Similarly, using $F_p=10$ ($n=5$ from each side) we divide the protein sequence in to 10 segments and extract 200 features in total ($10 \times 20=200$) and by using $F_p = 25$ ($n=2$ from each side), we extract 80 features in total ($4 \times 20=80$).

3.4. Segmented auto covariance (PSSM-SAC)

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 (Chou, 2000; Sharma et al., 2013b). To extract this information, the concept of auto covariance has been widely used in the literature in different ways (e.g. bi-gram (Du and Yu, 2013; Sharma et al., 2013b), tri-gram (Sun et al., 2010; Paliwal et al., 2014), auto correlation (Dehzangi and Sattar, 2013a, 2013b)). Among all these methods, pseudo amino acids composition has

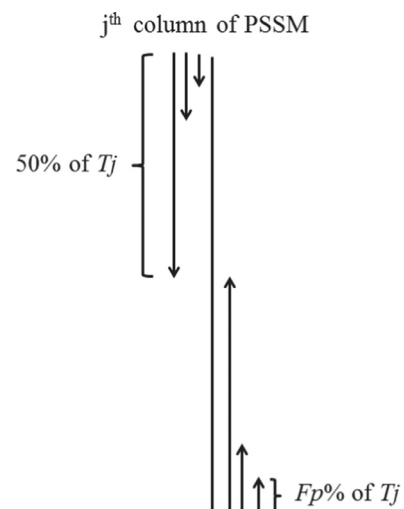


Fig. 1. The segmentation method used to extract PSSM-SD feature group from the j th column of PSSM.

attained the best results to extract local information (Chou et al., 2010; Wan et al., 2013; Esmaeili et al., 2010; Chou, 2001, 2005, 2011). In the present study, we extend the concept of segmented distribution features as described in the previous subsection to compute the auto covariance features from the segmented protein sequence (Dehzangi et al., 2013d). This is done to enforce local discriminatory information extracted from PSSM.

To extract this feature group, we calculate the auto covariance of the substitution probabilities of the amino acids using K_p as the distance factor in the following manner. Starting from the first row of PSSM, for the j th column of PSSM, we calculate K_p auto covariance features for the first I_j^1 . Similarly, we calculate auto covariance until reaching the first I_j^n amino acids. Then starting from the last row of PSSM for the j th column of PSSM, We repeat the same process for I_j^{n+1} and until reaching I_j^{2n} (I_j^1 to I_j^n and I_j^{n+1} until reaching to I_j^{2n} are calculated in the way that is explained in the previous subsection). This process is repeated for all 20 columns of PSSM and corresponding features are calculated as follows:

$$\text{PSSM-seg}_{q,m,j} = \frac{1}{(I_j^q - m)} \sum_{i=1}^{I_j^q - m} (P_{i,j} - P_{ave,j}) \times (P_{(i+m),j} - P_{ave,j}),$$

$$(q = 1, \dots, 2n \ \& \ m = 1, \dots, K_p \ \& \ j = 1, \dots, 20), \tag{3}$$

where $P_{ave,j}$ is the average substitution probability for the j th column in PSSM. Note that $n \times K_p$ auto covariance coefficients are computed in this manner by analyzing PSSM in the downward direction and $n \times K_p$ auto covariance coefficients are computed in this manner by analyzing PSSM in the upward direction ($2n \times K_p$ features in total). We also compute the global auto covariance coefficient (K_p features) of PSSM as follows:

$$\text{PSSM-AC}_{m,j} = \frac{1}{(L - m)} \sum_{i=1}^{L - m} (P_{i,j} - P_{ave,j}) \times (P_{(i+m),j} - P_{ave,j}),$$

$$(m = 1, \dots, K_p \ \& \ j = 1, \dots, 20). \tag{4}$$

Thus, we have extracted a total of $(nK_p + nK_p + K_p) = (2n + 1)K_p$ auto covariance features in this manner (for the j th column of the PSSM). For all 20 columns of the PSSM, segmented auto covariance of the substitution probabilities of the amino acids are extracted and combined to build the corresponding feature group which will be referred to as PSSM-SAC (PSSM-seg + PSSM-AC which consists of $20 \times (2n + 1) \times K_p$ features in total).

We will construct a feature vector consisting of our extracted features and call it PSSM-S (AAO + PSSM-AAO + PSSM-SAC + PSSM-SD = PSSM-S). We can represent our feature vector as

$$F = [f_{1,1}, f_{2,1}, \dots, f_{m,1}, f_{m+1,2}, \dots, f_{p,2}, f_{p+1,3}, \dots, f_{q,3}, f_{q+1,4}, \dots, f_{n,4}]^T, \tag{5}$$

where the superscript T indicates the transpose of the vector and in $f_{n,4}$ the first index (n) indicates the number of features and the second index (4) represents four feature groups used in this study (AAO, PSSM-AAO, PSSM-SAC, PSSM-SD), respectively. These feature vector can also be written in the form of Chou's general PseAAC (pseudo amino acid composition Xu et al., 2013b, 2013a). To do this, let us write this feature vector as

$$F = [\psi_1, \psi_2, \dots, \psi_m, \dots, \psi_\Omega]^T, \tag{6}$$

where Ω is the dimensionality of the feature vector F . The components of feature vector F can be expressed as the pseudo amino acid features (Du et al., 2014, 2012; Cao et al., 2013; Lin and

Lapointe, 2013) as follows:

$$\psi_m = \begin{cases} f_{-,1} & \text{from 1 to } m \in \text{AAO feature group} \\ f_{-,2} & \text{from } m+1 \text{ to } p \in \text{PSSM - AAO feature group} \\ f_{-,3} & \text{from } p+1 \text{ to } q \in \text{PSSM - SAC feature group} \\ f_{-,4} & \text{from } q+1 \text{ to } n \in \text{PSSM - SAC feature group} \end{cases} \tag{7}$$

4. Support Vector Machine

SVM is considered to be one of the best pattern recognition techniques (Vapnik, 1999). It is also widely used in Bioinformatics and has outperformed other classifiers and obtained promising results for protein subcellular localization as well as similar studies (Dehzangi et al., 2014a, 2014b; Dong et al., 2009; Yang and Chen, 2011; Lyon et al., 2014). 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 nonlinear cases). The classification of some known points in input space \mathbf{x}_i is y_i which is defined to be either -1 or $+1$. If \mathbf{x}' is a point in input space with unknown classification then:

$$y' = \text{sign} \left(\sum_{i=1}^n a_i y_i K(\mathbf{x}_i, \mathbf{x}') + b \right), \tag{8}$$

where y' is the predicted class of point \mathbf{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 (Chang and Lin, 2011). 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:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}, \tag{9}$$

where γ is the regularization parameter, \mathbf{x}_i and \mathbf{x}_j are input feature vectors. In this study, the γ in addition to the cost parameter C (also called the soft margin parameter) of the SVM classifier are optimized using a 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.

5. Results and discussion

We investigate the effectiveness of our proposed approaches in the following steps. First we study the effective values for the distribution factor (F_p) in the segmented distribution feature group and the distance factor (K_p) in the segmented auto covariance feature group. In the second step, we investigate the performance of our proposed feature groups by evaluating their effectiveness on the achieved results. Finally, we compare our achieved results with the previously reported results found in the literature.

5.1. Evaluation methods

To evaluate the performance of our proposed method, we use Jackknife cross validation as it has been widely used in the literature for this task and has been shown to be the most consistent and reliable method (Chou and Shen, 2006a; Shen and Chou, 2010a; Mohabatkar et al., 2011, 2013; Lin et al., 2008; Shen and Chou, 2010b; Wu et al., 2012; Xiao et al., 2011b). Jackknife cross validation iteratively uses all but one sample as

the training set and the remaining sample for testing until all the samples are used exactly once for testing purpose. For example, for a benchmark consists of n samples, Jackknife cross validation uses $n-1$ samples as training set and the remaining sample as the testing data in each iteration. It repeats this process n times until all the samples in the benchmark has been used once and only once as the test sample. Hence, it produces exactly one result for each sample and uses the maximum number of available samples ($n-1$) for training in each iteration.

To provide more information about the statistical significance of our achieved results, we will report *Sensitivity*, *Specificity*, and *Matthew's Correlation Coefficient (MCC)* for each subcellular location as well as for the overall benchmark (Hu et al., 2012; Yu et al., 2013; Dehzangi et al., 2014b; Marcin et al., 2012). Sensitivity, which is also referred to as the true positive rate, is a criterion used to evaluate the model as a metric of its ability to identify the correct samples. While specificity, which is also referred to as the true negative rate, is a criterion to specify the ability of the model to identify negative samples. These two parameters 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. In other words, it relates sensitivity and specificity parameters together and measures the correlation of the classification task. MCC varies between -1 and 1 . Having an MCC equal to 1 represents full classification correlation for the model, 0 represents a random classification correlation, and -1 represent the full negative classification correlation. These three parameters are calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100, \quad (10)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100, \quad (11)$$

$$\text{MCC} = \frac{(TN \times TP) - (TN \times FP)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (12)$$

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). Note that the average sensitivity for the whole sample set is also equal to the overall prediction accuracy which is the total number of correctly classified sample (C) over the total number of samples in the data set (N) and is calculated as follows:

$$Q = \frac{C}{N} \quad (13)$$

We will report the overall prediction accuracy as percentage (%).

5.2. Studying the effective values for F_p and K_p

In this step, we construct a feature vector consisting of our extracted features and call it PSSM-S (AAO + PSSM-AAO + PSSM-SAC + PSSM-SD = PSSM-S). The overall architecture of our proposed model is shown in Fig. 2. We investigate the largest effective value for K_p in the range between 1 and 10 with three different values for F_p (25, 10, and 5) as it was mentioned in Section 3. Investigating K_p until 10 is done because it has been shown as the effective value for this parameter for similar studies. It was shown that the effectiveness of this parameter does not vary significantly for values greater than 10, especially for the segmentation based method. Also, studying F_p for three values of 25, 10, and 5 is done because these numbers have been widely used in the literature as the distribution factors to extract distribution-based features. Hence, we produce 30 (10 values for K_p and 3 values for F_p) results

for each benchmark. The results achieved for Gram-positive and Gram-negative bacterial proteins benchmarks are shown in Figs. 3 and 4, respectively.

Note that we optimized the parameters (the regularization parameter (γ) and the cost parameter (C)) for our employed classifier (SVM classifier using RBF as its kernel function) using Gram-positive benchmark and for $F_p = 5$ and $K_p = 10$. It is done to avoid over fitting these two parameters. We have not used Gram-negative benchmark for optimization of these parameters to be able to investigate the generality of our proposed methods. The optimized values for γ and C are 0.005 and 3000 which are used for the rest of our experiments, respectively.

It is shown in Figs. 3 and 4 that by increasing the value of K_p from 1 to 10 (for all three values of F_p), the prediction accuracy declines. Increasing K_p from 1 to 10, increases the number of extracted features from 100 to 1000. Depending on the problem, the extra features can provide more discriminatory information about the correlation of amino acids and consequently increase the prediction accuracy. At the same time, they can also mask the impact of other features and increase the possibility of over fitting depending on the problem. Therefore, the number of feature being used is directly related to the case study. Our results shown that by increasing the distance factor, and consequently increasing the number of features, the prediction accuracy declines. This means increasing the distance factor (K_p from 1 to 10) does not provide additional discriminatory information. Our achieved results show that the prediction accuracy is maximum for $F_p = 25$ occurs when $K_p = 1$ while for $F_p = 10$ or $F_p = 5$ it occurs when $K_p = 2$ or $K_p = 3$, respectively.

As we increase the number of segments by reducing the distribution factor, we also increase the number of features (80 features for $F_p = 25$, 200 features for $F_p = 10$, and 400 features for $F_p = 5$). We can see in Figs. 3 and 4 that increasing the number of features in PSSM-SD (reducing the distribution factor from 25 to 5) requires the number of features in PSSM-SAC to be increased in the beginning to improve the prediction performance (K_p from 1 to 3). However, by increasing K_p to more than 3, the prediction accuracy decreases for all 3 values of F_p . Therefore, to explore potential discriminatory information extracted by PSSM-SD and PSSM-SAC, the number of features in these two feature groups should be consistent. As soon as the number of features in one of these feature groups increases significantly compared to the other one, the prediction accuracy declines.

It is shown in Figs. 3 and 4, for both Gram-positive and Gram-negative subcellular localization benchmarks, the highest results are achieved for $F_p = 25$ and $K_p = 1$. This is even higher than the results achieved by using $F_p = 10$ and $K_p = 2$ or $F_p = 5$ and $K_p = 3$. This shows that using $F_p = 25$ and $K_p = 1$ in PSSM-SD and PSSM-SAC, we are able to explore potential discriminatory information in PSSM for these two feature groups. Therefore, we extract PSSM-SD and PSSM-SAC feature groups for $F_p = 25$ and $K_p = 1$ as the effective values for these parameters for protein subcellular localization and for the rest of our experiments. For PSSM-S using $F_p = 25$ and $K_p = 1$ which produces 220 features in total (20 AAO features + 20 PSSM-AAO features + 80 PSSM-SD features + 100 PSSM-SAC features), we report 88.2% and 80.0% overall prediction accuracies for Gram-positive and Gram-negative bacterial proteins, respectively (Chou et al., 2010; Huang and Yuan, 2013; Pacharawongsakda and Theeramunkong, 2013; Chou and Shen, 2010).

5.3. The impact of our extracted features on the achieved results

To provide more insight to our achieved results and also to investigate the impact of the feature group proposed in this study, we use SVM on each of the extracted feature groups as well as their combinations to build PSSM-S. The sensitivity, specificity, and

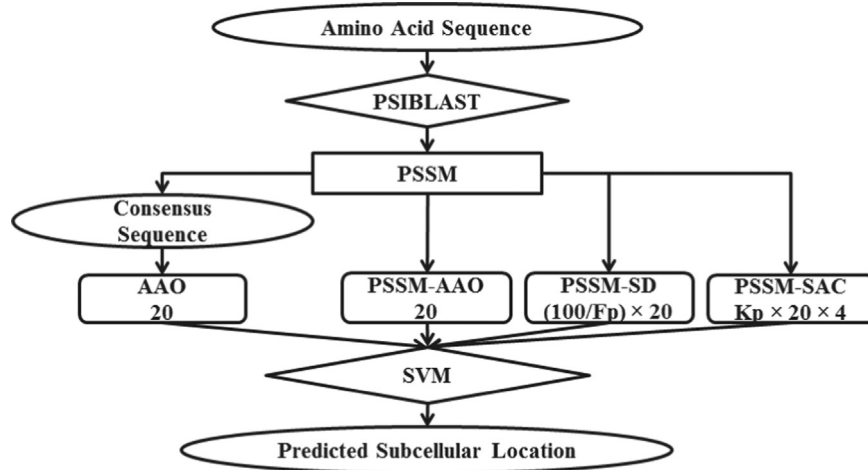


Fig. 2. The overall architecture of our proposed approach. The number of features extracted for each feature group is mentioned under its name.

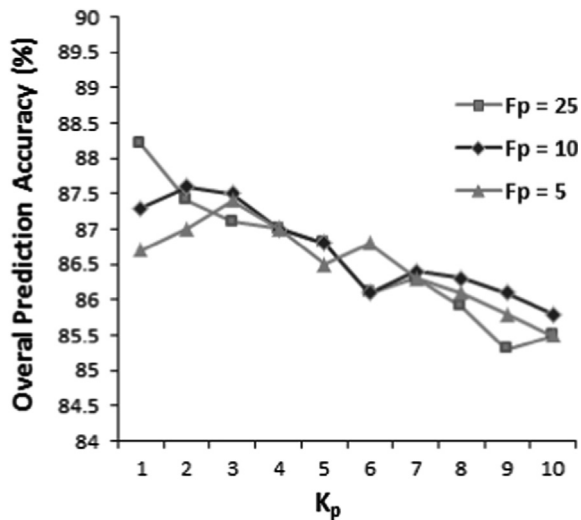


Fig. 3. The overall accuracies achieved for Gram-positive data set for K_p from 1 to 10 and F_p for 5, 10, and 25.

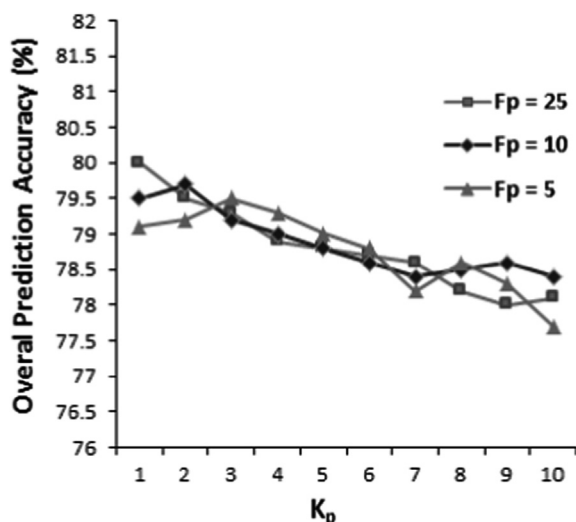


Fig. 4. The overall accuracies achieved for Gram-negative data set for K_p from 1 to 10 and F_p for 5, 10, and 25.

MCC for each subcellular location as well as overall values for Gram-positive and Gram-negative subcellular localization benchmarks are shown in Tables 2 and 3. As it is shown in these tables, the reported results for PSSM-S is better than the results achieved by using each of the feature groups explored in this study independently as well as any other combinations of these feature groups. This highlights the impact of our proposed feature groups.

High values achieved in Table 2 for sensitivity, specificity, and MCC highlights the statistical significance of our achieved results and effectiveness of PSSM-S for the Gram-positive protein subcellular localization (Chou et al., 2010; Shen and Chou, 2007; Huang and Yuan, 2013; Chou and Shen, 2010). For the Gram-negative benchmark, the sensitivity, and MCC are high for the cell inner membrane, cytoplasm, fimbrium, and periplasm while they are lower (over 0.5) for the cell outer membrane, and extracellular, and very low for the Flagellum, and Nucleoid. The poor results for the cell outer membrane, and extracellular locations are due to the difficulty of calculating these locations as it was emphasized in previous studies (Yu et al., 2013; Shen and Chou, 2010a; Marcin et al., 2012; Chou and Shen, 2010). For the flagellum, and nucleoid locations, the main reason for very low sensitivity and MCC is the number of samples belonging to each location. Fewer number of proteins in these two locations (12 samples in the flagellum, and 8 samples in the nucleoid) compared to the number of samples in other locations (557 samples in the cell inner membrane, and 410 samples in the cytoplasm) makes this benchmark, inconsistent (Chou and Shen, 2010). This reduces the prediction accuracy for locations with fewer samples as the SVM classifier weights these classes lower. Dissimilar to the cell inner membrane and cytoplasm, the performance for the fimbrium location is much better, despite relatively smaller number of samples in this location (32 samples) compared to the number of samples in the cell inner membrane location (557 samples). This is because of the simplicity of predicting samples in this location which has also been shown in previous studies (Xiao et al., 2011a; Shen and Chou, 2010a; Pacharawongsakda and Theeramunkong, 2013; Chou and Shen, 2010). As it is shown in Tables 3 and 4, the specificity for all the locations for Gram-positive, and Gram-negative proteins is near one which emphasizes the ability of our method to detect negative samples.

5.4. Comparison with previous studies

To be able to directly compare our results with the best results reported for these two benchmarks, we also produce our results

using 10-fold and 5-fold cross validations (Huang and Yuan, 2013; Pacharawongsakda and Theeramunkong, 2013). In 10-fold cross validation method, the input data set is randomly divided into 10 subsets (5 subsets for 5-fold cross validation). For each iteration, the combination of 9 subsets (4 subsets for 5-fold cross validation) is used as the training set and the remaining subset is used as the testing set. This process is repeated 10 time (5 times for 5-fold cross validation) until all the subsets are used once and only once as the testing set. As it was discusses earlier in subsection 4.1,

Jackknife cross validation is considered as the most effective evaluation method. However, due to its computational expense (performing $n - 1$ times for a benchmark consisting of n samples), some studies have used 10-fold or 5-fold cross validations criteria as an alternative evaluation method (performing 10 times instead of $n - 1$ times) (Huang and Yuan, 2013; Pacharawongsakda and Theeramunkong, 2013; Marcin et al., 2012; Liu et al., 2012). To the best of our knowledge, the best results achieved for Gram-positive and Gram-negative subcellular localizations using similar

Table 3

The sensitivity, specificity and MCC measurements for our proposed feature groups and their combination to build PSSM-S for Gram-positive bacterial proteins benchmark.

Feature vector	Subcellular locations				Overall
	Cell membrane	Cell wall	Cytoplasm	Extracellular	
(Sensitivity)					
AAO	0.609	0.167	0.889	0.756	0.734
PSSM-AAO	0.644	0.222	0.918	0.715	0.755
PSSM-SD	0.793	0.167	0.861	0.878	0.818
PSSM-SAC	0.736	0.222	0.894	0.724	0.778
AAO + PSSM-AAO	0.609	0.222	0.923	0.764	0.757
AAO + PSSM-AAO + PSSM-SD	0.851	0.222	0.913	0.935	0.874
PSSM-S	0.862	0.389	0.923	0.911	0.882
(Specificity)					
AAO	0.917	0.992	0.829	0.878	0.875
PSSM-AAO	0.934	0.990	0.829	0.885	0.883
PSSM-SD	0.940	0.990	0.873	0.928	0.912
PSSM-SAC	0.940	0.986	0.867	0.885	0.899
AAO + PSSM-AAO	0.943	0.992	0.825	0.880	0.883
AAO + PSSM-AAO + PSSM-SD	0.968	0.998	0.911	0.935	0.939
PSSM-S	0.966	0.992	0.914	0.953	0.943
(MCC)					
AAO	0.566	0.252	0.705	0.604	0.620
PSSM-AAO	0.622	0.298	0.733	0.583	0.646
PSSM-SD	0.751	0.233	0.728	0.777	0.730
PSSM-SAC	0.705	0.265	0.751	0.590	0.681
AAO + PSSM-AAO	0.608	0.318	0.734	0.614	0.650
AAO + PSSM-AAO + PSSM-SD	0.839	0.413	0.819	0.831	0.814
PSSM-S	0.843	0.484	0.831	0.845	0.826

Table 4

The sensitivity, specificity and MCC measurements for our proposed feature groups and their combination to build PSSM-S for Gram-negative bacterial proteins benchmark ((1) Cell inner membrane, (2) Cell outer membrane, (3) Cytoplasm, (4) Extracellular, (5) Fimbrium, (6) Flagellum, (7) Nucleoid, (8) Periplasm).

Feature vector	Subcellular locations								Overall
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
(Sensitivity)									
AAO	0.835	0.565	0.834	0.406	0.625	0	0	0.506	0.717
PSSM-AAO	0.840	0.540	0.861	0.481	0.656	0	0	0.550	0.736
PSSM-SD	0.873	0.508	0.885	0.451	0.625	0	0	0.583	0.753
PSSM-SAC	0.858	0.524	0.890	0.459	0.656	0	0	0.622	0.757
AAO + PSSM-AAO	0.846	0.516	0.859	0.496	0.719	0	0	0.600	0.745
AAO + PSSM-AAO + PSSM-SD	0.894	0.540	0.883	0.519	0.656	0.083	0	0.672	0.782
PSSM-S	0.910	0.581	0.890	0.549	0.750	0.167	0	0.678	0.800
(Specificity)									
AAO	0.895	0.979	0.840	0.960	0.995	0.994	0.998	0.958	0.904
PSSM-AAO	0.932	0.971	0.867	0.955	0.989	0.997	0.997	0.951	0.924
PSSM-SD	0.951	0.977	0.877	0.953	0.993	0.992	0.996	0.948	0.9343
PSSM-SAC	0.942	0.982	0.865	0.957	0.995	0.993	0.997	0.954	0.929
AAO + PSSM-AAO	0.924	0.978	0.874	0.953	0.992	0.994	0.997	0.955	0.924
AAO + PSSM-AAO + PSSM-SD	0.958	0.977	0.908	0.957	0.992	0.994	0.997	0.945	0.945
PSSM-S	0.957	0.979	0.911	0.959	0.994	0.994	0.997	0.957	0.947
(MCC)									
AAO	0.723	0.605	0.636	0.404	0.679	–	–	0.512	0.625
PSSM-AAO	0.780	0.552	0.692	0.450	0.610	–	–	0.524	0.660
PSSM-SD	0.832	0.551	0.725	0.420	0.638	–	–	0.544	0.687
PSSM-SAC	0.808	0.590	0.714	0.439	0.695	–	–	0.589	0.688
AAO + PSSM-AAO	0.775	0.564	0.699	0.457	0.690	–	–	0.574	0.669
AAO + PSSM-AAO + PSSM-SD	0.858	0.580	0.766	0.487	0.649	0.084	–	0.602	0.727
PSSM-S	0.870	0.618	0.776	0.519	0.733	0.176	–	0.639	0.748

Table 5

Comparison of our achieved results (%) with the similar studies found in the literature. For Gram-positive and Gram-negative subcellular localizations, the results using 5-fold and 10-fold cross validations are used (in column two and three), respectively.

Study	Gram-positive benchmark	Gram-negative benchmark
Huang and Yuan (2013)	83.7	–
Pacharawongsakda and Theeramunkong (2013)	–	73.2
PSSM-S	86.9	79.6

experiment have been achieved by Huang and Yuan (2013) and Pacharawongsakda and Theeramunkong (2013), respectively. In Huang and Yuan (2013), they used 5-fold cross validation evaluation criterion while in Pacharawongsakda and Theeramunkong (2013), they used 10-fold cross validation criterion. Therefore, in this study, to provide more consistent results, we use Jackknife cross validation to report our results and also use 10-fold and 5-fold cross validations for direct comparison of our results with the previously reported results for these two benchmarks.

Using SVM for PSSM-S with using 10-fold cross validation evaluation method, we achieve 87.7% and 79.6% overall prediction accuracies for Gram-positive and Gram-negative subcellular localization benchmarks, respectively. These are slightly lower than 88.2% and 88.0% overall prediction accuracies using Jackknife cross validation for these two benchmarks. To compare our results directly with Huang and Yuan (2013) for the Gram-positive subcellular localization benchmark, we use 5-fold cross validation evaluation method and report the results. Using 5-fold cross validation, we achieve 86.9% prediction accuracy which is 3.2% better than 83.7% prediction accuracy that was reported in Huang and Yuan (2013). Similarly, to compare our results directly with Pacharawongsakda and Theeramunkong (2013) for the Gram-negative subcellular localization benchmark, we use 10-fold cross validation evaluation method and report the results. Using 10-fold cross validation, we achieve 79.6% prediction accuracy which is 6.4% better than 73.2% prediction accuracy that was reported in Pacharawongsakda and Theeramunkong (2013). These results are achieved by using only 220 features in total. Note that these enhancements achieved by using evolutionary-based features extracted from PSSM compared to the results reported using features extracted from GO (Wan et al., 2013; Yu et al., 2013; Pacharawongsakda and Theeramunkong, 2013; Marcin et al., 2012). It highlights the importance of our method to explore potential discriminatory information embedded in PSSM and introduce reliable features to tackle the protein subcellular localization prediction problem (Table 5).

6. Conclusion and future works

In this study, we have proposed a novel technique to explore potential discriminatory information embedded in PSSM to tackle Gram-positive and Gram-negative subcellular localizations. To do this, we first extracted PSSM matrix for the employed benchmarks and extract two occurrence-based features and two segmentation-based feature groups namely, consensus-based occurrence (AAO), semi-occurrence (PSSM-AAO), segmented distribution (PSSM-SD) and segmented auto covariance (PSSM-SAC). We then tuned the distribution parameters (distribution F_p and distance factors K_p) by applying SVM to the combination of these four feature groups (called PSSM-S) and investigated different values for these two parameters. Finally, by using SVM to our extracted feature groups we showed the importance of all these feature groups to explore potential

discriminatory information embedded in PSSM. We reported 88.2% and 80.0% prediction accuracies for Gram-positive and Gram-negative subcellular localizations using just 220 features in total.

Reported results in this study are 3.2% and 6.4% better than previously reported results for Gram-positive, and Gram-negative bacterial proteins benchmarks which also used GO for feature extraction (Huang and Yuan, 2013; Pacharawongsakda and Theeramunkong, 2013). These enhancements highlight the effectiveness of our method to explore potential discriminatory information embedded in PSSM for protein subcellular localizations and produce effective, and reliable features for this task.

For our future work, we aim at investigating the impact of segmentation-based feature extraction technique to explore other sources for feature extractions such as structural, and physicochemical-based information. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen, 2009; Lin and Lapointe, 2013), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgments

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 available at: <http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/>.

References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* 17, 3389–3402.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al., 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25 (1), 25–29.
- Cao, D.S., Xu, Q.S., Liang, Y.Z., 2013. Propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29 (7), 960–962.
- Chang, Chih-Chung, Lin, Chih-Jen, 2011. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* 2 (3), 27.
- Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C., Chou, K.C., 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One* 7 (10), e47843.
- Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013a. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucl. Acids Res.* 41 (6), e68.
- Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013b. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucl. Acids Res.* 1;41 (6), e68. <http://dx.doi.org/10.1093/nar/gks1450>.
- Chou, K.C., 2000. Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci.* 1, 171–208.
- Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins Struct. Funct. Bioinform.* 43, 246–255.
- Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21 (1), 10–19.
- Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273 (1), 236–247.
- Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* 9 (6), 1092–1100.
- Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* 277 (48), 45765–45769.
- Chou, K.C., Elrod, D.W., 1999. Protein subcellular location prediction. *Protein Eng.* 12 (2), 107–118.
- Chou, K.C., Shen, H.B., 2006a. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* 347 (1), 150–157.
- Chou, K.C., Shen, S.B., 2006b. Large-scale predictions of gram-negative bacterial protein subcellular locations. *J. Proteome Res.* 5 (12), 3420–3428.
- Chou, K.C., Shen, H.B., 2007a. Recent progress in protein subcellular location prediction. *Anal. Biochem.* 370 (1), 1–16.
- Chou, K.C., Shen, H.B., 2007b. MemType-2l: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* 360 (2), 339–345.

- Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3 (2), 153–162.
- Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.* 1 (02), 63.
- Chou, K.C., Shen, H.B., 2010. Plant-mPLOC: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One* 5 (6), e11335.
- Chou, Kuo-Chen, Shen, Hong-Bin, et al., 2010. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci.* 2 (10), 1090.
- Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One* 6 (3), e18258.
- Chou, K.C., Wu, C.Z., Xiao, X., 2012. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* 8 (2), 629–641.
- Dehzangi, A., Phon-Amnuaisuk, S., 2011. Fold prediction problem: the application of new physical and physicochemical-based features. *Protein Pept. Lett.* 18 (2), 174–185.
- Dehzangi, A., Sattar, A., 2013a. Ensemble of diversely trained support vector machines for protein fold recognition. In: *Proceedings of the 5th Asian Conference on Intelligent Information and Database Systems, ACIDS05*. Springer-Verlag, pp. 335–344.
- Dehzangi, A., Sattar, A., 2013b. Protein fold recognition using segmentation-based feature extraction model. In: *Proceedings of the 5th Asian Conference on Intelligent Information and Database Systems, ACIDS05*. Springer-Verlag, pp. 345–354.
- Dehzangi, A., Paliwal, K.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. *IEEE Trans. Comput. Biol. Bioinform.* 10 (3), 564–575.
- Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A., 2013b. Enhancing protein fold prediction accuracy using evolutionary and structural features. In: *Proceedings of the Eighth IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB*, pp. 196–207.
- Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A., 2013c. Exploring potential discriminatory information embedded in PSSM to enhance protein structural class prediction accuracy. In: *Proceedings of the Eighth IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB*, pp. 208–219.
- Dehzangi, A., Paliwal, K.K., Sharma, A., Lyons, J., Sattar, A., 2013d. Protein fold recognition using an overlapping segmentation approach and a mixture of feature extraction models. In: *AI 2013: Advances in Artificial Intelligence*, Springer, pp. 32–43.
- Dehzangi, A., Lyons, J., Sharma, A., Paliwal, K.K., Sattar, A., 2014a. A segmentation-based method to extract structural and evolutionary features for protein fold recognition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (99), 1–11.
- Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A., 2014b. Proposing a highly accurate protein structural class predictor using segmentation-based features. *BMC Genomics* 15 (Suppl 1), S2.
- Dong, Q., Zhou, S., Guan, G., 2009. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* 25 (20), 2655–2662.
- Du, P., Wang, X., Xu, C., Gao, Y., 2012. PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* 425 (2), 117–119.
- Du, P., Gu, S., Jiao, Y., 2014. PseAAC-general: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.* 15 (3), 3495–3506.
- Emanuelsson, O., Nielsen, H., Brunak, S., Heijne, G.V., 2000. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol.* 300 (4), 1005–1016.
- Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* 263 (2), 203–209.
- Fan, G.L., Li, Q.Z., 2012. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* 304, 88–95.
- Fan, Y.N., Xiao, X., Min, J.L., Chou, K.C., 2014. iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking. *Int. J. Mol. Sci.* 15 (3), 4915–4937.
- Gardy, J.L., Brinkman, F.S.L., 2006. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4 (1), 741–751.
- Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W., Chou, K.C., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, btu083.
- Huang, C., Yuan, J., 2013. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *Biosystems* 113 (1), 50–57.
- Hu, Y., Li, T., Sun, J., Tang, S., Xiong, W., Li, D., Chen, G., Cong, P., 2012. Predicting gram-positive bacterial protein subcellular localization based on localization motifs. *J. Theor. Biol.* 308, 135–140.
- Hui, Ding, En-Ze, Deng, Lu-Feng, Yuan, Li, Liu, Hao, Lin, Wei, Chen, Kuo-Chen, Chou, 2014. CTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. vol. 2014. Article ID 286419, pp. 10. (<http://dx.doi.org/10.1155/2014/286419>).
- Li, J., Xiong, L., Schneider, J., Murphy, R.F., 2012. Protein subcellular location pattern classification in cellular images using latent discriminative models. *Bioinformatics* 28 (12), i32–i39.
- Lin, S.X., Lapointe, J., 2013. Theoretical and experimental biology in one—a symposium in honour of professor Kuo-Chen Chou's 50th anniversary and professor Richard Giegé's 40th anniversary of their scientific careers. *J. Biomed. Sci. Eng.* 6, 435.
- Lin, H., Ding, H., Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept. Lett.* 15 (7), 739–744.
- Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2013. iLoc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.* 9 (4), 634–644.
- Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., 2012. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. *Amino Acids* 42, 2243–2249.
- Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., Chou, K.C., 2014. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30 (4), 472–479.
- Lyons, J., Biswas, N., Sharma, A., Dehzangi, A., Paliwal, K.K., 2014. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. *J. Theor. Biol.* 354, 137–145. <http://dx.doi.org/10.1016/j.jtbi.2014.03.033>.
- Marcin, M., Marcin, P., Bujnicki, J.M., 2012. MetaLocGramN: a meta-predictor of protein subcellular localization for gram-negative bacteria. *Biochim. Biophys. (BBA)—Proteins Proteomics* 1824 (12), 1425–1433.
- Matsuda, S., Vert, J.P., Saigo, H., Ueda, N., Toh, H., Akutsu, T., 2005. A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.* 14 (11), 2804–2813.
- Mei, S., 2012. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.* 310, 80–87.
- Mohabatkar, H., Beigi, M.M., Esmaeili, A., 2011. Prediction of GABA_A receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* 281 (1), 18–23.
- Mohabatkar, H., Beigi, M.M., Abdolahi, K., Mohsenzadeh, S., 2013. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.* 9 (1), 133–137.
- Nakai, K., Kanehisa, M., 1991. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Struct. Funct. Bioinform.* 11 (2), 95–110.
- Nakai, K., Kanehisa, M., 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14 (4), 897–911.
- Nakashima, H., Nishikawa, K., 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* 238 (1), 54–61.
- Nanni, L., Lumini, A., Gupta, D., Garg, A., 2012. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9 (2), 467–475.
- Nanni, L., Lumini, A., Brahnam, S., 2013a. An empirical study on the matrix-based protein representations and their combination with sequence-based approaches. *Amino Acid J.* 44 (3), 887–901.
- Nanni, L., Brahnam, S., Ghidoni, S., Menegatti, E., Barrier, T., 2013b. A comparison of methods for extracting information from the co-occurrence matrix for subcellular classification. *Expert Syst. Appl.* 40 (18), 7457–7467.
- Pacharawongsakda, E., Theeramunkong, T., 2013. Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC. *IEEE Trans. Nanobiosci.* 12 (4), 311–320.
- Paliwal, K.K., Sharma, A., Lyons, J., Dehzangi, A., 2014. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. *IEEE Trans. NanoBiosci.* 13 (1), 44–50.
- Pierleoni, A., Martelli, P.L., Casadio, R., 2011. MemLoc: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics* 27 (9), 1224–1230.
- Pufeng, Du, Yuan, Yu, 2013. SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. *BioMed Res. Int.* 2013, 7 (Article ID 263829). (<http://dx.doi.org/10.1155/2013/263829>).
- Qiu, W.R., Xiao, X., Chou, K.C., 2014a. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* 15 (2), 1746–1766.
- Qiu, W.R., Xiao, X., Lin, W.Z., Chou, K.C., 2014b. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *BioMed Res. Int.*
- Sharma, A., Paliwal, K.K., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., 2013a. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. *BMC Bioinform.* 14 (233), 11.
- Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K., 2013b. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *J. Theor. Biol.* 320 (0), 41–46.
- Shen, H.B., Chou, K.C., 2007. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins. *Protein Eng. Des. Sel.* 20 (1), 39–46.

- Shen, H.B., Chou, K.C., 2010a. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of gram-negative bacterial proteins. *J. Theor. Biol.* 264 (2), 326–333.
- Shen, H.B., Chou, K.C., 2010b. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J. Biomol. Struct. Dyn.* 28 (2), 175–186.
- Sun, C., Zhao, X.M., Tang, W., Chen, L., 2010. FGsub: fusarium graminearum protein subcellular localizations predicted from primary structures. *BMC Syst. Biol.* 4 (Suppl 2), S12.
- Taguchi, Y.H., Gromiha, M.M., 2007. Application of amino acid occurrence for discriminating different folding types of globular proteins. *BMC Bioinform.* 8 (1), 404.
- Tahir, M., Khan, A., Kaya, H., 2014. Protein subcellular localization in human and hamster cell lines: employing local ternary patterns of fluorescence microscopy images. *J. Theor. Biol.* 340, 85–95.
- Tang, S., Li, T., Cong, P., Xiong, W., Wang, Z., Sun, J., 2013. PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif. *Nucl. Acids Res.* 41 (W1), W441–W447.
- Tantoso, E., Li, K.B., 2008. AalndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. *Amino Acids* 35 (2), 345–353.
- Vapnik, V.N., 1999. *The Nature of Statistical Learning Theory*. Springer-Verlag.
- Wan, S., Mak, M.-W., Kung, S.-Y., 2013. GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *J. Theor. Biol.* 323, 40–48.
- Wu, Z.C., Xiao, X., Chou, K.C., 2011. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst.* 7 (12), 3287–3297.
- Wu, Z.C., Xiao, X., Chou, K.C., 2012. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein Pept. Lett.* 19 (1), 4–14.
- Xiao, X., Wu, Z.C., Chou, K.C., 2011a. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PLoS One* 6 (6), e20592.
- Xiao, X., Wu, Z.C., Chou, K.C., 2011b. iLoc-virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.* 284 (1), 42–51.
- Xu, Y., Shao, X.J., Wu, L.Y., Deng, N.Y., Chou, K.C., 2013a. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine s-nitrosylation sites in proteins. *PeerJ* 1, e171.
- Xu, Y., Ding, J.W., Ling, Y., Chou, K.C., 2013b. iSNO-PseAAC: predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* 8 (2), e55844.
- Xu, Y., Wen, X., Shao, X.J., Deng, N.Y., Chou, K.C., 2014. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci.* 15 (5), 7594–7610.
- Yang, J.Y., Chen, X., 2011. Improving taxonomy-based protein fold recognition by using global and local features. *Proteins: Struct. Funct. Bioinform.* 79 (7), 2053–2064.
- Yu, L., Luo, J., Guo, Y., Li, Y., Pu, X., Li, M., 2013. In silico identification of gram-negative bacterial secreted proteins from primary sequence. *Comput. Biol. Med.* 43 (9), 1177–1181.
- Zhang, S.W., Liu, Y.F., Yu, Y., Zhang, T.H., Fan, X.N. 2013. MSLoc-DT: a new method for predicting the protein subcellular location of multispecies based on decision templates. *Anal. Biochem.*