IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 23, 2019, accepted March 14, 2019, date of publication March 27, 2019, date of current version April 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907642

An Improved Process for Generating Uniform PSSMs and Its Application in Protein Subcellular Localization via Various Global Dimension Reduction Techniques

SHUNFANG WANG[1], WENJIA LI[1], YU FEI[2], ZICHENG CAO[1], DONGSHU XU[1], AND HUANYU GUO[3]

[1]Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming 650504, China
[2]School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China
[3]Department of Communication Engineering, School of Information Science and Engineering, Yunnan University, Kunming 650504, China

Corresponding authors: Shunfang Wang (sfwang_66@ynu.edu.cn), Yu Fei (feiyukm@aliyun.com), and Zicheng Cao (caozichengtom@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 11661081, Grant 11561071, and Grant 11461079, in part by the Natural Science Foundation of Yunnan Province under Grant 2017FA032, and in part by the Young and Middle-Aged Academic and Technical Leaders Training Plan of Yunnan Province under Grant 2018HB031.

ABSTRACT This paper proposes an improved protein feature expression called segmented amino acid composition in position-specific scoring matrix (PSSM-SAA) in the field of subcellular localization prediction. Since there has been sufficient local information in the PSSM-SAA vector with high dimensionality, four global algorithms of dimensional reduction are suggested, including linear discriminant analysis (LDA), median LDA (MDA), generalized Fisher discriminant analysis (GDA), and median–mean line-based discriminant analysis (MMLDA). PSSM-SAA is also compared with three important expressions: PSSM-S, DipCPSSM, and PsePSSM. Numerical experiments involving the overall success rate (OSR) show that PSSM-SAA is much better than PSSM-S and DipCPSSM and slightly better than or equal in performance to PsePSSM regardless of which dimension reduction algorithm is used. LDA is finally recommended for PSSM-SAA through comparison among four techniques of dimensional reduction. Other popular evaluation indexes also confirm the effectiveness of PSSM-SAA with LDA. Next, the suggested model is compared with the state-of-the-art predictors to further evaluate its validity. Finally, a new user-friendly local software for implementing PSSM-SAA is provided, which can be found at https://www.github.com/caozicheng/PSSM SAA-Builder.

INDEX TERMS Dimensional reduction, feature expression, linear discriminant analysis, protein subcellular localization, segmented amino acid composition in PSSM.

I. INTRODUCTION

Subcellular localization refers to the specific location of a protein or expression product of a functional gene in the cell. For example, for two types of bacterial proteins, Gram-positive and Gram-negative [1], possible locations could be the extracellular matrix, cytoplasm, cell wall or cell membrane of Gram-positive proteins or perhaps the periplasm, nucleoid, flagellum, fimbrium, extracellular matrix, cytoplasm, cell outer membrane or cell inner membrane for Gram-negative proteins. Protein subcellular localization is important to biological research because location errors will greatly affect biological function, which plays a key role in drug design and other applications.

Computational methods for predicting subcellular location have become a hot topic in recent years. Many research results suggest that feature expression methods are key techniques for protein classification prediction [2]–[8]. All possible information that can be extracted by a classification model is determined by a specific feature expression. Thus,

The associate editor coordinating the review of this manuscript and approving it for publication was Ivan Lee.

it is reasonable to believe that a feature expression determines the latent upper limit of prediction accuracy, while efficient classification models can only closely approach this limit [9]–[12]. Based on this consideration, this paper focuses on finding an efficient feature expression for protein subcellular location.

Feature expression is the first step in subcellular localization. In 1986, Nakashima *et al.* expressed protein sequence information with a 20-dimensional amino acid composition (AAC) frequency [2]. In 2000, Chou proposed the concept of pseudo amino acid composition (PseAA) [3], which not only contains amino acid sequence information but also information pertaining to the physical and chemical characteristics of amino acids. Based on the amino acid composition method and the pseudo amino acid composition method, some researchers have accepted the concept of combining the two [4], [5]. In 1999, Jones first proposed a position-specific scoring matrix (PSSM) for protein secondary structure prediction [6], which searched for homologous proteins by PSI-BLAST and was introduced into protein subcellular localization by Bhasin and Raghava in 2004 [7]. A PSSM is a type of matrix whose number of columns is the length of a protein sequence. Thus, proteins with different lengths produce PSSMs of different sizes. Consequently, studies have examined methods for constructing uniform PSSMs. In 2007, Chou and Shen proposed a normalization processing method for PSSMs and called it PsePSSM [8]. Wei *et al.* proposed a model of physicochemical properties based on the PSSM and the k-skip-n-gram [13]. To gain more information, researchers have adopted fusion methods [9]–[12]. Wang and Yang used PseAA-PSSM, which combines pseudo amino acid composition and PSSM methods, in 2009 [9]. Wang and Liu combined dipeptide composition and the PSSM to form DipCPSSM in 2015 [10]. Recently, many PSSM-based feature extraction methods have also been proposed and applied to predict different protein attributes, such as subcellular localization, protein-protein interaction, protein remote homology prediction, and protein structure type. Specifically, Juan *et al.* proposed a method called DP-PSSM to extract features for subcellular localization in Gram-negative proteins in 2009 [14]. In 2013, a new feature extraction method named D-FPSSM was proposed for predicting protein-protein interactions, which is based on PSSM evolutionary information, in [15]. In 2014, Dehzangi *et al.* proposed a feature expression method that extracts discriminative evolutionary features from PSSMs and named it PSSM-S [16]. In the same year, Liu *et al.* proposed an optimal means to incorporate evolutionary information into profiles, which was then applied to protein remote homology prediction [17]. Some researchers selected features by maximizing relevance and minimizing redundancy [18], [19]. To predict the type of protein structure, the RPSSM feature expression method was proposed by Ding *et al.* and Kurgan *et al.*, which can effectively solve problems of inaccuracy that arise in certain methods such as SCPRED [20], [21]. In the aforementioned method, however, AAC, dipeptide composition, PSSM, and other feature expressions are often separately used [22]–[25], or sometimes two of them are conjugated together by simply stitching. All these applications pertaining to PSSMs have shown that the methods based on evolution information in PSSMs have extracted strong features for classification prediction. Therefore, determining how to use uniform PSSMs to form more effective features remains an outstanding issue. In this paper, we implement a different method for fusing PSSM and AAC by means of segment distribution to form a new expression, PSSA-SAA, which features uniform PSSMs and better fuses feature expressions to some extent.

With increasing protein feature information, data dimensions have become much higher than before, for example, the feature expression PSSM-SAA proposed in this paper contains 1600 features. Therefore, it is imperative to eliminate the data redundancy in these feature expressions to form new favorable features [26], which are achieved by mapping PSSM-SAA features onto low-dimensional spaces in this paper. Regarding dimension reduction algorithms, some recent studies have been dedicated to protein subcellular localization to reduce the redundancy in data [9], [10], [27]. In particular, Zou *et al.* proposed a hierarchical feature reduction strategy in 2016 that could further improve the performance of certain predictors of protein attributes [28], [29]. Shan *et al.* combined a feature method with a discriminant analysis method for the prediction of the secondary structure of proteins [30]. However, compared with other pattern recognition research, such as that pertaining to face recognition, studies on protein data dimensionality reduction are relatively scarce. For example, although linear discriminant analysis (LDA), a classic dimension reduction algorithm, has been applied in protein subcellular localization [27], its derivative global algorithms, such as median linear discriminant analysis (MDA) [31], generalized Fisher discriminant analysis (GDA) [32] and median-mean line-based discriminant analysis (MMLDA) [33], must be explored in terms of their performance in protein research. These derivative algorithms are effective supplements to LDA when data contain certain outliers or their within-class covariance matrix is singular. Since the data characteristics of proteins are often unknown, we try to use these derivative algorithms to reduce the dimensionality of our proposed expression PSSM-SAA in this paper for protein subcellular localization.

II. MATERIALS AND METHODS

A. DATA SET

Among all types of proteins, bacterial proteins are essential for basic research and drug design because they have the ability to grow rapidly and have certain special characteristics. We use two datasets in this paper, one pertaining to Gram-positive protein sequences and another pertaining to Gram-negative protein sequences, which can be obtained from http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc/.

The Gram-positive set [34]–[37] contains 519 different bacterial proteins in four subcellular locations, among which

4 proteins belong to two locations and 515 proteins belong to one location. Therefore, there are actually 523 ''locative proteins''. In [38], the concept of ''locative proteins'', together with the difference and relationship between ''protein'' and ''locative protein'', was established. Among the Gram-positive proteins, only 0.77% are located in two or more subcellular location. Compared with more than 15% of the high-level human proteins, which are located in more than one subcellular location [38], [39], proteins of low-level bacteria often belong to one subcellular location, which is a perfect example of how the life-sustaining mechanism of low-level organisms is even simpler than that of high-level organisms [39]. Therefore, for this low-level Gram-negative bacterial organism, we mainly consider the single-label method for subcellular localization. A protein with two labels has been used as two single label samples. This research route is consistent with many recent types of research of this type, such as [16], [40], [41]. That is, in a classification task, a multilabel protein is used as several single-label samples according to the number of their labels in the benchmark. We can guarantee the worst case for predicting the location in this manner [16], [40], [41]. Based on this consideration, for our focus on feature expression and dimension reduction, we simplify this subcellular location problem as a single-label classification problem instead of a multilabel problem.

The Gram-negative set [39], [42], [43] contains 1456 locative proteins. Among them, there are 1392 different bacterial proteins in eight subcellular locations, among which 64 proteins belong to two locations and 1328 proteins belong to one location.

TABLE 1. Detailed information regarding each location in the Gram-positive and Gram-negative sets.

Subcellular location	Number of proteins
Gram-positive	
Cell membrane	174
Cell wall	18
Cytoplasm	208
Extracellular	123
Gram-negative	
Cell inner membrane	557
Cell outer membrane	124
Cytoplasm	410
Extracellular	133
Fimbrium	32
Flagellum	12
Nucleoid	8
Periplasm	180

The details of the Gram-positive and the Gram-negative datasets are provided in Table 1.

B. AN IMPROVED PROCESS FOR OBTAINING UNIFORM PSSMS WITH SEGMENTED AMINO ACID COMPOSITION IN PSSM (PSSM-SAA)

In this section, we further present an improved expression, PSSM-SAA, to provide uniform PSSMs. Each PSSM contains information about the evolution of proteins obtained by the PSI-BLAST algorithm for each protein sequence. For the search, the number of iterations and the E-value are set to 3 and 0.001, respectively. Since the lengths of different protein sequences are different, the final PSSM is distinct, as expressed in Equation (1).

$$P_{PSSM} = \begin{bmatrix} M_{1\to1} & M_{1\to2} & \cdots & M_{1\to20} \\ M_{2\to1} & M_{2\to2} & \cdots & M_{2\to20} \\ \cdots & \cdots & \cdots & \cdots \\ M_{i\to1} & M_{i\to2} & \cdots & M_{i\to20} \\ \cdots & \cdots & \cdots & \cdots \\ M_{L\to1} & M_{L\to2} & \cdots & M_{L\to20} \end{bmatrix} \tag{1}$$

where L is the number of amino acids in the protein sequence and $M_{i\to j}$ is the score describing how the ith amino acid evolves into a j type of amino acid. Thus, we can normalize $M_{i\to j}$ with $P_{ij} = \frac{M_{i\to j} - \min\limits_{1\le i\le L, 1\le j\le 20}(M_{i\to j})}{\max\limits_{1\le i\le L, 1\le j\le 20}(M_{i\to j}) - \min\limits_{1\le i\le L, 1\le j\le 20}(M_{i\to j})}$, where $P_{ij}\epsilon[0, 1]$ is the probability that the ith amino acid is substituted by a j type of amino acid via evolution.

Dehzangi *et al.* [16] proposed a segmented distribution in PSSM (PSSM-SD) approach, but they did not consider the local frequency information of each of the 20 types of amino acids in each segment. This paper proposes an improved feature expression method, PSSM-SAA (segmented amino acid composition in PSSM), which adds more amino acid composition distribution information [4] to the feature expression vector as a necessary local information supplement. Thus, the new method is an innovative approach based on the methods reported in [4] and [16], whose details are as follows.

Similarly to the segmentation method in [16], we divide the protein sequence into several unequal length subsegments. Each subsegment represents a distribution feature. The segmentation process is as follows.

1) Calculate the sum of rows for the jth column of the PSSM and denote it as $T_j \hat{=} \sum_{i=1}^{L} P_{ij}$, $(j = 1, \ldots, 20)$.

2) For the jth column of the PSSM, start from the first row to calculate the row label I_j^1 and the number of amino acids in the first segment. Establish a percentage F_p and n in advance satisfying $0 < F_p < 0.5$ and $n \times F_p = 0.5$. According to $\left(\sum_{i=1}^{I_j^1} P_{ij}\right) \Big/ T_j \le F_p$ and $\left(\sum_{i=1}^{I_j^1+1} P_{ij}\right) \Big/ T_j > F_p$, we can obtain the value of I_j^1.

3) Repeat step 2) to obtain $I_j^2 I_j^3, \ldots, I_j^n$. That is, $\left(\sum_{i=1}^{I_j^2} P_{ij}\right) \Big/ T_j \le 2 \times F_p$ and $\left(\sum_{i=1}^{I_j^2+1} P_{ij}\right) \Big/ T_j > 2 \times F_p, \ldots, \left(\sum_{i=1}^{I_j^n} P_{ij}\right) \Big/ T_j \le n \times F_p$ and $\left(\sum_{i=1}^{I_j^n+1} P_{ij}\right) \Big/ T_j > n \times F_p$. Thus, we obtain an n-dimensional vector $(I_j^1, I_j^2, \ldots, I_j^n)$.

4) For the jth column of the PSSM, start from the last row, and repeat steps 2) and 3) to calculate $(I_j^{n+1}, I_j^{n+2}, \ldots, I_j^{2n})$, which satisfies $\left(\sum_{i=1}^{I_j^{2n}} P_{ij}\right) \Big/ T_j \leq n \times F_p$ and $\left(\sum_{i=1}^{I_j^{2n}+1} P_{ij}\right) \Big/ T_j > n \times F_p$. Thus, there are $2n$ segments in the jth column of the PSSM, which are segmented at the $I_j^1\ th, \cdots, I_j^n\ th$, $I_j^{n+1}\ th$, $\ldots, I_j^{2n}\ th$ row labels. In total, there are $2n \times 20$ segments in the PSSM since there are 20 columns in the PSSM.

Based on the abovementioned segmentation results, the core process of our new proposed PSSM-SAA method, which is also the main difference from that reported in [16], is as follows. For each of the $2n \times 20$ segments in the PSSM, each of the 20 types of amino acids is first counted and then divided by the length of the protein sequence$L$. That is, we obtain the frequency of each of the 20 types of amino acids in each segment with the formula $f_i^{(j,g)} = A_i^{(j,g)} \Big/ L$, where $A_i^{(j,g)} (i = 1, 2, \ldots, 20)$ is the number of the ith amino acid in the gth segment of the jth column ($j = 1, 2, \ldots, 20$; $g = 1, 2, \ldots, 2n$). Therefore, the feature expression vector of PSSM-SAA is expressed as indicated in Equation (2).

$$PSSM-SAA = [PSSM-SAA^{(1,1)}, \cdots, PSSM-SAA^{(j,g)}, \cdots, PSSM-SAA^{(20,2n\times 20)}] \tag{2}$$

where $PSSM - SAA^{(j,g)} = [f_1^{(j,g)}, f_2^{(j,g)}, \ldots, f_{20}^{(j,g)}]^T$, $(j = 1, 2, \ldots, 20; g = 1, 2, \ldots, 2n)$.

The PSSM demonstrates the evolution of the information associated with a protein sequence. PSSM-SAA can extract some information reflecting the distribution of local amino acid composition depending on the knowledge of the PSSM. That is, PSSM-SAA combines both the evolution information and the amino acid composition information of proteins and consequently contains more detailed information. Thus, the performance of PSSM-SAA in protein subcellular localization is examined, as detailed in section 5.

In this paper, we specifically let $F_p = 25$ and $n = 2$. Then, each column is divided into 4 segments. For the total of 20 columns in a PSSM, we can obtain 80 segments corresponding to $(I_j^1, I_j^2, I_j^3, I_j^4)$, $j = 1, 2, \cdots, 20$. According to our proposed PSSM-SAA method, a 1600-dimensional vector, where $1600 = 20 \times 20 \times 4$, can be extracted from a protein sequence.

C. FOUR DIMENSION REDUCTION ALGORITHMS

With sufficient local information extracted, PSSM-SAA greatly increases the number of feature dimensions. Therefore, we recommend that PSSM-SAA be used in conjunction with a dimension reduction algorithm to improve the classification efficiency. Since PSSM-SAA extracts and emphasizes local protein information, this paper uses the global statistical dimension reduction algorithm to achieve an information balancing effect. In this paper, four different types of global dimensionality reduction algorithms—linear discriminant analysis (LDA) [27], median linear discriminant analysis (MDA) [31], generalized Fisher discriminant analysis (GDA) [32], and median mean line-based discriminant analysis (MMLDA) [33]—are combined with PSSM-SAA to comprehensively explore the properties of PSSM-SAA from different perspectives. Few studies have employed pattern recognition to compare these linear dimensional reduction algorithms at the same time; thus, we first provide this comparison in the field of protein subcellular localization.

1. Linear discriminant analysis (LDA) [27] is used to find a set of optimal discriminant vectors such that the projection of the samples from the same class on these vectors is concentrated and samples from different classes are separated far from each other. The definitions of the between-class scatter matrix and within-class scatter matrix are as follows:

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{3}$$

$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \tag{4}$$

where c is the number of sample categories, N_i is the number of samples of the category i, x_{ij} is the jth sample of the category i, $\mu_i = (1/N_i) \sum_{j=1}^{N_i} x_{ij}$ is the mean of samples of the category i, $\mu = (1/N) \sum_{i=1}^{c} \sum_{j=1}^{N_i} x_{ij}$ is the mean of all samples, and N is the total number of samples. The objective function of LDA is defined as follows:

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{5}$$

To satisfy the maximum ratio of the between-class scatter matrix and the within-class scatter matrix, W is required to maximize $J(W)$. This expression can be converted to solve for the generalized eigenvalues in Equation (6) (see [44]).

$$S_B W = \lambda S_W W \tag{6}$$

In Equation (6), $W = [w_1, w_2, \ldots, w_d]$ represents the feature vectors corresponding to the d largest nonzero eigenvalues and the optimal solution for Equation (5), $d < c$.

2. LDA is not sufficient to provide an accurate center for classification when some classes are far away from other classes. Therefore, Yang *et al.* proposed a method involving class median instead of class mean [45]. In this paper, an improved median linear analysis method (MDA) [31] is used for subcellular localization. The steps of MDA are as follows:

1) Calculate the median of the category i. Let $X^i = [x_1, x_2, \ldots, x_{N_i}]^T$ be the samples of the category i, which can be expressed as a matrix $X^i = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{N_i,1} & \cdots & x_{N_i,m} \end{bmatrix}$. Each row represents a sample containing m elements. All elements of the first column are arranged in ascending order. The median $b_{i,1}$is then determined. If m is even, $b_{i,1}$ is the average of the two middle values.

2) According to step 1), determine the median $b_{i,2}$ of the second column until the median ($b_i = [b_{i,1}, b_{i,2}, \ldots, b_{i,m}]^T$) of the category i is obtained after m times.

3) Similarly to the method of calculating b_i, we can obtain the median of the total samples, $B = [B_1, B_2, \ldots, B_m]^T$. In the MDA method, b_i is used to replace μ_i in Equation (3).

4)B is used to replace μ in Equation (4). The classification rule is the same as that in LDA, which can lead to a robust class center in theory.

3. A method based on generalized discriminant analysis (GDA) was proposed by Liu *et al.* [32], which is executed as a small sample size problem. In addition, the formula is as follows: $J_t(W) = \frac{W^T S_B W}{W^T S_t W}$, where $S_t = S_W + S_B$.

4. To overcome the negative effect of outliers on the center of the class, Chou and Shen [34] proposed an adaptive class model, median-mean line-based discriminant analysis (MMLDA), whose steps are as follows:

1) For each sample x_{ij}, calculate the Euclidean distances $d(x_{ij}, x_{sl})$ from x_{ij} to the samples of the category s. Sort these N_s values of $d(x_{ij}, x_{sl})$, $(i = 1, 2, \ldots, c; s = 1, 2, \ldots, c; l = 1, 2, \ldots, N_s)$ in ascending order and determine the median. If N_s is odd, the sample corresponding to the median distance is called the median sample of x_{ij} for category$s$ and denoted by M_{ij}^s. If N_s is even, M_{ij}^s is the average of the two samples corresponding to the two middle distances.

2) Calculate the mean m^s of the category s. m^s and M_{ij}^s can be connected by a straight line, denoted as $\overline{M_{ij}^s m^s}$. The projection of x_{ij} on $\overline{M_{ij}^s m^s}$, which is denoted as $\hat{x}_{ij}^s$, can be calculated by the following formula:

$$\hat{x}_{ij}^s = (1-\alpha) M_{ij}^s + \alpha m^s \quad , \alpha \in [0, 1] \tag{7}$$

Through some algebra, it is easy to obtain $\alpha = (x_{ij} - M_{ij}^s) \cdot (m^s - M_{ij}^s) \Big/ (m^s - M_{ij}^s) \cdot (m^s - M_{ij}^s)$.

3) The goal of MMLDA is to find a projection axis φ by maximizing the following criterion function:

$$J_{MML}(\varphi) = \frac{\varphi^T V_B \varphi}{\varphi^T V_W \varphi} \tag{8}$$

where V_B is the between-class median–mean linear scatter matrix and V_W is the within-class median–mean linear scatter matrix. The terms can be expressed as follows:

$$V_B = \frac{1}{N} \sum_{\substack{ij \\ s \neq i}} (x_{ij} - \hat{x}_{ij}^s)(x_{ij} - \hat{x}_{ij}^s)^T \tag{9}$$

$$V_W = \frac{1}{N} \sum_{\substack{ij \\ s = i}} (x_{ij} - \hat{x}_{ij}^s) / (x_{ij} - \hat{x}_{ij}^s)^T \tag{10}$$

The optimal solution $\varphi = [\varphi_1, \varphi_2, \ldots, \varphi_d]$ of Equation (8) is the feature vector corresponding to the top d eigenvalues of $V_B \varphi = \lambda V_W \varphi$.

D. CUTTING-EDGE FEATURE EXPRESSIONS USED FOR COMPARISON WITH PSSM-SAA

In experimental practice, three currently popular expressions, PSSM-S, DipCPSSM, and PsePSSM, are used to compare and compete with our proposed method PSSM-SAA.

PSSM-S [16] involves the methods of consensus sequence-based occurrence (AAO), semi occurrence (PSSM-AAO), segmented auto covariance (PSSM-SAC) and segmented distribution (PSSM-SD), and contains global evolution information and local discriminative evolutionary information. AAO, PSSM-AAO, PSSM-SAC, and PSSM-SD are extracted from a PSSM. Therefore, PSSM-S is a 220-dimensional feature vector composed of AAO (20 dimensions), PSSM-AAO (20 dimensions), PSSM-SD (80 dimensions) and PSSM-SAC (100 dimensions).

DipCPSSM refers to the fusion of dipeptide composition and PSSM methods [10]. The 400 dipeptide residue pairs from 20 amino acids are counted in the protein sequence [46], together with the 20-dimensional amino acid composition vector, to form a 420-dimensional feature vector [10]. At the same time, a form of normalization processing is performed for the PSSM. That is, the transposed matrix of the PSSM is multiplied by the PSSM to obtain a 20×20 asymmetric matrix (see [47], [48]), which has 210 effective elements located in the lower (or upper) triangular block. Finally, a protein sequence is represented as a 630-dimensional vector with this fusion method DipCPSSM.

PsePSSM [3] is also a type of normalization processing method for PSSMs that is represented as $P_{PsePSSM}^{\xi} = [\overline{M}_1, \overline{M}_2, \ldots, \overline{M}_{20}, G_1^{\xi}, G_2^{\xi}, \ldots, G_{20}^{\xi}]^T \quad (\xi < L)$, where $\overline{M}_j = (1/L) \sum_{i=1}^{L} M_{i \to j} \quad (j = 1, 2, \ldots, 20)$ and $G_j^{\xi} = (1/(L-\xi)) \sum_{i=1}^{L-\xi} [M_{i \to j} - M_{(i+\xi) \to j}]^2$, $(j = 1, 2, \ldots, 20; \xi < L)$. Since the shortest length of the protein sequences in the protein database is 50, the value of ξ is less than 50. Therefore, a protein sequence can be expressed as a 20-dimensional vector ($\xi = 0$) and forty-nine 40-dimensional vectors ($\xi = 1, 2, \ldots, 49$). By removing duplicate elements, $P_{PsePSSM}$ is a 1000-dimensional $(20 + 20 \times 49)$ feature vector.

E. CLASSIFIER, MODEL VALIDATION, AND EVALUATION INDEX

Classifiers have a great influence on the prediction of subcellular location; common classifiers include the k-nearest neighbor algorithm (KNN) [49]–[52] and support vector machine (SVM) [53]–[56]. According to studies investigating similar classification problems in protein subcellular localization, such as [27], the accuracy rate for the Gram-negative dataset with the KNN classifier can reach as high as 93.57%. Another study [57] performed to predict subcellular location with the distance weighted KNN also demonstrated the great advantage of using an improved KNN classifier compared to traditional SVM. In addition, the data treated by the dimensionality reduction algorithm of LDA have the characteristics of maximizing the between-class distance and minimizing the within-classes distance, which is consistent with the theory of KNN, to predict the category according to the type of the nearest samples. Therefore, KNN is chosen instead of SVM in this paper. In the feature space, if most of the K nearest neighbors of a sample belong to a certain category, the sample

also belongs to this category. Euclidean distance is used in this paper.

Regarding model validation, in this paper, we use the Jackknife cross-validation method [58], [59], which is considered to be most reasonable. For a given dataset with N individuals, the basic objective of the Jackknife test is to systematically leave out each individual as the target set and the other $N-1$ individuals as the training set. For the 523 locative proteins addressed in this paper, each protein is taken as the test sequence, and the remaining 522 sequences are used as a training set circularly. A result is obtained after 523 cycles.

The evaluation indexes used in this paper include the overall success rate (OSR)accuracy rate (ACC)Matthews correlation coefficient (MCC)sensitivity (Sen) and specificity (Spe), which are listed as follows (11)–(15), as shown at the bottom of the next page, where $TP(i)$ is the number of samples correctly classified into category i; $FN(i)$ is the number of samples incorrectly classified into category non-i, where non-i indicates all the categories but category i; $TN(i)$ is the number of samples correctly classified into category non-i; and $FP(i)$ is the number of samples incorrectly classified into category i. $Sen(i)$ is the proportion of the samples correctly classified into category i. $Spe(i)$ is the proportion of the samples correctly classified into category non-i. MCC ranges from -1 to 1; the closer the value is to 1, the better the performance of the classifier becomes.

F. EXPERIMENTAL STEPS

The experimental steps are as follows, which are also shown in Figure 1.

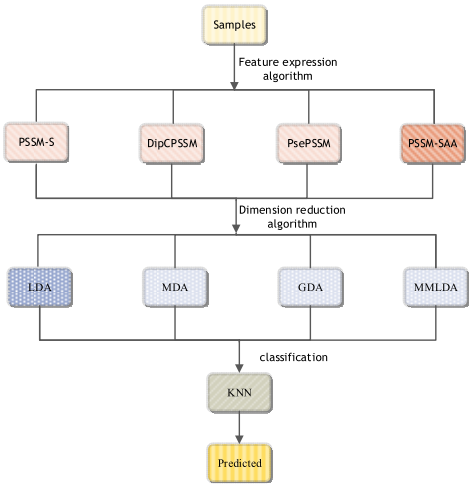

FIGURE 1. Experimental flow chart.

Step 1: Use expression methods PSSM-S, DipCPSSM, PsePSSM, and PSSM-SAA to extract features of the Gram-positive dataset and Gram-negative dataset.

Step 2: Reduce the feature redundancy by using four types of dimension reduction algorithms: LDA, MDA, GDA, and MMLDA.

Step 3: Employ KNN to classify the test samples.

III. RESULTS AND DISCUSSION

A. EXPERIMENTAL COMPARISON

First, we compare PSSM-SAA and three feature expressions with dimension reduction. In this paper, the predicted results are affected by two parameters: the dimension of reduction (d) and the number of nearest neighbors (k). Table 2 lists two OSRs for the Gram-negative set under 16 cases combining four feature expressions and four dimension reduction algorithms. These two types of OSRs include the highest OSR (H-OSR) among different combinations of k and d (indicated in red) and the regular OSR (R-OSR) with certain fixed k and d($k=1$ and $d=7$ here). The contents of Table 3 are similar to those in Table 2 except that the dataset is the Gram-positive set and the R-OSR features the parameters $k=1$ and $d=3$.

Table 2 shows that, overall, PSSM-SAA with dimension reduction performs excellently and reaches OSRs higher than 91% for the Gram-negative dataset. In contrast, the R-OSRs are lower than 90% for PsePSSM and lower than 80% for DipCPSSM and PSSM-S. For an appropriate parameter combination of k and d, PSSM-SAA with dimension reduction can reach H-OSRs higher than 95%, outperforming the other three feature expressions regardless of which dimensionality reduction algorithm is implemented.

Table 3 suggests that PSSM-SAA and PsePSSM yield similar results for the Gram-positive dataset based on the fact that their R-OSRs are both greater than 98% and their H-OSRs are greater than 99%. In Table 3, the experimental results of both PSSM-SAA and PsePSSM appear stable, while the OSRs of DipCPSSM and PSSM-S vary widely from approximately 50% to 90%. PSSM-SAA and PsePSSM show stable OSRs across the different dimension reduction algorithms used. However, for DipCPSSM and PSSM-S, the results for LDA and GDA are much better than those for MDA and MMDA. The results for LDA and GDA agree overall, indicating a small possibility of a singularity of the within-class covariance.

TABLE 2. The OSRs of the Gram-negative set with different feature expressions and dimension reduction algorithms.

Feature expression*	Index	No dimension reduction	Dimension reduction algorithm			
			LDA	MDA	GDA	MMLDA
(1)	R-OSR	0.6690	0.7795	0.7637	0.7816	0.5316
	H-OSR	0.6916	0.8413	0.8290	0.8455	0.6291
(2)	R-OSR	0.6164	0.7328	0.7170	0.7479	0.4341
	H-OSR	0.6806	0.8310	0.8214	0.8420	0.5357
(3)	R-OSR	0.6030	0.8977	0.8942	0.8977	0.7788
	H-OSR	0.6834	0.9499	0.9485	0.9497	0.8668
(4)	R-OSR	0.5639	0.9155	0.9155	0.9155	0.9155
	H-OSR	**0.6882**	**0.9519**	**0.9519**	**0.9554**	**0.9540**

*(1) PSSM-S (2) DipCPSSM, (3) PsePSSM, (4) PSSM-SAA

Overall, as indicated in Tables 2∼3, PSSM-SAA gives the largest H-OSR compared to PSSM-S, DipCPSSM and PsePSSM. The maximum increase in H-OSR with LDA

TABLE 3. The OSRs of the Gram-positive set with different feature expressions and dimension reduction algorithms.

Feature expression*	Index	No dimension reduction	Dimension reduction algorithm			
			LDA	MDA	GDA	MMLDA
(1)	R-OSR	0.5793	0.8528	0.8184	0.8566	0.4857
	H-OSR	0.5870	0.9082	0.8719	0.9063	0.5354
(2)	R-OSR	0.6520	0.8967	0.8528	0.9254	0.6004
	H-OSR	0.7055	0.9235	0.8987	0.9541	0.6291
(3)	R-OSR	0.5985	0.9847	0.9847	0.9847	0.9847
	H-OSR	0.6711	0.9924	0.9924	0.9924	0.9885
(4)	R-OSR	0.5679	0.9847	0.9847	0.9847	0.9847
	H-OSR	**0.6864**	**0.9924**	**0.9924**	**0.9924**	**0.9924**

*(1) PSSM-S (2) DipCPSSM, (3) PsePSSM, (4) PSSM-SAA

is 12.09% for the Gram-negative set, and 8.42% for the Gram-positive. The results of MDA, GDA and MMLDA showed no increase compared to those of LDA. The reason may be that these protein data do not have the small sample size problem and do not contain obvious outliers; thus, there is no need to use the median instead of the mean.

This paper also investigates the results concerning how the OSR varies with the k value for the four feature expressions for the two datasets, as described in the APPENDIX. To summarize, PSSM-SAA performs better than the other three feature expressions. To briefly describe the advantage of PSSM-SAA, Fig. 2 provides the average OSR (abbreviated as A-OSR) across 420 combinations of k and d for the Gram-negative set (k:1-60, d1-7) in subgraph (a) and the A-OSR across 90 combinations for the Gram-positive set (k:1-30, d1-3) in subgraph (b). In Figure 2, the comparison between the four feature expressions and four dimension reduction algorithms suggests that PSSM-SAA is an effective expression method and LDA is a suitable complementary dimensional reduction algorithm. The combination of PSSM-SAA and the dimension reduction algorithm can significantly improve the overall success rate (OSR), and different dimension reduction algorithms have little influence on this result. Although in some cases PSSM-SAA shows performance equal to or slightly lower than that of PsePSSM, most of the time the former exceeds the latter. Table 2 and Figure 2 (a) illustrate how PSSM-SAA outperforms the other three methods for the Gram-negative set. Table 3 and Figure 2 (b) suggest that PSSM-SAA shows the same or slightly better performance than PsePSSM for the Gram-positive set.

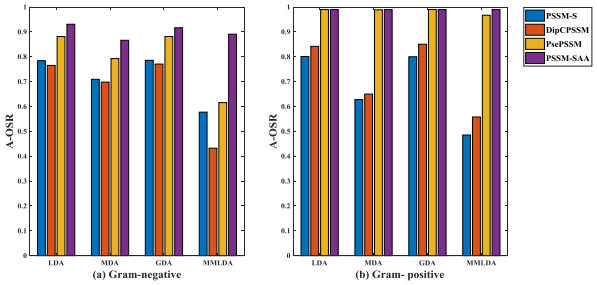

FIGURE 2. The A-OSRs of different feature expressions and dimension reduction algorithms.

Based on the abovementioned numerical results, we aim to gain insight into the four feature expressions with other popular indexes besides the OSR. Since the differences among the dimension reduction algorithms are not obvious, for simplicity and versatility, we only use LDA as the dimensional reduction method for its efficiency in the following experiment to compare with other indexes, which are shown in Table 4 and Table 5. As shown in Table 4 and Table 5, the four indexes of ACC, Sen, Spe, and MCC are used to further evaluate the performance of the feature expressions for the Gram-negative and Gram-positive datasets, respectively, and the number of dimensions is uniformly reduced to 7 (Table 4) or 3 (Table 5) by LDA. In Table 4, the dash means that the value is not in the range $[-1, 1]$. It can be observed that the MCC values of both PsePSSM and PSSM-SAA for the location of the fimbrium both reach up to 1. The situations in which the Sen value is equal to 0 or the MCC value is not in the range $[-1, 1]$ are due to the fact that the sample size associated with the location of the flagellum and nucleoid is smaller than the sample sizes associated with the other six locations. Table 5 shows that many values are the same, which may be caused by the fewer locations in the Gram-positive set. The results of PSSM-SAA, whose multiple values are the same as those of PsePSSM, are still significantly higher than those of

$$OSR = \frac{\sum_{i=1}^{c} TP(i)}{\sum_{i=1}^{c} [TP(i) + FN(i)]} \tag{11}$$

$$ACC(i) = \frac{TP(i) + TN(i)}{TP(i) + TN(i) + FP(i) + FN(i)} \tag{12}$$

$$MCC(i) = \frac{TP(i) \times TN(i) - FP(i) \times FN(i)}{\sqrt{[TP(i) + FN(i)][TP(i) + FP(i)][TN(i) + FP(i)][TN(i) + FN(i)]}} \tag{13}$$

$$Sen(i) = \frac{TP(i)}{TP(i) + FN(i)} \tag{14}$$

$$Spe(i) = \frac{TN(i)}{TN(i) + FP(i)} \tag{15}$$

PSSM-S and DipCPSSM. In Tables 4∼5, the values of PSSM-SAA that are less than those of the other feature expressions (here, only PsePSSM) are marked in red, which represent only a small minority overall, suggesting the superiority of PSSM-SAA.

TABLE 4. The ACC, Sen, Spe, and MCC values of four feature expressions with LDA for the Gram-negative set.

	Protein subcellular location*	PSSM-S	DipCPSSM	PsePSSM	PSSM-SAA
ACC	(1)	0.9045	0.9299	0.9794	0.9835
	(2)	0.9505	0.965	0.9924	0.9952
	(3)	0.9691	0.899	0.9657	0.9718
	(4)	0.9382	0.9437	0.9794	0.9815
	(5)	0.9897	0.9904	1	1
	(6)	0.9918	0.9918	0.9918	0.9918
	(7)	0.9945	0.9945	0.9945	0.9945
	(8)	0.921	0.9341	0.9911	0.9938
Sen	(1)	0.912	0.8761	0.9659	0.9874
	(2)	0.6129	0.6774	0.9355	0.9597
	(3)	0.9634	0.9098	0.9561	0.9634
	(4)	0.5564	0.7669	0.9474	0.9173
	(5)	0.625	0.5938	1	1
	(6)	0	0	0	0
	(7)	0	0	0	0
	(8)	0.75	0.7444	0.9722	0.9667
Spe	(1)	0.8999	0.9633	0.9878	0.9811
	(2)	0.982	0.9917	0.9977	0.9985
	(3)	0.9713	0.8948	0.9694	0.9751
	(4)	0.9766	0.9615	0.9826	0.9879
	(5)	0.9979	0.9993	1	1
	(6)	1	1	1	1
	(7)	1	1	1	1
	(8)	0.9451	0.9068	0.9937	0.9976
MCC	(1)	0.8021	0.8511	0.9563	0.9653
	(2)	0.6565	0.7563	0.9509	0.9689
	(3)	0.9247	0.7686	0.9162	0.9311
	(4)	0.5936	0.6843	0.884	0.8903
	(5)	0.7324	0.747	1	1
	(6)	-	-	-	-
	(7)	-	-	-	-
	(8)	0.6579	0.6986	0.9591	0.9713

Next, we further compare our final model (PSSM-SAA with LDA) with some state-of-the-art predictors. Saini *et al.* [60] discussed subcellular localization for the same Gram-positive and Gram-negative datasets using a linear interpolation smoothing model. Wang and Yang [27] also

TABLE 5. The ACC, Sen, Spe and MCC values of four feature expressions with LDA for the Gram-positive set.

	Protein subcellular location	PSSM-S	DipCPSSM	PsePSSM	PSSM-SAA
ACC	Cell membrane	0.9216	0.9465	0.9962	0.9962
	Cell wall	0.9885	0.9885	0.9981	0.9981
	Cytoplasm	0.9369	0.9541	0.9962	0.9962
	Extracellular	0.935	0.9388	0.9943	0.9943
Sen	Cell membrane	0.8448	0.8736	0.9943	0.9943
	Cell wall	0.7778	0.7222	0.9444	0.9444
	Cytoplasm	0.9663	0.9567	0.9904	0.9952
	Extracellular	0.8455	0.9268	1	0.9919
Spe	Cell membrane	0.9599	0.9828	0.9971	0.9971
	Cell wall	0.996	0.998	1	1
	Cytoplasm	0.9175	0.9524	1	0.9968
	Extracellular	0.9625	0.9425	0.9925	0.995
MCC	Cell membrane	0.8214	0.8788	0.9914	0.9914
	Cell wall	0.8191	0.8135	0.9709	0.9709
	Cytoplasm	0.8727	0.9049	0.992	0.992
	Extracellular	0.8174	0.8385	0.9843	0.9841

TABLE 6. A comparison of the results obtained by the proposed method with recently published results obtained by the jackknife test.

Scheme	Results	
	Gram-Positive	Gram-Negative
Saini et al. [60]	80.1%	83.4%
Wang et al. [27]	-	93.57%
This paper	99.24%	95.19%

predicted the subcellular location of Gram-negative bacteria by the LDA method with a sequence encoding scheme by fusing PSSM and PseAA. In [27], the KNN classifier was employed to identify subcellular location based on reduced low-dimensional feature vectors. Our H-OSR results are compared with the success rates reported in [27], [60] by the Jackknife test in Table 6, which suggests that the results of this paper are better than those in [27], [60].

B. SOFTWARE AND USER GUIDE

The PSSM-SAA method proposed in this paper shows good performance according to the abovementioned experimental results. To verify the effectiveness of PSSM-SAA for readers and users and thus use it for prospective protein feature extraction, the software localization service is provided here to implement PSSM-SAA expression, which is called PSSM-SAA Builder. PSSM-SAA Builder provides users

with usage documents, two data sets as sample examples and the source code, whose homepage is shown in Figure 3.

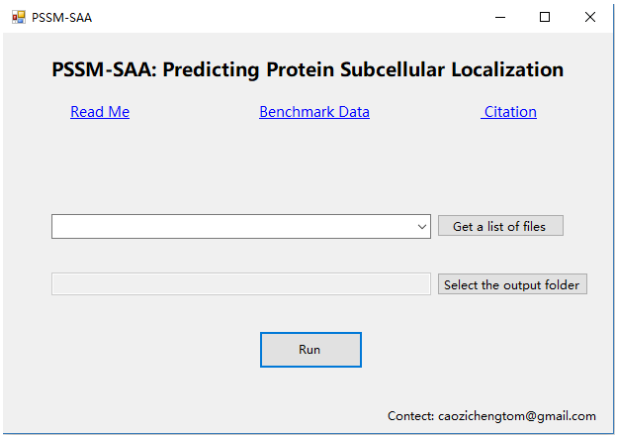

FIGURE 3. A screenshot of software homepage showing PSSM-SAA Builder.

The following steps can help readers understand and use PSSM-SAA Builder.

Step 1: Go to the website https://www.github.com/caozicheng/PSSM_SAA-Builder to find and download the software installation package, sample sets, source code, and user document.

Step 2: Follow the instructions to complete the installation and enter the software homepage as shown in Figure 3, which provides readers with relevant experimental theories and the standard dataset used in this experiment.

Step 3: Click ''Get a list of files'' to select a PSSM file, then click ''Select the output folder'' to customize the file path of PSSM-SAA.

Step 4: Click ''Run''. The PSSM of each protein sequence can generate a 1600-dimensional vector jumping automatically to the relevant path of the generated feature file.

IV. CONCLUSIONS

To summarize, compared to PSSM-S, DipCPSSM and PsePSSM, the proposed PSSM-SAA contains more detailed information as it can extract the distribution information of each amino acid in each segment. This novel approach for extracting local amino acid composition information, depending on the global evolutionary information in a PSSM, is likely the reason for the good performance of PSSM-SAA in protein subcellular localization. Furthermore, besides LDA, other three global dimension reduction algorithms including MDA, GDA and MMLDA are firstly applied for subcellular localization, to provide a contrast with LDA. Totally, with four feature expressions based on four dimension reduction algorithms, the experimental results for the Gram-negative and Gram-positive datasets show that PSSM-SAA with LDA is a promising method in protein classification prediction.

Note that MDA, GDA and MMLDA may be used to treat other specific protein data according to the research needs, which is why we discuss them in this subcellular location. In addition, due to the linearity of PSSM-SAA, four dimension reduction methods considered in this paper are also linear. Generally, there are many nonlinear dimensional reduction algorithms, such as kernel-based methods (kernel LDA, kernel PCA, combinational kernel methods) [61]–[65] and other popular technologies used in bioinformatics [66]–[68], that can be used to study other nonlinear feature expressions if necessary, to form new and interesting directions of research.

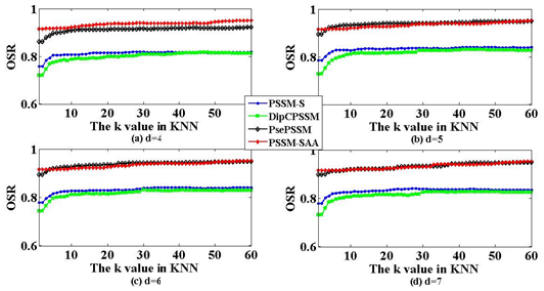

FIGURE 4. The variation of OSR versus k value for LDA and Gram-negative.

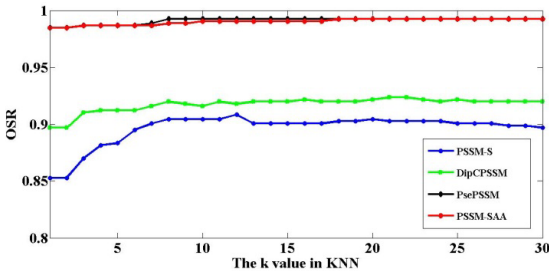

FIGURE 5. The variation of OSR versus k value for LDA with $d = 3$ and Gram-positive.

APPENDIX
FIGURES PERTAINING TO OSR VARY WITH k AND d

In this section, we describe the results indicating how OSR varies with k for the four feature expression methods and two datasets, as shown in Figures 4 $\sim$ 11. Through many numerical experiments, we observe that for the Gram-negative dataset, the H-OSRs of the four feature expressions are mainly distributed in dimensions 4, 5, 6 and 7 for both LDA and GDA and dimensions 5 and 7 for both MDA and MMLDA. For the Gram-negative dataset, the H-OSRs of the four feature expressions are mainly distributed in dimension 3 with LDA, dimensions 2 and 3 with MDA, dimensions 2 and 3 with GDA, and dimensions 2 and 3 with MMLDA. Therefore, we set the fixed dimensions referring to this dimensional distribution to further observe the variation of the OSR.

Specifically, Figure 4 presents the OSRs of the four feature expression methods for the Gram-negative dataset with LDA when the numbers of dimensions are 4, 5, 6 and 7. Figure 5 shows the OSRs of the four feature expressions of

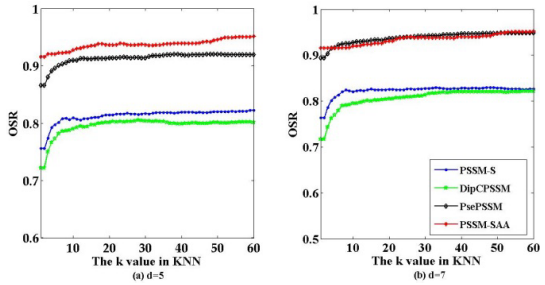

FIGURE 6. The variation of OSR versus k value for MDA and Gram-negative.

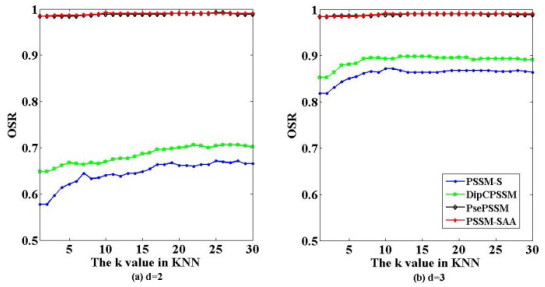

FIGURE 7. The variation of OSR versus k value for MDA and Gram-positive.

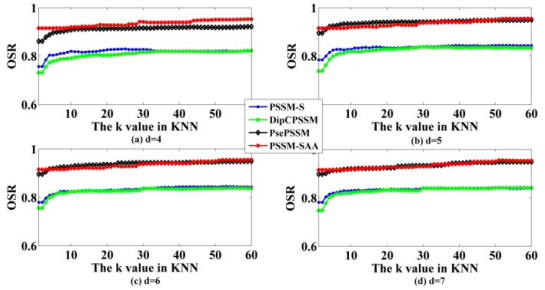

FIGURE 8. The variation of OSR versus k value for GDA and Gram-negative.

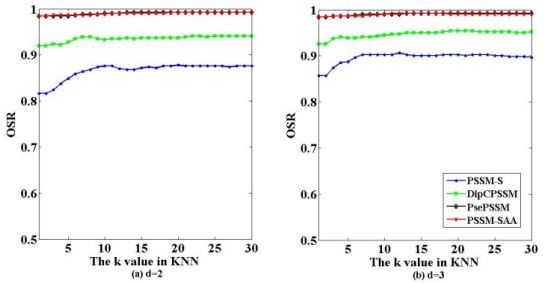

FIGURE 9. The variation of OSR versus k value for GDA and Gram-positive.

the Gram-positive dataset with LDA when the number of dimensions is 3. Figure 6 demonstrates the OSRs of the four feature expressions of the Gram-negative dataset with MDA when the numbers of dimensions are 5 and 7. Figure 7 gives the OSRs of the four feature expressions of the Gram-positive dataset with MDA when the numbers of dimensions are 2 and 3. The results shown in Figures 6 $\sim$ 7 are similar to those in Figures 4 $\sim$ 5, except for the case in which GDA is

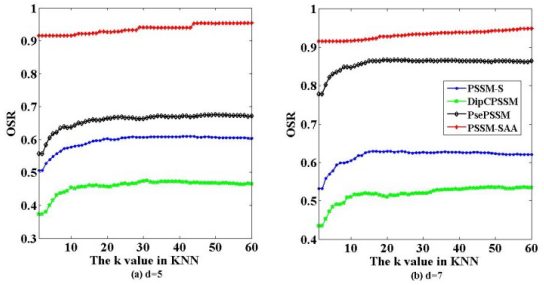

FIGURE 10. The variation of OSR versus k value for MMLDA and Gram-negative.

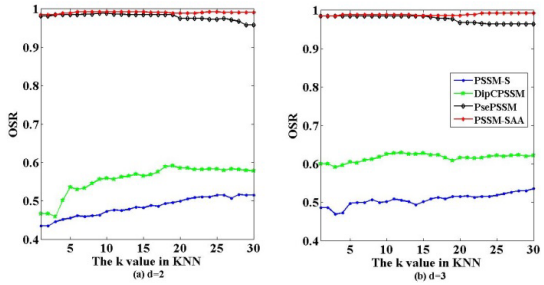

FIGURE 11. The variation of OSR versus k value for MMLDA and Gram-positive.

used, as well as those in Figures 8 $\sim$ 9, which correspond to the use of the MMLDA dimensional reduction algorithm.

For different combinations of d and k, we can draw the similar conclusion that the prediction results of PSSM-SAA are the same or even better than those of the classic PsePSSM with changes in k and d. Although some of the results by PsePSSM are better than those obtained by PSSM-SAA, the largest OSR is obtained using PSSM-SAA. Both PSSM-SAA and PsePSSM perform much better than PSSM-S and DipCPSSM do with changes in k and d.

ACKNOWLEDGMENT

(Shunfang Wang and Wenjia Li contributed equally to this work as co-first authors.)

REFERENCES

[1] X. Xiao, Z.-C. Wu, and 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*, vol. 6, no. 6, Jun. 2011, Art. no. e20592.

[2] H. Nakashima, K. Nishikawa, and T. Ooi, "The folding type of a protein is relevant to the amino acid composition," *J. Biochem.*, vol. 99, no. 1, pp. 153–162, Jan. 1986.

[3] K.-C. Chou, "Prediction of protein subcellular locations by incorporating quasi-sequence-order effect," *Biochem. Biophys. Res. Commun.*, vol. 278, no. 2, pp. 477–483, Nov. 2000.

[4] J. Shi, Q. Pan, S. Zhang, and Y. Cheng, "Classification of protein homo–oligomers using amino acid composition distribution," *Acta Biophysica Sinica*, vol. 22, no. 1, pp. 49–56, Jan. 2006.

[5] H. Yang, Y. M. Cheng, S. W. Zhang, and Q. Pan, "Prediction of protein subcellular localization using a novel feature extraction method: Sequence-segmented pseudo amino acid composition," *Acta Biophysica Sinica*, vol. 24, no. 3, pp. 232–238, Mar. 2008.

[6] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 1999.

[7] M. Bhasin and G. P. S. Raghava, "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST," *Nucleic Acids Res.*, vol. 32, pp. W414–W419, Jul. 2004.
[8] K.-C. Chou and H.-B. Shen, "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochem. Biophys. Res. Commun.*, vol. 360, no. 2, pp. 339–345, Aug. 2007.
[9] T. Wang and J. Yang, "Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins," *Mol. Diversity*, vol. 13, no. 4, pp. 475–481, Nov. 2009.
[10] S. Wang and S. Liu, "Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA," *Int. J. Mol. Sci.*, vol. 16, no. 12, pp. 30343–30361, Dec. 2015.
[11] J. Chen, H. Xu, P.-A. He, Q. Dai, and Y. Yao, "A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously," *Biosystems*, vol. 139, pp. 37–45, Jan. 2016.
[12] Y. Liang and S. Zhang, "Identify Gram-negative bacterial secreted protein types by incorporating different modes of PSSM into Chou's general PseAAC via Kullback–Leibler divergence," *J. Theor. Biol.*, vol. 454, pp. 22–29, Oct. 2018.
[13] L. Wei, M. Liao, X. Gao, J. Wang, and W. Lin, "mGOF-loc: A novel ensemble learning method for human protein subcellular localization prediction," *Neurocomputing*, vol. 217, pp. 73–82, Dec. 2016.
[14] E. Y. T. Juan, W. J. Li, J. H. Jhang, and C. H. Chiu, "Predicting protein subcellular localizations for gram-negative bacteria using DP-PSSM and support vector machines," in *Proc. Int. Conf. Complex, Intell. Softw. Intensive Syst.*, Fukuoka, Japan, Mar. 2009, pp. 836–841.
[15] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, "PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information," *Genomics*, vol. 102, no. 4, pp. 237–242, Oct. 2013.
[16] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou?s general PseAAC," *J. Theor. Biol.*, vol. 364, pp. 284–294, Jan. 2015.
[17] B. Liu *et al.*, "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, Feb. 2014.
[18] I. Naseem, S. Khan, R. Togneri, and M. Bennamoun, "ECMSRC: A sparse learning approach for the prediction of extracellular matrix proteins," *Current Bioinf.*, vol. 12, no. 4, pp. 361–368, 2017.
[19] B.-Q. Li, Y.-H. Zhang, M.-L. Jin, T. Huang, and Y.-D. Cai, "Prediction of protein-peptide interactions with a nearest neighbor algorithm," *Current Bioinf.*, vol. 13, no. 1, pp. 14–24, 2018.
[20] S. Ding, Y. Li, Z. Shi, and S. Yan, "A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile," *Biochimie*, vol. 97, no. 2, pp. 60–65, Feb. 2014.
[21] L. Kurgan, K. Cios, and K. Chen, "SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinformatics*, vol. 9, no. 1, p. 226, May 2008.
[22] B. Liu, J. Chen, and X. Wang, "Application of learning to rank to protein remote homology detection," *Bioinformatics*, vol. 31, no. 21, pp. 3492–3498, Nov. 2015.
[23] B. Liu, S. Wang, Q. Dong, S. Li, and X. Liu, "Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning," *IEEE Trans. Nanobiosci.*, vol. 15, no. 4, pp. 328–334, Jun. 2016.
[24] B. Liu, H. Wu, and 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 Sci.*, vol. 9, no. 4, pp. 67–91, Apr. 2017.
[25] L. Wei, P. Xing, G. Shi, Z.-L. Ji, and Q. Zou, "Fast prediction of protein methylation sites using a sequence-based feature selection technique," *IEEE-ACM Trans. Comput. Biol. Bioinform.*, to be published.
[26] S. Liang, A. J. Ma, S. Yang, Y. Wang, and Q. Ma, "A review of matched-pairs feature selection methods for gene expression data analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 88–97, Feb. 2018.
[27] T. Wang and J. Yang, "Predicting subcellular localization of gram-negative bacterial proteins by linear dimensionality reduction method," *Protein Pept. Lett.*, vol. 17, no. 1, pp. 32–37, Jan. 2010.
[28] Q. Zou, S. Wan, Y. Ju, J. Tang, and X. Zeng, "Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy," *BMC Syst. Biol.*, vol. 10, no. 4, p. 114, Dec. 2016.
[29] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, Jan. 2016.
[30] K. G. Shan, Z. Wei, Y. E. Feng, and L. Z. Yuan, "Using quadratic discriminant analysis to predict protein secondary structure based on chemical shifts," *Current Bioinf.*, vol. 12, no. 1, pp. 52–56, Jan. 2017.
[31] F. Zhang, X. Chen, B. Zhang, and S. Wang, "Improved median linear discriminant analysis for face recognition," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, Hangzhou, China, Dec. 2013, pp. 1051–1055.
[32] K. Liu, Y.-Q. Cheng, J.-Y. Yang, and X. Liu, "An efficient algorithm for Foley-Sammon optimal set of discriminant vectors by algebraic method," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 6, no. 5, pp. 817–829, Dec. 1992.
[33] J. Xu, J. Yang, Z. Gu, and N. Zhang, "Median–mean line based discriminant analysis," *Neurocomputing*, vol. 123, pp. 233–246, Jan. 2014.
[34] K.-C. Chou and H.-B. Shen, "Large-scale predictions of gram-negative bacterial protein subcellular locations," *J. Proteome Res.*, vol. 5, no. 12, pp. 3420–3428, Dec. 2006.
[35] K.-C. Chou and H.-B. Shen, "Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, Jan. 2008.
[36] H.-B. Shen and K.-C. Chou, "Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins," *J. Theor. Biol.*, vol. 264, no. 2, pp. 326–333, May 2010.
[37] H.-B. Shen and K.-C. Chou, "Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites," *J. Biomol. Struct. Dyn.*, vol. 28, no. 2, pp. 175–186, Oct. 2010.
[38] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, May 2008.
[39] Z.-C. Wu, X. Xiao, and K.-C. Chou, "iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of Singleplex and multiplex Gram-positive bacterial proteins," *Protein Peptide Lett.*, vol. 19, no. 1, pp. 4–14, Jan. 2012.
[40] C. Huang and J. Yuan, "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*, vol. 113, no. 1, pp. 50–57, Jul. 2013.
[41] E. Pacharawongsakda and T. Theeramunkong, "Predict subcellular locations of Singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC," *IEEE Trans. Nanobiosci.*, vol. 12, no. 4, pp. 311–320, Dec. 2013.
[42] H.-B. Shen and K.-C. Chou, "Gpos-PLoc: An ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins," *Protein Eng. Des. Sel.*, vol. 20, no. 1, pp. 39–46, Jan. 2007.
[43] H.-B. Shen and K.-C. Chou, "Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins," *Protein Peptide Lett.*, vol. 16, no. 12, pp. 1478–1484, Dec. 2009.
[44] J. Yang and J.-Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognit.*, vol. 36, no. 2, pp. 563–566, 2003.
[45] J. Yang, J. Yang, and D. Zhang, "Median Fisher discriminator: A robust feature extraction method with applications to biometrics," *Front. Comput. Sci.*, vol. 2, no. 3, pp. 295–305, Sep. 2008.
[46] P. Petrilli, "Classification of protein sequences by their dipeptide composition," *Bioinformatics*, vol. 9, no. 2, pp. 205–209, Apr. 1993.
[47] Q.-B. Gao, Z.-Z. Wang, C. Yan, and Y.-H. Du, "Prediction of protein subcellular location using a combined feature of sequence," *FEBS Lett.*, vol. 579, no. 16, pp. 3348–3444, Jun. 2005.
[48] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS One*, vol. 6, no. 3, Mar. 2011, Art. no. e18258.
[49] Q.-B. Gao and Z.-Z. Wang, "Using nearest feature line and tunable nearest neighbor methods for prediction of protein subcellular locations," *Comput. Biol. Chem.*, vol. 29, no. 5, pp. 388–392, Oct. 2005.
[50] L.-Y. Liu, Y.-H. Chen, B.-X. Ma, and Y. Cao, "Prediction of protein subnuclear location using evolutionary fuzzy k-nearest neighbors and its ensemble," *J. Univ. Jinan*, vol. 4, p. 11, Apr. 2010.
[51] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.
[52] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. Syst., Man, Cybern*, vol. 25, no. 5, pp. 804–813, May 1995.

[53] Y.-D. Cai, X.-J. Liu, X.-B. Xu, and G.-P. Zhou, "Support vector machines for predicting protein structural class," *Bmc Bioinformatics*, vol. 2, no. 1, p. 3, Jun. 2001.
[54] Y.-D. Cai, X.-J. Liu, X.-B. Xu, and K.-C. Chou, "Prediction of protein structural classes using support vector machines," *Comput Chem.*, vol. 26, no. 3, pp. 293–296, Feb. 2002.
[55] K.-C. Chou and Y.-D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *J. Biol. Chem.*, vol. 227, no. 48, pp. 45765–45769, Nov. 2002.
[56] Y.-D. Cai, S.-L. Lin, and K.-C. Chou, "Support vector machines for prediction of protein signal sequences and their cleavage sites," *Peptides*, vol. 24, no. 1, pp. 159–161, Jan. 2003.
[57] X. Wang, H. Li, Q. W. Zhang, and R. Wang, "Predicting subcellular localization of apoptosis proteins combining GO features of homologous proteins and distance weighted KNN classifier," *Biomed Res. Int.*, vol. 2016, no. 2, Mar. 2016, Art. no. 1793272.
[58] K.-C. Chou, W.-M. Liu, G. M. Maggiora, and C.-T. Zhang, "Prediction and classification of domain structural classes," *Proteins*, vol. 31, no. 1, pp. 97–103, Apr. 1998.
[59] K.-C. Chou and G. M. Maggiora, "Domain structural class prediction," *Protein Eng.*, vol. 11, no. 7, pp. 523–538, Jul. 1998.
[60] H. Saini, G. Raicar, A. Dehzangi, S. Lal, and A. Sharma, "Subcellular localization for Gram positive and Gram negative bacterial proteins using linear interpolation smoothing model," *J. Theor. Biol.*, vol. 386, pp. 25–33, Dec. 2015.
[61] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. 7th Int. Conf. Artif. Neural Netw. (ICANN)*. Berlin, Germany: Springer-Verlag, 1997, pp. 583–588.
[62] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Müllers, "Fisher discriminant analysis with kernels," in *Proc. Neural Netw. Signal Process. IX, IEEE Signal Process. Soc. Workshop*, Madison, WI, USA, Aug. 1999, pp. 41–48.
[63] J. H. Zhao, S. F. Wang, and F. L. Zhang, "Face recognition study with combination-kernel-based KPCA," *Comput. Eng. Des.*, vol. 35, no. 2, pp. 631–635, Apr. 2014.
[64] S. Wang, B. Nie, K. Yue, Y. Fei, W. J. Li, and D. S. Xu, "Protein subcellular localization with Gaussian kernel discriminant analysis and its kernel parameter selection," *Int. J. Mol. Sci.*, vol. 18, no. 12, p. 2718, Dec. 2017.
[65] S. Wang and Y. Yue, "Protein subnuclear localization based on a new effective representation and intelligent kernel linear discriminant analysis by dichotomous greedy genetic algorithm," *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0195636.
[66] S. Wan, Y. Duan, and Q. Zou, "HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source," *Proteomics*, vol. 17, Sep. 2017, Art. no. 1700262.
[67] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *J. Parallel Distrib. Comput.*, vol. 117, pp. 212–217, Jul. 2018.
[68] Y. You, H. Cai, and J. Chen, "Low rank representation and its application in bioinformatics," *Current Bioinf.*, vol. 13, no. 5, pp. 508–517, 2018.

SHUNFANG WANG received the Ph.D. degree in probability theory and mathematical statistics from Yunnan University, China, in 2005. Since 2016, she has been supervising Ph.D. students. She is currently a Professor with the School of Information Science and Engineering, Yunnan University. Her research interests include machine learning, bioinformatics, and applied statistics.

WENJIA LI received the master's degree in computer application technology from Yunnan University. Her research interests include deep learning and bioinformatics.

YU FEI received the Ph.D. degree in probability theory and mathematical statistics from Yunnan University, in 2003. Since 2015, he has been supervising Ph.D. students. He is currently a Professor with the School of Statistics and Mathematics, Yunnan University of Finance and Economics, China. His research interests include longitudinal data analysis, statistical diagnostics, and applied statistics.

ZICHENG CAO received the master's degree in computer technology from Yunnan University. His research interests include bioinformatics, machine learning, and systems biology.

DONGSHU XU received the master's degree in computer technology from Yunnan University. His research interests include machine learning and computer vision.

HUANYU GUO is currently pursuing the master's degree in communication engineering with Yunnan University, China. His research interests include machine learning and artificial intelligence.