Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 947416, 12 pages
http://dx.doi.org/10.1155/2014/947416

Research Article

iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach

Wang-Ren Qiu,[1] Xuan Xiao,[1,2,3] Wei-Zhong Lin,[1] and Kuo-Chen Chou[3,4]

[1] *Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333046, China*
[2] *Information School, ZheJiang Textile & Fashion College, Ningbo 315211, China*
[3] *Gordon Life Science Institute, Boston, MA 02478, USA*
[4] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence should be addressed to Xuan Xiao; xxiao@gordonlifescience.org

Received 15 February 2014; Revised 26 April 2014; Accepted 29 April 2014; Published 22 May 2014

Academic Editor: Liam McGuffin

Copyright © 2014 Wang-Ren Qiu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Before becoming the native proteins during the biosynthesis, their polypeptide chains created by ribosome's translating mRNA will undergo a series of "product-forming" steps, such as cutting, folding, and posttranslational modification (PTM). Knowledge of PTMs in proteins is crucial for dynamic proteome analysis of various human diseases and epigenetic inheritance. One of the most important PTMs is the Arg- or Lys-methylation that occurs on arginine or lysine, respectively. Given a protein, which site of its Arg (or Lys) can be methylated, and which site cannot? This is the first important problem for understanding the methylation mechanism and drug development in depth. With the avalanche of protein sequences generated in the postgenomic age, its urgency has become self-evident. To address this problem, we proposed a new predictor, called iMethyl-PseAAC. In the prediction system, a peptide sample was formulated by a 346-dimensional vector, formed by incorporating its physicochemical, sequence evolution, biochemical, and structural disorder information into the general form of pseudo amino acid composition. It was observed by the rigorous jackknife test and independent dataset test that iMethyl-PseAAC was superior to any of the existing predictors in this area.

1. Introduction

Posttranslational modifications (PTMs) of proteins are crucial for understanding the dynamic proteome and various signaling pathways or networks in cells. As one of the most important PTMs, protein methylation typically occurs on arginine (Arg) or lysine (Lys) residues in the protein sequence [1]. In fact, there are growing evidences indicating that protein Arg-methylation is capable of providing important regulatory mechanisms for gene expression in a wide variety of biological contexts [2] and that Lys-methylation is correlated with either gene activation or repression depending on the site and degree of methylation [3]. Owing to their important roles in gene regulation (Figure 1), the Arg-methylation and Lys-methylation as well as their regulatory enzymes are implicated in a variety of human disease states, such as cancer [4], coronary heart disease [5], multiple sclerosis [6], rheumatoid arthritis [7], and neurodegenerative disorders [8]. Furthermore, epigenetic inheritance due to methylation can occur through either DNA methylation or protein methylation. Many researches on humans have shown that repeated high-level activation of the body's stress system (particularly in early childhood) could alter methylation processes, leading to changes in the chemistry of the individual's DNA. The chemical changes could disable genes and prevent the brain from properly regulating its response to stress. Researchers and clinicians have drawn a link between this neurochemical dysregulation and the development of chronic health problems such as depression [9], obesity [10], diabetes [11], and hypertension [12]. Therefore, it would certainly provide very useful information or clues for drug discovery to study and analyze the mechanisms that govern these basic epigenetic phenomena.

Although the full extent of regulatory roles of protein methylation is still under elusive investigation, many efforts have been made to determine the methylation sites with experimental approaches, such as mutagenesis of potential

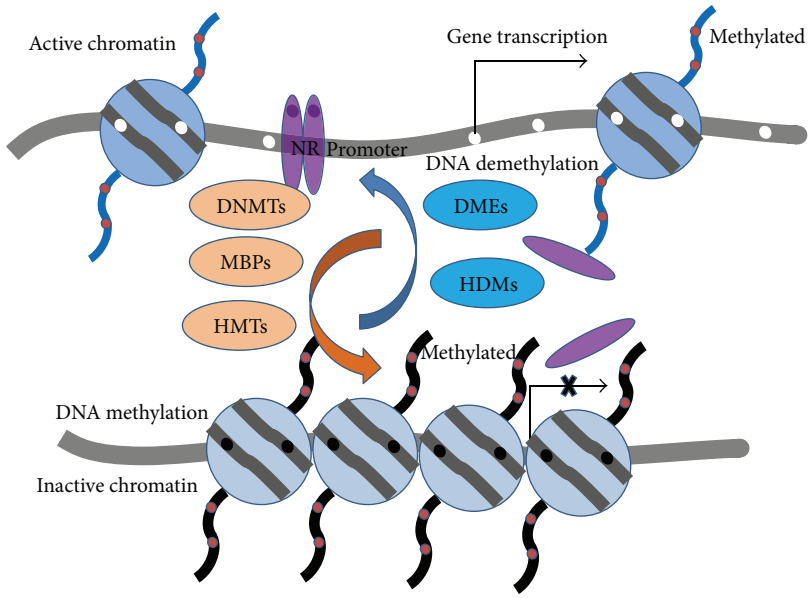

Figure 1: Schematic drawing to show the involvement of the Arg-methylation and Lys-methylation in gene regulation (adapted from [13] with permission).

methylated residues, methylation-specific antibodies [14], and mass spectrometry [15, 16]. The results obtained from these experimental methods have not only provided reliable methylation sites but also indicated that the Arg-methylation and Lys-methylation were closely correlated with the local downstream and upstream residues from the central Arg and Lys, respectively. Unfortunately, even if the number of local residues was limited at $\xi = 5, 6,$ or 7 for both downstream and upstream, it is by no means easy to determine all the methylation sites. This is because the number of possible peptide sequence $\mathbb{N}$ thus formed from 20 amino acids runs into

$$\mathbb{N} = 20^{2\xi} = 10^{2\xi \log(20)} = \begin{cases} 1.0240 \times 10^{13}, & \text{when } \xi = 5 \\ 4.0960 \times 10^{15}, & \text{when } \xi = 6 \\ 1.6384 \times 10^{18}, & \text{when } \xi = 7, \end{cases} \quad (1)$$

which is an astronomical figure for any of the above three cases! It would be exhausting to purely utilize the experimental approaches to determine the large-scale methylation sites. With the avalanche of protein sequences generated in the postgenomic age, it is highly desired to develop automated methods for rapidly and reliably identifying the methylation sites in proteins.

Actually, considerable efforts have been made in this regard. For instance, Daily et al. [17] developed a method for predicting Arg- and Lys-methylation sites using Support Vector Machine (SVM) based on the hypothesis that PTMs preferentially occurred in intrinsically disordered regions [18]. Chen et al. [19] built a web server called MeMo for identifying methylation sites by using the orthogonal binary coding scheme to formulate the protein sequence fragments and SVM to operate the prediction. Using Bi-profile Bayes feature extraction approach, Shao et al. [20] developed a predictor called BPB-PPMS to identify protein methylation sites. Meanwhile, Shien et al. [21] proposed a methylation site prediction method called MASA, in which both sequence information and structural characteristics, such as accessible surface area (ASA) and secondary structure of residues surrounding the methylation sites, were taken into account. Two years later, another method in this area was presented by Hu et al. [22] using the feature selection approach and nearest neighbor algorithm. Recently, Shi et al. [23] developed a method called PMeS to improve the prediction of protein methylation sites based on an enhanced feature encoding scheme and SVM. Although each of the aforementioned methods has its own merit and did play a role in stimulating the development of this area, they all need improvement from one or more of the following aspects: (i) the benchmark dataset used by the previous investigators needs to be updated by incorporating some new and experiment-confirmed data, or improved by removing redundancy and duplicate sequences; (ii) further enhancing the prediction quality by introducing the state-of-the-art machine learning techniques; (iii) making the formulation of all the statistical samples purely based on the sequence information alone because some of the existing methods also needed the structural information that was not always available and hence would unavoidably suffer from some limitation; and (iv) establishing user-friendly and public-accessible web servers because most of the existing methods did not have any web server whatsoever or the web server did not work.

The present study was initiated with an attempt to develop a new predictor for identifying protein methylation sites by focusing on the abovementioned four aspects.

According to a recent comprehensive review [24] and demonstrated by a series of recent publications (see, e.g., [25–28]), to establish a really useful statistical predictor for a protein or peptide 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 protein or peptide 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.

2. Materials and Methods

2.1. Benchmark Dataset. To develop a statistical predictor, it is fundamentally important to establish a reliable and stringent benchmark dataset to train and test the predictor. If the benchmark dataset contains some errors, the predictor trained by it must be unreliable and the accuracy tested by it would be completely meaningless. The benchmark dataset used by Hu et al. [22] contained many duplicate peptide sequences and self-conflicting data. As shown in Part I of the Online Supporting Information S1 available online at http://dx.doi.org/10.1155/2014/947416, of the 180 samples in their positive Arg-methylation learning dataset, 5 were duplicates; of the 2,171 negative learning dataset, 64 were duplicates; of the 10 samples in their positive Arg-methylation testing dataset, 3 were duplicates; of the 206 samples in the negative testing dataset, 46 were duplicates. Similarly, as shown in Part II of the supporting information, of the 262 samples in their positive Lys-methylation learning dataset, 3 were duplicates; of the 2,569 negative learning dataset, 506 were duplicates; of the 48 samples in their positive Lys-methylation testing dataset, 24 were duplicates; of the 243 samples in the negative testing dataset, 111 were duplicates. Also, in their benchmark dataset [22], there were many self-conflicting samples. As shown in Part I of the Online Supporting Information S2, of the 2,351 samples in their learning dataset for Arg-methylation, 8 occur in both positive and negative subsets. Similarly, as shown in Part II of the supporting information, of the 2,831 samples in their learning dataset for Lys-methylation, 60 occur in both positive and negative subsets. Of the 291 samples in their testing dataset for Lys-methylation, 5 occur in both positive and negative subsets. Therefore, the first important thing is to construct a new and reliable benchmark dataset by getting rid of all the duplicates or self-conflicting sequence data. The concrete procedures can be summarized as follows.

In this study the benchmark dataset was derived from the Swiss-Prot database (version 2013_06). Collected were those proteins that had clear experimental annotations about their Arg-methylation and Lys-methylation sites. For facilitating description later, let us adopt the Chou's peptide formulation that was used for studying HIV protease cleavage sites [29, 30], specificity of GalNAc-transferase [31], and signal peptide cleavage sites [32]. According to Chou's scheme, a peptide with Arg (namely R in its single-letter code) or Lys (namely K) located at its center (Figure 2) can be expressed as

$$\begin{aligned}\mathbf{P}(\mathbb{R}) &= \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{R}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi}\\ \mathbf{P}(\mathbb{K}) &= \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{K}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi},\end{aligned} \tag{2}$$

where the subscript ξ is an integer (cf. (1)), $\mathrm{R}_{-\xi}$ represents the ξth downstream amino acid residue from the center, R_{ξ} the ξth upstream amino acid residue, and so forth (Figures 2(a) and 2(b)). Peptides $\mathbf{P}(\mathbb{R})$ and $\mathbf{P}(\mathbb{K})$ with the profile of (2) can be further classified into the following categories:

$$\begin{aligned}&\mathbf{P}(\mathbb{R})\\ &\in \begin{cases}\text{Arg-methylation peptide,} & \text{if its center is a methylation site}\\ \text{non-Arg-methylation peptide,} & \text{otherwise,}\end{cases}\\ &\mathbf{P}(\mathbb{R})\\ &\in \begin{cases}\text{Lys-methylation peptide,} & \text{if its center is a methylation site}\\ \text{non-Lys-methylation peptide,} & \text{otherwise,}\end{cases}\end{aligned} \tag{3}$$

where $\in$ represents "a member of" in the set theory.

As pointed out in a comprehensive review [34], there is no need to separate a benchmark dataset into a training dataset and a testing dataset for validating a prediction method if it is tested by the jackknife or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset for the current study can be formulated as

$$\begin{aligned}\mathbb{S}_{\mathrm{R}} &= \mathbb{S}_{\mathrm{R}}^{+} \cup \mathbb{S}_{\mathrm{R}}^{-}\\ \mathbb{S}_{\mathrm{K}} &= \mathbb{S}_{\mathrm{K}}^{+} \cup \mathbb{S}_{\mathrm{K}}^{-},\end{aligned} \tag{4}$$

where $\mathbb{S}_{\mathrm{R}}$ is the benchmark dataset for Arg-methylation, $\mathbb{S}_{\mathrm{K}}$ is the benchmark dataset for Lys-methylation, $\cup$ is the symbol for "union" in the set theory, $\mathbb{S}_{\mathrm{R}}^{+}$ contains the samples for Arg-methylation peptides only, $\mathbb{S}_{\mathrm{R}}^{-}$ contains the samples for non-Arg-methylation peptides only (cf. (3)), and so forth.

After some preliminary trials and also considering the treatment by the previous investigators [17–20, 22, 23], we chose $\xi = 5$ (cf. (2)) to construct the samples for the benchmark datasets $\mathbb{S}_{\mathrm{R}}$ and $\mathbb{S}_{\mathrm{K}}$, respectively. The detailed procedure was as follows. If the upstream or downstream in a protein was less than 5, the lacking residues were filled with the same residue of its closest neighbor. The peptide samples thus obtained were subject to a screening procedure to winnow those that were identical to any other. Excluded from our benchmark dataset were also those that were self-conflict, namely, occurring in both methylation group and nonmethylation group.

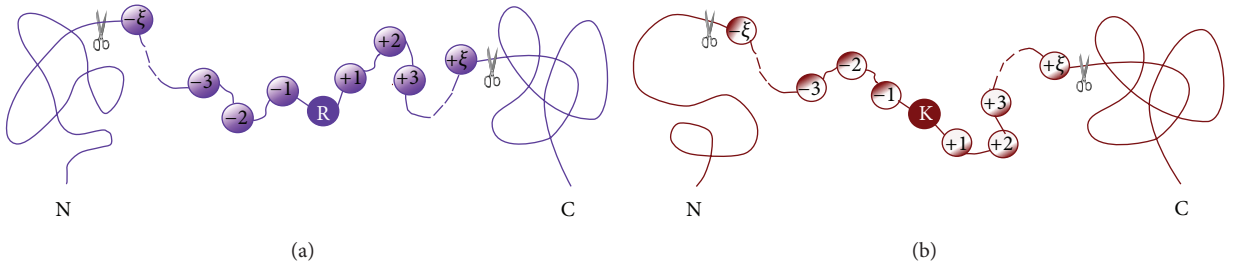

FIGURE 2: Schematic drawing to show the Chou's peptide formulation for studying (a) Arg-methylation and (b) Lys-methylation (adapted from [32, 33] with permission).

Finally, we obtained 1,481 peptide samples for $\mathbb{S}_R$, of which 185 samples were of Arg-methylation belonging to the positive dataset $\mathbb{S}_R^+$, and 1,296 samples of non-Arg-methylation belonging to the negative dataset $\mathbb{S}_R^-$. The Arg-methylation sites and their corresponding $(2\xi+1) = 11$ amino acids along the protein chain are given in the Online Supporting Information S3. Similarly, we also obtained 1,884 peptide samples for $\mathbb{S}_K$, of which 226 samples were of Lys-methylation belonging to the positive dataset $\mathbb{S}_K^+$, and 1,518 samples of non-Lys-methylation belonging to the negative dataset $\mathbb{S}_K^-$. The Lys-methylation sites and their corresponding $(2\xi + 1) = 11$ amino acids along the protein chain are given in the Online Supporting Information S4.

2.2. Sample Formulation. One of the most important but also most difficult problems in computational biology is how to formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. This is because all the existing operation engines, such as "Correlation Angle" method [35–37], "Optimization Approach" [38], "Component Coupled" algorithm [39, 40], "Covariance Discriminant" or CD algorithm [41–44], "Neural Network" algorithm [45, 46], Support Vector Machine or SVM algorithm [27, 47], "Random Forest" algorithm [48], "Conditional Random Field" algorithm [44], "Nearest Neighbor" algorithm [49], "K-Nearest Neighbor" or KNN algorithm [50], "Optimized Evidence-Theoretic K-Nearest Neighbor" or OET-KNN algorithm [51], and "Fuzzy K-Nearest Neighbor" algorithm [26, 52], can only handle vector but not sequence samples. However, a vector defined in a discrete model may completely lose all the sequence-order information [53]. Therefore, in developing a statistical method for predicting the attribute of a peptide in protein, an important task is to formulate the peptide with a vector that can truly reflect its key feature by incorporating some of its sequence information.

To realize this, various feature vectors (see, e.g., [26, 44, 54–64]) were proposed to express proteins or peptides by extracting their different features into the pseudo amino acid composition [53, 65] or Chou's PseAAC [66–68] or general form of PseAAC [24, 69].

According to [24], the general form of PseAAC for a protein or peptide $\mathbf{P}$ can be formulated by

$$\mathbf{P} = [\psi_1 \ \ \psi_2 \ \ \cdots \ \ \psi_u \ \ \cdots \ \ \psi_\Omega]^{\mathbf{T}}, \tag{5}$$

where $\mathbf{T}$ is the transpose operator, while Ω an integer to reflect the vector's dimension. The value of Ω as well as the components ψ_u $(u = 1, 2, \ldots, \Omega)$ in (5) will depend on he protein or peptide sequence. Below, let us describe how to extract the useful information from the benchmark datasets $\mathbb{S}_R$ and $\mathbb{S}_K$ to define the peptide samples via (5). Actually, we are to approach this problem from the following four aspects: (i) position specific scoring matrices (PSSM), (ii) grey-PSSM approach, (iii) amino acid factors (AAF), and (iv) disorder score (DS).

Biology is a natural science with historic dimension. All biological species have developed beginning from a very limited number of ancestral species. It is true for protein sequence as well [70]. Their evolution involves changes of single residues, insertions and deletions of several residues [71], gene doubling, and gene fusion. To incorporate this kind of evolution information into (5), let us consider the following.

According to [72], the sequence evolution information for a peptide with 11 amino acid residues can be expressed by a 11×20 matrix, as given by

$$\mathbf{P}_{\text{PSSM}}^0 = \begin{bmatrix} m_{1,1}^0 & m_{1,2}^0 & \cdots & m_{1,20}^0 \\ m_{2,1}^0 & m_{2,2}^0 & \cdots & m_{2,20}^0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{11,1}^0 & m_{11,2}^0 & \cdots & m_{11,20}^0 \end{bmatrix}, \tag{6}$$

where $m_{i,j}^0$ represents the original score of amino acid residue in the ith $(i = 1, 2, \ldots, 11)$ sequential position of the peptide that is being changed to amino acid type j $(j = 1, 2, \ldots, 20)$ during the evolution process. Here, the numerical codes $1, 2, \ldots, 20$ are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes [73]. The 11×20 scores in (6) were generated by using PSI-BLAST [72] to search the UniProtKB/Swiss-Prot database (Release 2011_05) through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of the peptide $\mathbf{P}$. In order to make every element

in (6) within the range of 0-1, a conversion was performed through the standard sigmoid function to make it become

$$\mathbf{P}_{\text{PSSM}} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,20} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ m_{11,1} & m_{11,2} & \cdots & m_{11,20} \end{bmatrix}, \quad (7)$$

where

$$m_{i,j} = \frac{1}{1 + e^{-m^0_{i,j}}} \quad (1 \le i \le 11, 1 \le j \le 20). \quad (8)$$

Now let us use each of $11 \times 20 = 220$ elements in (7) to represent the 1st 220 components of (5),

$$\psi_u = \begin{cases} m_{1,1} & \text{when } u = 1 \\ \vdots & \vdots \\ m_{11,1} & \text{when } u = 11 \\ \vdots & \vdots \\ m_{1,20} & \text{when } u = 210 \\ \vdots & \vdots \\ m_{11,20} & \text{when } u = 220. \end{cases} \quad (9)$$

Next, let us use the grey model approach to extract more useful information from (7) to define some additional components in (5). According to the grey system theory [74], if the information of a system investigated is fully known, it is called a "white system;" if completely unknown, a "black system;" if partially known, a "grey system". The model developed on the basis of such a theory is called "grey model," which is a kind of nonlinear and dynamic model formulated by a differential equation. The grey model is particularly useful for solving complicated problems that are lack of sufficient information, or need to process uncertain information and to reduce random effects of acquired data. Following the same approach as done by Lin et al. [25], besides the 220 components as defined in the above equation, we can add the following $3 \times 20 = 60$ additional components for (5):

$$\psi_u = \begin{cases} a^1_1 & \text{when } u = 221 \\ a^1_2 & \text{when } u = 222 \\ b^1 & \text{when } u = 223 \\ \vdots & \vdots \\ a^{20}_1 & \text{when } u = 278 \\ a^{20}_2 & \text{when } u = 279 \\ b^{20} & \text{when } u = 280, \end{cases} \quad (10)$$

where

$$\begin{bmatrix} a^j_1 \\ a^j_2 \\ b^j \end{bmatrix} = \left(\mathbf{B}^{\mathbf{T}}_j \mathbf{B}_j\right)^{-1} \mathbf{B}^{\mathbf{T}}_j \mathbf{U}_j \quad (j = 1, 2, \ldots, 20). \quad (11)$$

In the above equation

$$\mathbf{B}_j = \begin{bmatrix} -m_{2,j} & -m_{1,j} - 0.5m_{2,j} & 1 \\ -m_{3,j} & -\sum_{i=1}^{2} m_{i,j} - 0.5m_{3,j} & 1 \\ \vdots & \vdots & \vdots \\ -m_{k,j} & -\sum_{i=1}^{k-1} m_{i,j} - 0.5m_{k,j} & 1 \\ \vdots & \vdots & \vdots \\ -m_{11,j} & -\sum_{i=1}^{11-1} m_{i,j} - 0.5m_{11,j} & 1 \end{bmatrix}, \quad (12)$$

$$\mathbf{U}_j = \begin{bmatrix} m_{2,j} - m_{1,j} \\ m_{3,j} - m_{2,j} \\ \vdots \\ m_{k,j} - m_{k-1,j} \\ \vdots \\ m_{11,j} - m_{10,j} \end{bmatrix}.$$

The structure and function of proteins are largely dependent on the composition of various physicochemical properties of the 20 amino acids. These properties were described with the following five factors by Atchley et al. [75, 76]: (i) polarity (AAF-1), (ii) secondary structure (AAF-2), (iii) molecular volume (AAF-3), (iv) codon diversity (AAF-4), and (v) electrostatic charge (AAF-5). They were used to predict posttranslational modification sites [22, 77, 78]. Thus, using the AAIndex data [79, 80], we can add $5 \times 11 = 55$ components for (5) as formulated below

$$\psi_u = \begin{cases} r^1_1 & \text{when } u = 281 \\ \vdots & \vdots \\ r^1_5 & \text{when } u = 285 \\ \vdots & \vdots \\ r^{11}_1 & \text{when } u = 331 \\ \vdots & \vdots \\ r^{11}_5 & \text{when } u = 335, \end{cases} \quad (13)$$

where r^ℓ_k $(k = 1, 2, \ldots, 5; \ell = 1, 2, \ldots, 11)$ is the kth AAindex for the ℓth amino acid residue of the peptide concerned as given in Table 1 [76].

The functional importance of the disordered regions in proteins has been increasingly recognized [81, 82] and used to predict protein structures and functions [81, 83, 84]. According to Sickmeier et al. [85], they also play various roles in signaling and regulation by multiple binding of proteins and high-specificity low affinity interactions. To incorporate

this kind of information into the PaeAAC of (5), the following 11 components were defined:

$$\psi_u = \begin{cases} @^1 & \text{when } u = 336 \\ @^2 & \text{when } u = 337 \\ \vdots & \vdots \\ @^{11} & \text{when } u = 346, \end{cases} \quad (14)$$

where $@^\ell$ is the disorder score calculated by VSL2 [86] for the ℓth ($\ell = 1, 2, \ldots, 11$) amino residue on the peptide sample.

Finally, we obtained the PseAAC with $\Omega = 346$ components (cf. (5)), of which 220 were defined by (9), 60 by (10), 55 by (13), and 11 by (14). And such 346-D feature vector was used to represent the peptide samples for further study.

2.3. Operation Engine. In this study, we used the SVM (Support Vector Machine) [87, 88] as the operation engine for conducting predictions. SVM is a powerful and popular method for pattern recognition that has been successfully used in the realm of bioinformatics (see, e.g., [64, 89–91]. The basic idea of SVM is to transform the data into a high dimensional feature space and then determine the optimal separating hyperplane using a kernel function. To handle a multiclass problem, "one-versus-one (OVO)" and "one-versus-rest (OVR)" are generally applied to extend the traditional SVM. For a brief formulation of SVM and how it works, see the papers [89, 92]. For more details about SVM, see a monograph [93].

The SVM software used in this paper was downloaded from the LIBSVM package [94], which provided a simple interface. Due to its advantages, the users can easily perform classification prediction by properly selecting the built-in parameters c and γ. In order to maximize the performance of the SVM algorithm, the two parameters in the RBF kernel were preliminarily optimized through a grid search strategy, as briefed as follows. As indicated in (9), (10), (13), and (14), each peptide sample in the current study was a 346-D vector containing $\Omega = 220 + 60 + 55 + 11 = 346$ components. These 346 components were used as the input for each of the peptide samples investigated. The class values were set to 1 for methylation sites and −1 for nonmethylation sites. The threshold used to identify the positive (methylation) or negative (nonmethylation) peptide was set to 0 by default. For this kind of two-group classification, SVM would separate the classes with a surface that maximizes the margin between them. Because the ratio between the numbers of samples in the two groups was about one to seven (the samples in $\mathbb{S}_R^+$ were 185, and the samples in $\mathbb{S}_R^-$ were 1296, while the samples in $\mathbb{S}_K^+$ were 226, and the samples in $\mathbb{S}_K^-$ were 1518), the negative datasets were randomly divided into seven subsets for $\mathbb{S}_R^-$ and $\mathbb{S}_K^-$, respectively. During training process, the jackknife operations were conducted on such 14 datasets to optimize the SVM parameters using the search function SVMcgForClass, which was downloaded from http://www.matlabsky.com/.

The predictor obtained via the aforementioned procedures is called iMethyl-PseAAC.

How to properly and quantitatively measure the quality of a new predictor [95] and how to make it user-friendly for the public are the two key issues that have important impacts on its application value [96]. Below, let us address these two problems.

2.4. A Set of Metrics for Examining Prediction Quality. In literature the following four metrics are often used for examining the performance quality of a predictor

$$\begin{aligned} \text{Sn} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Sp} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{MCC} &= \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}}, \end{aligned} \quad (15)$$

where TP represents the number of the true positive; TN, the number of the true negative; FP, the number of the false positive; FN, the number of the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; and MCC, the Mathew's correlation coefficient. To most biologists, however, the four metrics as formulated in (15) are not quite intuitive and easy to understand, particularly for the Mathew's correlation coefficient. Here let us adopt the formulation proposed recently in [27, 44] based on the symbols introduced by Chou [33] in predicting signal peptides. According to the formulation, the same four metrics can be written as

$$\begin{aligned} \text{Sn} &= 1 - \frac{N_-^+}{N^+} \\ \text{Sp} &= 1 - \frac{N_+^-}{N^-} \\ \text{Acc} &= 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\ \text{MCC} &= \frac{1 - \left(N_-^+/N^+ + N_+^-/N^-\right)}{\sqrt{\left(1 + \left(N_+^- - N_-^+\right)/N^+\right)\left(1 + \left(N_-^+ - N_+^-\right)/N^-\right)}}, \end{aligned} \quad (16)$$

where N^+ is the total number of the Arg-methylation (or Lys-methylation) peptides investigated, while N_-^+ is the number of the peptides incorrectly predicted as the non-Arg-methylation peptides, and N^- is the total number of the non-Arg-methylation investigated, while N_+^- is the number of the non-Arg-methylation incorrectly predicted as the Arg-methylation peptides [97].

Now, it is crystal clear from (16) that when $N_-^+ = 0$ meaning none of the Arg-methylation peptides was incorrectly predicted to be a non-Arg-methylation peptide, we have the sensitivity Sn = 1. When $N_-^+ = N^+$ meaning that all the Arg-methylation peptides were incorrectly predicted to

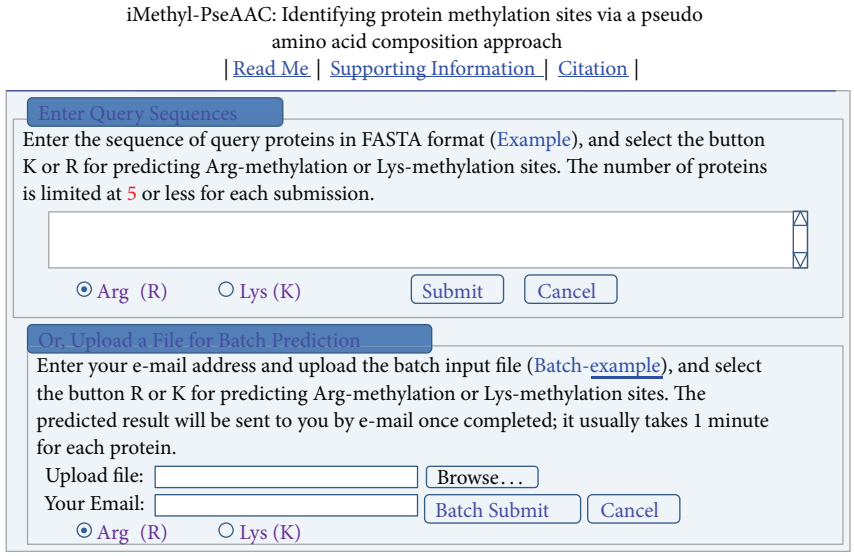

Figure 3: A semiscreenshot to show the top page of iMethyl-PseAAC. Its web-site address is at http://www.jci-bioinfo.cn/iMethyl-PseAAC.

be the non-Arg-methylation peptides, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning none of the non-Arg-methylation peptides was incorrectly predicted to be the Arg-methylation peptide, we have the specificity Sp = 1, whereas $N_+^- = N^-$ meaning all the non-Arg-methylation peptides were incorrectly predicted as the Arg-methylation peptides, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of Arg-methylation peptides in the positive dataset and none of the non-Arg-methylation peptides in the negative dataset was incorrectly predicted, we have the overall accuracy Acc = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the Arg-methylation peptides in the positive dataset and all the non-Arg-methylation peptides in the negative dataset were incorrectly predicted, we have the overall accuracy Acc = 0 and MCC = −1, whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have Acc = 0.5 and MCC = 0 meaning no better than random prediction. As we can see from the above discussion based on (16), the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier-to-understand.

2.5. Web Server and User Guide. For the convenience of the vast majority of biological scientists, a web server for iMethyl-PseAAC was established. Here, let us provide a step-by-step guide on how to use the web server to get the desired results without the need to follow the mathematic equations that were presented just for the integrity in developing the predictor.

Step 1. Open the web server at http://www.jci-bioinfo.cn/iMethyl-PseAAC and you will see the top page of the predictor on your computer screen, as shown in Figure 3. Click on the *Read Me* button to see a brief introduction about iMethyl-PseAAC predictor and the caveat when using it.

Step 2. Either type or copy/paste the sequences of query proteins into the input box located at the center of Figure 3. The input should be in the FASTA format; only the 20 native amino acid codes are allowed in the protein sequences. Click the *Example* button to see the input format.

Step 3. Check on the "Arg" button for predicting the Arg-methylation sites, or "Lys" button for the Lys-methylation sites.

Step 4. Click the *Submit* button to see the predicted result. For example, if you use the sequences of the two query proteins in the *Example* window as the input and check the Arg button on, after clicking the *Submit* button, you will see the following predicted results. The total number of Arg (R) in the 1st protein (P62805) is 14, and the Arg at the sequence positions 4 and 41 (highlighted in red) is the methylation site, but the Arg at all the other 12 sites is not. The total number of Arg (R) in the 2nd protein (P68431) is 18, and the Arg at the sequence positions 3, 9, and 18 (highlighted in red) is the methylation site, but the Arg at all the other 15 sites is not. However, if you check the Lys button for the two query proteins, after clicking the *Submit* button, you will see that the total number of Lys (K) in the 1st protein (P62805) is 11, and the Lys at the sequence positions 13, 17, and 21 (highlighted in red) is the methylation site, but the Lys at all the other 8 sites is not, and that the total number of Lys (K) in the 2nd protein (P68431) is 13, of which, except the sequence position 116, the Lys at all the other 12 positions is the methylation site. A comparison of these predicted results with the experimental observations will be given in the Results and Discussion section. It takes about 30 seconds for the above computation before the predicted results appear on the computer screen; the more number of query proteins and longer of each sequence, the more time it is usually needed.

The number of proteins is limited at 5 or less for each such direct submission.

Step 5. As shown on the lower panel of Figure 3, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the "Browse" button. To see the sample of batch input file, click on the button *Batch-example*. After clicking the button *Batch-submit*, you will see "Your batch job is under computation; once the results are available, you will be notified by e-mail."

Step 6. Click the Citation button to see the relevant papers that document the detailed development and algorithm of iMethyl-PseAAC.

Step 7. Click on the Supporting Information button to download the benchmark dataset used to train and test the iMethyl-PseAAC predictor.

Caveat. To obtain the predicted result with the anticipated success rate, the entire sequence of the query protein rather than its fragment should be used as an input.

3. Results and Discussion

In statistical prediction, the following three cross-validation methods are often used to evaluate the anticipated accuracy of a predictor: independent dataset test, subsampling (K-fold cross-validation) test, and jackknife test [98]. However, as elucidated by a comprehensive review [24], among the three cross-validation methods, the jackknife test was deemed the least arbitrary and most objective because it could always yield a unique result for a given benchmark dataset and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (see, e.g., [60, 61, 90, 99–101]). Therefore, in this study, we also adopted the jackknife test to examine the prediction quality of the iMethyl-PseAAC predictor.

It is instructive to point out that the number of positive samples and that of negative samples in the current benchmark dataset, for either Arg- or Lys-methylation system, are highly imbalanced. As shown in Online Supporting Information S3 and Online Supporting Information S4, the number of negative samples is about seven times the number of the positive samples. A general approach to treat this kind of highly sample-imbalanced system is to randomly separate the large set into several subsets and make each of them have about the same size of the small set.

The details for the subsets thus obtained for the current Arg-methylation and Lys-methylation systems are given in Online Supporting Information S5 and Online Supporting Information S6, respectively.

The jackknife rates achieved by iMethyl-PseAAC for the Arg-methylation system and Lys-methylation system are given in Tables 1 and 2, respectively. As we can see from the two tables, the average accuracy achieved by iMethyl-PseAAC for the Arg-methylation system was 76.19% and that for the Lys-methylation system was 70.74%. Meanwhile, we can also see that the corresponding MCCs (cf. (16)) were 52.74% and 41.66%, respectively, indicating that the prediction accuracy of iMethyl-PseAAC was quite stable, fully consistent with its sensitivity Sn and specificity Sp.

Table 1: The metrics rates obtained by the jackknife test on the Arg-methylation system, where the positive dataset contains 185 samples (see Online Supporting Information S3), while the negative dataset consists of seven subsets with each containing 185 samples except for the 6th subset that contains 186 samples (see Online Supporting Information S5).

Negative subset	Acc (%)	MCC	Sn (%)	Sp (%)
1	72.16	0.45	64.32	80.00
2	78.65	0.57	79.46	77.84
3	73.24	0.47	66.49	80.00
4	78.71	0.57	78.38	79.03
5	77.30	0.55	70.81	83.78
6	75.68	0.51	75.68	75.68
7	77.57	0.56	67.57	87.57
Average	76.19	0.53	71.81	80.56

Table 2: The metrics rates obtained by the jackknife test on the Lys-methylation system, where the positive dataset contains 226 samples (see Online Supporting Information S4), while the negative dataset consists of seven subsets with each containing 217 samples except for the 5th subset that contains 216 samples (see Online Supporting Information S6).

Negative subset	Acc (%)	MCC	Sn (%)	Sp (%)
1	73.36	0.47	72.57	74.19
2	65.01	0.30	63.27	66.82
3	71.11	0.42	70.35	71.89
4	67.27	0.35	63.72	70.97
5	73.76	0.48	73.01	74.54
6	72.23	0.45	64.16	80.65
7	72.46	0.45	73.01	71.89
Average	70.74	0.42	68.58	72.99

To further demonstrate its power, let us compare iMethyl-PseAAC with the existing predictors in this area. Only those predictors with a publicly accessible web server were qualified to be included in this study. Thus, the comparison will be made among the three predictors whose web servers are BPB-PPMS [20], PMeS [23], and iMethyl-PseAAC. Also, the best way to compare them is through practical application. To realize this, let us construct two independent datasets. One was for comparing the accuracy in identifying the Arg-methylation sites, and the other for Lys-methylation. The former contains 75 samples of which 20 are positive and 55 negative (see Online Supporting Information S7), while the latter contains 40 samples of which 14 are positive and 26 negative (see Online Supporting Information S8). To avoid the memory effect or bias in favor with iMethyl-PseAAC, none of the samples in the two independent datasets occurs in the datasets used to train the iMethyl-PseAAC predictor.

Listed in Tables 3 and 4 were the outcomes obtained by the three web-server predictors on the two independent

Table 3: Comparison of iMethyl-PseAAC with the existing web-server predictors when tested for identifying Arg-methylation sites by the independent dataset (see Online Supporting Information S7).

Predictor	Acc (%)	MCC	Sn (%)	Sp (%)
PMeS[a]	76.00	0.45	70.00	78.18
BPB-PPMS[b]	93.33	0.83	85.00	96.36
iMethyl-PseAAC	97.33	0.94	100.00	96.36

[a]From [23].
[b]From [20].

Table 4: Comparison of iMethyl-PseAAC with the existing web-server predictors when tested for identifying Lys-methylation sites by the independent dataset (see Online Supporting Information S8).

Predictor	Acc (%)	MCC	Sn (%)	Sp (%)
PMeS[a]	65.00	0.35	78.57	57.69
BPB-PPMS[b]	70.00	0.36	64.29	73.08
iMethyl-PseAAC	75.00	0.60	100.00	61.54

[a]See footnote a of Table 3.
[b]See footnote b of Table 3.

datasets. As we can see from the two tables, the scores of the four metrics (cf. (16)) achieved by iMethyl-PseAAC were all remarkably higher than those by its counterparts except the rate of Sp for which iMethyl-PseAAC was tied with BPB-BPMS (see column 5 of Table 3) and about 11% lower than that of BPB-BPMS (see column 5 of Table 4). These results have clearly indicated that iMethyl-PseAAC is superior to its counterparts in predicting the Arg-methylation and Lys-methylation sites in proteins.

Finally, it is instructive to present an in-depth analysis to compare the experimental results with those reported in Step 4 of the "Web Server and User Guide." According to experimental observations, the protein (P62805) has 103 amino acid residues and 14 Arg sites, of which only the 1st Arg (or the one located at the sequence position 4) is methylated, while all the other 13 Arg residues (or those located at the sequence positions 18, 20, 24, 36, 37, 40, 41, 46, 56, 68, 79, 93, and 96) are not methylated. Thus, we have $N^+ = 1$ and $N^- = 13$ (cf. (16)). Since none of methylated Arg sites was incorrectly predicted as nonmethylated site and only one of the 13 nonmethylated Arg sites was incorrectly predicted as methylated sites, we have $N_-^+ = 0$ and $N_+^- = 1$. Substituting these data into (16), we obtain Sn = 1, Sp = 0.92, Acc = 0.93, and MCC = 0.68.

The 2nd protein (P68431) has 136 amino acid residues and 18 Arg residues, of which the first three Arg residues (or those located at the sequence positions 3, 9, and 18) are methylated according to experimental observations. Thus, we have $N^+ = 3$ and $N^- = 15$. Since none of the 3 methylated Arg sites was incorrectly predicted as nonmethylated and none of the 15 nonmethylated Arg sites was incorrectly predicted as methylated, we have $N_-^+ = 0$ and $N_+^- = 0$. Substituting these data into (16), we obtain Sn = 1, Sp = 1, Acc = 1, and MCC = 1, meaning that the predicted result by iMethyl-PseAAC in the aforementioned Step 4 for protein (P68431) is perfectly correct.

Similar analysis can also be extended for the Lys-methylation. For example, the protein (P62805) has 11 Lys sites, of which only the 5th Lys (or the one located at the sequence position 21) was the methylated and all the other Lys residues (or those located at the sequence positions 6, 9, 13, 17, 32, 45, 60, 78, 80, and 92) were not according to experimental observations. Accordingly, its 3rd and 4th Lys residues were overpredicted by iMethyl-PseAAC as methylated. Thus we have $N^+ = 1$, $N^- = 10$, $N_-^+ = 0$, and $N_+^- = 2$. Substituting these data into (16), we obtain Sn = 1, Sp = 0.80, Acc = 0.82, and MCC = 0.63.

The 2nd protein (P68431) has 13 Lys residues, of which only the 3rd Lys (or the one located at sequence position 15) and 12th Lys (or the one located at the sequence position 116) are not methylated while all the other Lys residues (or those located at 5, 10, 19, 24, 28, 37, 38, 57, 65, 80, and 123) are methylated according to experimental observations. Thus, we have $N^+ = 11$ and $N^- = 2$. Since none of the 11 methylated Lys sites was incorrectly predicted as nonmethylated site and only one of the 2 nonmethylated Lys sites was incorrectly predicted as the methylated site, we have $N_-^+ = 0$ and $N_+^- = 1$. Substituting these data into (16), we obtain Sn = 1, Sp = 0.5, Acc = 0.92, and MCC = 0.68.

4. Conclusion

To timely acquire the information of Arg- and Lys-methylation sites in proteins is important for studying epigenetic inheritance in depth, analyzing various human diseases, and developing new drugs. It is anticipated that the iMethyl-PseAAC predictor may become a very useful high throughput tool in this regard. Its user-friendly web server and the step-by-step guide can help users easily to get their desired data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors wish to thank the editor for taking time to edit this paper and thank the anonymous reviewers for their constructive comments, which were very useful in strengthening the presentation of this paper. This work was partially supported by the National Nature Science Foundation of China (nos. 31260273 and 61261027), the Jiangxi Provincial Foreign Scientific and Technological Co-operation Project (no. 20120BDH80023), Natural Science Foundation of Jiangxi Province, China (nos. 2010GQS0127, 20114BAB211013, 20122BAB211033, 20122BAB201044, and 20122BAB2010), the Department of Education of Jiangxi Province (GJJ12490), the LuoDi plan of the Department of Education of Jiangxi Province (KJLD12083), and the Jiangxi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

References

[1] C. Walsh, *Posttranslational Modifications of Proteins: Expanding Nature's Inventory*, Roberts and Company Publishers, Greenwood Village, Colo, USA, 2006.

[2] M. T. Bedford and S. Richard, "Arginine methylation: an emerging regulator of protein function," *Molecular Cell*, vol. 18, no. 3, pp. 263–272, 2005.

[3] W. K. Paik, D. C. Paik, and S. Kim, "Historical review: the field of protein methylation," *Trends in Biochemical Sciences*, vol. 32, no. 3, pp. 146–152, 2007.

[4] R. A. Varier and H. T. M. Timmers, "Histone lysine methylation and demethylation pathways in cancer," *Biochimica et Biophysica Acta*, vol. 1815, no. 1, pp. 75–89, 2011.

[5] X. Chen, F. Niroomand, Z. Liu et al., "Expression of nitric oxide related enzymes in coronary heart disease," *Basic Research in Cardiology*, vol. 101, no. 4, pp. 346–353, 2006.

[6] F. G. Mastronardi, D. D. Wood, J. Mei et al., "Increased citrullination of histone H3 in multiple sclerosis brain and animal models of demyelination: a role for tumor necrosis factor-induced peptidylarginine deiminase 4 translocation," *Journal of Neuroscience*, vol. 26, no. 44, pp. 11387–11396, 2006.

[7] A. Suzuki, R. Yamada, and K. Yamamoto, "Citrullination by peptidylarginine deiminase in rheumatoid arthritis," *Annals of the New York Academy of Sciences*, vol. 1108, pp. 323–339, 2007.

[8] V. D. Longo and B. K. Kennedy, "Sirtuins in aging and age-related disease," *Cell*, vol. 126, no. 2, pp. 257–268, 2006.

[9] C. Caldji, I. C. Hellstrom, T.-Y. Zhang, J. Diorio, and M. J. Meaney, "Environmental regulation of the neural epigenome," *The FEBS Letters*, vol. 585, no. 13, pp. 2049–2058, 2011.

[10] F. A. Champagne and J. P. Curley, "How social experiences influence the brain," *Current Opinion in Neurobiology*, vol. 15, no. 6, pp. 704–709, 2005.

[11] F. A. Champagne, I. C. G. Weaver, J. Diorio, S. Dymov, M. Szyf, and M. J. Meaney, "Maternal care associated with methylation of the estrogen receptor-α1b promoter and estrogen receptor-α expression in the medial preoptic area of female offspring," *Endocrinology*, vol. 147, no. 6, pp. 2909–2915, 2006.

[12] V. J. Felitti, R. F. Anda, D. Nordenberg et al., "Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the adverse childhood experiences (ACE) study," *The American Journal of Preventive Medicine*, vol. 14, no. 4, pp. 245–258, 1998.

[13] X. Zhang and S.-M. Ho, "Epigenetics meets endocrinology," *Journal of Molecular Endocrinology*, vol. 46, no. 1, pp. R11–R32, 2011.

[14] F.-M. Boisvert, J. Côté, M.-C. Boulanger, and S. Richard, "A proteomic analysis of arginine-methylated protein complexes," *Molecular & Cellular Proteomics*, vol. 2, no. 12, pp. 1319–1330, 2003.

[15] S.-E. Ong, G. Mittler, and M. Mann, "Identifying and quantifying in vivo methylation sites by heavy methyl SILAC," *Nature Methods*, vol. 1, no. 2, pp. 119–126, 2004.

[16] C. C. Wu, M. J. MacCoss, K. E. Howell, and J. R. Yates III, "A method for the comprehensive proteomic analysis of membrane proteins," *Nature Biotechnology*, vol. 21, no. 5, pp. 532–538, 2003.

[17] K. M. Daily, P. Radivojac, and A. K. Dunker, "Intrinsic disorder and prote in modifications: building an SVM predictor for methylation," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '05)*, vol. 5, pp. 475–481, November 2005.

[18] T. Huang, Z. S. He, W. R. Cui et al., "A sequence-based approach for predicting protein disordered regions," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 243–248, 2013.

[19] H. Chen, Y. Xue, N. Huang, X. Yao, and Z. Sun, "MeMo: a web tool for prediction of protein methylation modifications," *Nucleic Acids Research*, vol. 34, pp. W249–W253, 2006.

[20] J. Shao, D. Xu, S.-N. Tsai, Y. Wang, and S.-M. Ngai, "Computational identification of protein methylation sites through Bi-profile Bayes feature extraction," *PLoS ONE*, vol. 4, no. 3, Article ID e4920, 2009.

[21] D.-M. Shien, T.-Y. Lee, W.-C. Chang et al., "Incorporating structural characteristics for identification of protein methylation sites," *Journal of Computational Chemistry*, vol. 30, no. 9, pp. 1532–1543, 2009.

[22] L.-L. Hu, Z. Li, K. Wang et al., "Prediction and analysis of protein methylarginine and methyllysine based on Multisequence features," *Biopolymers*, vol. 95, no. 11, pp. 763–771, 2011.

[23] S. P. Shi, J. D. Qiu, X. Y. Sun, S. B. Suo, S. Y. Huang, and R. P. Liang, "PMeS: prediction of methylation sites based on enhanced feature encoding scheme," *PLoS ONE*, vol. 7, no. 6, Article ID e38772, 2012.

[24] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[25] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, no. 4, pp. 634–644, 2013.

[26] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[27] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[28] Y. Xu, X. J. Shao, L. Y. Wu, N. Y. Deng, and K.-C. Chou, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, article e171, 2013.

[29] K.-C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins," *The Journal of Biological Chemistry*, vol. 268, no. 23, pp. 16938–16948, 1993.

[30] K.-C. Chou, "Review: prediction of human immunodeficiency virus protease cleavage sites in proteins," *Analytical Biochemistry*, vol. 233, no. 1, pp. 1–14, 1996.

[31] K.-C. Chou, "A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase," *Protein Science*, vol. 4, no. 7, pp. 1365–1383, 1995.

[32] K.-C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[33] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering*, vol. 14, no. 2, pp. 75–79, 2001.

[34] K.-C. Chou and H.-B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[35] K.-C. Chou and C.-T. Zhang, "A correlation-coefficient method to predicting protein-structural classes from amino acid compositions," *European Journal of Biochemistry*, vol. 207, no. 2, pp. 429–433, 1992.

[36] J. J. Chou, "A formulation for correlating properties of peptides and its application to predicting human immunodeficiency

virus protease-cleavable sites in proteins," *Biopolymers*, vol. 33, no. 9, pp. 1405–1414, 1993.

[37] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[38] C.-T. Zhang and K.-C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Science*, vol. 1, no. 3, pp. 401–408, 1992.

[39] K.-C. Chou and C.-T. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions," *The Journal of Biological Chemistry*, vol. 269, no. 35, pp. 22014–22020, 1994.

[40] K.-C. Chou, "Does the folding type of a protein depend on its amino acid composition?" *The FEBS Letters*, vol. 363, no. 1-2, pp. 127–131, 1995.

[41] K.-C. Chou and D. W. Elrod, "Using discriminant function for prediction of subcellular location of prokaryotic proteins," *Biochemical and Biophysical Research Communications*, vol. 252, no. 1, pp. 63–68, 1998.

[42] G. P. Zhou and N. Assa-Munt, "Some insights into protein structural class prediction," *Proteins: Structure, Function and Genetics*, vol. 44, no. 1, pp. 57–59, 2001.

[43] Y. Gao, S. Shao, X. Xiao et al., "Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter," *Amino Acids*, vol. 28, no. 4, pp. 373–376, 2005.

[44] Y. Xu, J. Ding, L. Y. Wu, and K.-C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[45] K.-Y. Feng, Y.-D. Cai, and K.-C. Chou, "Boosting classifier for predicting protein domain structural class," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 213–217, 2005.

[46] T. B. Thompson, K.-C. Chou, and C. Zheng, "Neural network prediction of the HIV-1 protease cleavage sites," *Journal of Theoretical Biology*, vol. 177, no. 4, pp. 369–379, 1995.

[47] X. Xiao, P. Wang, and K.-C. Chou, "iNR-physchem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix," *PLoS ONE*, vol. 7, no. 2, Article ID e30869, 2012.

[48] K. K. Kandaswamy, K.-C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[49] K.-C. Chou and Y.-D. Cai, "Prediction of protease types in a hybridization space," *Biochemical and Biophysical Research Communications*, vol. 339, no. 3, pp. 1015–1020, 2006.

[50] K.-C. Chou and H.-B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," *Journal of Proteome Research*, vol. 5, no. 8, pp. 1888–1897, 2006.

[51] K.-C. Chou and H.-B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1728–1734, 2007.

[52] X. Xiao, P. Wang, and K.-C. Chou, "GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions," *Molecular BioSystems*, vol. 7, no. 3, pp. 911–919, 2011.

[53] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function and Genetics*, vol. 43, no. 3, pp. 246–255, 2001.

[54] Y. Zhao, C. Pinilla, D. Valmori, R. Martin, and R. Simon, "Application of support vector machines for T-cell epitopes prediction," *Bioinformatics*, vol. 19, no. 15, pp. 1978–1984, 2003.

[55] G. L. Zhang, A. M. Khan, K. N. Srinivasan, J. T. August, and V. Brusic, "Neural models for predicting viral vaccine targets," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 5, pp. 1207–1225, 2005.

[56] L. Nanni and A. Lumini, "A new encoding technique for peptide classification," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3185–3191, 2011.

[57] L. Nanni, S. Brahnam, and A. Lumini, "Wavelet images and Chou's pseudo amino acid composition for protein classification," *Amino Acids*, vol. 43, no. 2, pp. 657–665, 2012.

[58] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "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 Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[59] D. N. Georgiou, T. E. Karakasidis, and A. C. Megaritis, "A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory," *The Open Bioinformatics Journal*, vol. 7, pp. 41–48, 2013.

[60] Y.-K. Chen and K.-B. Li, "Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 318, pp. 1–12, 2013.

[61] M. Khosravian, F. K. Faramarzi, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods," *Protein & Peptide Letters*, vol. 20, no. 2, pp. 180–186, 2013.

[62] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[63] S. Mondal and P. P. Pai, "Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction," *Journal of Theoretical Biology*, vol. 356, pp. 30–35, 2014.

[64] H. Mohabatkar, M. M. Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[65] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[66] S.-X. Lin and J. Lapointe, "Theoretical and experimental biology in one," *Journal of Biomedical Science and Engineering*, vol. 6, pp. 435–442, 2013.

[67] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[68] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[69] P. Du, S. Gu, and Y. Jiao, "PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *International Journal of Molecular Sciences*, vol. 15, no. 3, pp. 3495–3506, 2014.

[70] K.-C. Chou, "Structural bioinformatics and its impact to biomedical science," *Current Medicinal Chemistry*, vol. 11, no. 16, pp. 2105–2134, 2004.

[71] K.-C. Chou, "The convergence-divergence duality in lectin domains of selectin family and its implications," *The FEBS Letters*, vol. 363, no. 1-2, pp. 123–126, 1995.

[72] A. A. Schäffer, L. Aravind, T. L. Madden et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[73] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular BioSystems*, vol. 8, no. 2, pp. 629–641, 2012.

[74] J. Deng, "Introduction to grey system theory," *The Journal of Grey System*, vol. 1, no. 1, pp. 1–24, 1989.

[75] W. R. Atchley and J. Zhao, "Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins," *Molecular Biology and Evolution*, vol. 24, no. 1, pp. 192–202, 2007.

[76] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drüke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, pp. 6395–6400, 2005.

[77] J. Peng, D. Schwartz, J. E. Elias et al., "A proteomics approach to understanding protein ubiquitination," *Nature Biotechnology*, vol. 21, no. 8, pp. 921–926, 2003.

[78] M. Matsumoto, S. Hatakeyama, K. Oyamada, Y. Oda, T. Nishimura, and K. I. Nakayama, "Large-scale analysis of the human ubiquitin-related proteome," *Proteomics*, vol. 5, no. 16, pp. 4145–4151, 2005.

[79] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Research*, vol. 36, no. 1, pp. D202–D205, 2008.

[80] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 28, no. 1, article 374, 2000.

[81] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[82] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 321–331, 1999.

[83] J. Liu, H. Tan, and B. Rost, "Loopy proteins appear conserved in evolution," *Journal of Molecular Biology*, vol. 322, no. 1, pp. 53–64, 2002.

[84] P. Tompa, "Intrinsically unstructured proteins," *Trends in Biochemical Sciences*, vol. 27, no. 10, pp. 527–533, 2002.

[85] M. Sickmeier, J. A. Hamilton, T. LeGall et al., "DisProt: the database of disordered proteins," *Nucleic Acids Research*, vol. 35, no. 1, pp. D786–D793, 2007.

[86] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein in intrinsic disorder," *BMC Bioinformatics*, vol. 7, article 208, 2006.

[87] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[88] M. A. Hearst, "Support vector machines," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[89] K.-C. Chou and Y.-D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *The Journal of Biological Chemistry*, vol. 277, no. 48, pp. 45765–45769, 2002.

[90] S. Wan, M.-W. Mak, and S.-Y. Kung, "GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 323, pp. 40–48, 2013.

[91] P.-M. Feng, W. Chen, H. Lin, and K.-C. Chou, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[92] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical Journal*, vol. 84, no. 5, pp. 3257–3263, 2003.

[93] N. Cristianini and J. Shawe-Taylor, *An Introduction of Support Vector Machines and Other Kernel-Based Learning Methodds*, Cambridge University Press, Cambridge, UK, 2000.

[94] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[95] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[96] K.-C. Chou and H.-B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 1, no. 2, pp. 63–92.

[97] K.-C. Chou, "Prediction of protein signal sequences and their cleavage sites," *Proteins: Structure, Function and Genetics*, vol. 42, no. 1, pp. 136–139, 2001.

[98] K.-C. Chou and C.-T. Zhang, "Review: prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[99] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[100] S. Mei, "Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning," *Journal of Theoretical Biology*, vol. 310, pp. 80–87, 2012.

[101] S.-W. Zhang, Y.-L. Zhang, H.-F. Yang, C.-H. Zhao, and Q. Pan, "Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies," *Amino Acids*, vol. 34, no. 4, pp. 565–572, 2008.