Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, 2019, *20*, 275-292

RESEARCH ARTICLE

iMethylK-PseAAC: Improving Accuracy of Lysine Methylation Sites Identification by Incorporating Statistical Moments and Position Relative Features into General PseAAC *via* Chou's 5-steps Rule

Sarah Ilyas[1], Waqar Hussain[1], Adeel Ashraf[1], Yaser Daanial Khan[1,*], Sher Afzal Khan[2,4] and Kuo-Chen Chou[3]

[1]*Department of Computer Science, School of Systems and Technology, University of Management and Technology, P.O. Box 10033, C-II, Johar Town, Lahore 54770, Pakistan;* [2]*Faculty of Computing and Information Technology in Rabigh, Jeddah, 21577, KSA;* [3]*Gordon Life Science Institute, Boston, MA 02478, USA;* [4]*Department of Computer Sciences, Abdul Wali Khan University, Mardan, Pakistan*

ARTICLE HISTORY

Received: May 08, 2019
Revised: July 02, 2019
Accepted: July 26, 2019

DOI: 10.2174/1389202920666190809095206

Abstract: ***Background*:** Methylation is one of the most important post-translational modifications in the human body which usually arises on lysine among the most intensely modified residues. It performs a dynamic role in numerous biological procedures, such as regulation of gene expression, regulation of protein function and RNA processing. Therefore, to identify lysine methylation sites is an important challenge as some experimental procedures are time-consuming.

***Objective*:** Herein, we propose a computational predictor named iMethylK-PseAAC to identify lysine methylation sites.

***Methods*:** Firstly, we constructed feature vectors based on PseAAC using position and composition relative features and statistical moments. A neural network is trained based on the extracted features. The performance of the proposed method is then validated using cross-validation and jackknife testing.

***Results*:** The objective evaluation of the predictor showed accuracy of 96.7% for self-consistency, 91.61% for 10-fold cross-validation and 93.42% for jackknife testing.

***Conclusion*:** It is concluded that iMethylK-PseAAC outperforms the counterparts to identify lysine methylation sites such as iMethyl-PseACC, BPB-PPMS and PMeS.

Keywords: Methylation, lysine methylation, PseAAC, statistical moments, 5-steps rule, prediction.

1. INTRODUCTION

The overall process of generating new proteins is called protein synthesis, which is governed by the loss of cellular proteins with the help of dissemination. The process of synthesizing a protein from an mRNA is known as translation. Post-translational modification is one of the most significant modifications in biology that highlights the covalent and most commonly enzymatic modification of proteins throughout the process of protein biosynthesis [1]. Covalent post-translational modifications (PTMs) of proteins produce a complicated layer of the proteome. High-throughput proteomics with focused investigation on site-specific PTM and protein adjusting enzymes have revealed insight into the extent of these alterations over a diverse range of organisms.

Among the 20 amino acids, lysine is a standout amongst the most intensely modified residues. Nowadays, lysine residues are identified to be covalently altered by acetyl, hydroxyl, glycosyl, propionyl, butyryl, crotonyl, malonyl, succinyl, and methyl. Among all these modifications, lysine methylation is the most complex and difficult PTM that has, in any case, the possibility to transform the capacity of the altered protein [2]. A well-known PTM, which includes the modification of up to three methyl groups to the e-amine of a lysine residue, has attracted remarkable attention in recent years. Lysine methylation has been observed in both atomic and cytoplasmic proteins and is currently viewed as a common modification in eukaryotes, prokaryotes and archaea.

Protein methylation is a form of post-translational modification which arises on lysine residues and is basically catalysed by enzymes [3]. The methyltransferases transform amino acids by adding a methyl group as shown in Fig. **1**. Methylation has been mostly observed in the histones, where the transferal of methyl groups is catalysed by histone methyltransferases. Histones which are methylated on solid residues can perform epigenetically to suppress or activate gene expression [4]. It takes place on nitrogen side-chains in arginine and lysine residues and also on the carboxyl-terminus of different proteins. Protein methylation which occurs on nitrogen atoms at the N-terminus cannot be inverted and reduces the protein activity, whereas methylation on

*Address correspondence to this author at the Department of Computer Science, School of Systems and Technology, University of Management and Technology, P.O. Box 10033, C-II, Johar Town, Lahore, Pakistan;
Tel: +923054440271; E-mail: yaser.khan@umt.edu.pk

1389-2029/19 $58.00+.00

the C-terminus residues can increase the activity of a protein which is regularly performed in a body [5]. The process of protein methylation is basically the accumulation of a methyl group in proteins that usually takes place on the lysine residues in the protein sequence. Protein methylation is involved in the modification of substantial metals, regulation of gene expression, regulation of protein function, and RNA processing. De-methylation is the event in contrast to methylation. Lysine can be methylated once, twice, or thrice through lysine methyltransferases. Lysine methylation acts as an essential part in how histones work together with proteins [6].

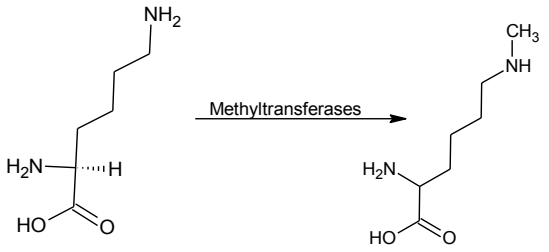

Fig. (1). Process of methyltransferase.

Proper methylation is essential for dangerous reactions in the body. It can cause a number of serious health issues including heart diseases [7], brain diseases such as depression and migraines [8], cancer [9, 10], and many more. It can occur on the nitrogen atoms of either back or sidechain of lysine (K), Arginine (R), histidine (H), alanine (A), proline (P), and glutamine (Q) residues, oxygen atoms of aspartate (D) and glutamate (E) residues as well as on the sulphur atom of cysteine (C) residue [11]. Among all of them, lysine (K) is the most frequently observed amino acid of methylation. Lysine residues accept up to three methyl groups forming mono-, di-and tri-methylated proteins, which are the main types of lysine methylation and are severe in several biological procedures [12].

It is an important challenge to identify lysine methylation sites experimentally without disturbing the overall sequence of the protein structure. Such type of modifications can be performed experimentally which is a time-consuming and expensive technique. In this regard, various researchers have used computational methodologies to identify lysine methylation by using the primary structure of the protein. In recent decades, researchers have made various contributions to improve numerous computational models in order to predict an element of a protein [13-22]. A number of computational models have been proposed for the prediction of PTM sites [23-29]. In the last few years, many studies have been conducted by previous researchers in the field of bioinformatics and computational biology, which help in identifying the function and characteristics of proteins [2, 14-16, 19-21, 23-27, 30-60]. Besides these, various papers have been reported targeting the prediction of PTM [2, 17, 20, 21, 23-28, 30-35, 37-51, 53-58, 61-68]. Struggles have been made for the prediction of protein domains. A novel method was introduced by merging the methods of RF, mRMR and IFS including the features of physicochemical and biological properties, sequence conservation, residual disorder, and solvent accessibility [69]. Sun and others noticed that the methylation position of CpG sites is another essential problem while studying gene regulation and indicates a robust relationship with the transcription factor binding sites (TFBS) involved. They established the models that are used to compare the modifications among regions and tissues [70]. Shao *et al.* established a predictor named BPB-PPMS, to identify methylation sites *via* Bi-profile Bayes feature extraction approach. It was only designed to predict the methylation position for lysine and arginine residues [71]. In this regard, another method was proposed using feature selection technique and nearest neighbour algorithm by Hu *et al.* [72]. This method works as a useful tool for biologists to find the possible methylated sites of proteins. A method called PMeS was developed by Shi *et al*. to increase the prediction of methylation sites depending upon an enhanced feature encoding scheme and SVM. When PMeS is used with other existing approaches, it provides enhanced predictive performance and greater strength [73]. Valavanis *et al.* studied the computational structure for the logarithmic ratio of methylated as well as un-methylated sites, quoted as M-value. The consequences presented here are linked to those derived by applying typical pre-processing and statistical selection procedures [74]. Li and co-workers in 2014 proposed a novel predictor called Methyl-SVMIACO, depending upon the Support Vector Machine and enhanced IACO Algorithm, to discover methylation sites. This algorithm is basically used to find the best feature subgroup and a parameter of SVM, whereas SVM is active in finding the methylation sites [75]. Qiu *et al*. determined a new predictor called iMethyl-PseAAC. In this prediction scheme, a sample of the peptide was framed *via* 346-dimensional vector designed as a result of combining physicochemical, biochemical and physical disorder data with overall pseudo amino acid composition [48]. Karagod and Sinha discovered a machine learning structure that produces greater accuracy than the previous MS-SPCA and EVORA methods for predicting the phases of different diseases like cancer using CpG data [76].

Although various efforts have been made for the prediction of methylation sites which have its own merit but problems still exist; they all need perfection to develop improved methods for the following characteristics: (i) the standard dataset, which was considered by previous researchers, needs to be restructured or updated by combining some novel and computational-based data, (ii) can be boosted by eliminating redundancy and duplicate sequences as compared to the existing ones, (iii) constructing statistical samples that are totally dependent on the sequence data, (iv) accuracy of the current models should be enhanced as some earlier models, (v) further enhancing the prediction quality as compared to others by using the computational method. For efficient prediction of methylation, it would be suitable to propose more exact models after considering these inefficiencies.

In the present study, a computational method is proposed using a broad feature extraction procedure for the prediction of Lysine (K) methylation. The dataset for this prediction is collected from two well-known sources namely, dbPTM and UniProt. Features which are relevant to the post-translational modification sites are extracted. A Neural Network (NN) is trained based on the extracted features using backpropagation technique [18, 77-79]. Consequently, the validation of the proposed model is achieved with various methods consisting of accuracy metrics, Mathew's correla-

tion coefficient (MCC), Sensitivity (Sn), Specificity (Sp) and Cross-validation testing. The first aim of this study is to make it easier for experimental researchers to acquire their expected data while the second is to simulate previous studies related to some extent. To understand these two aims, we follow the 5-step rule [80]. The whole process is carried out with the aid of Chou's 5-step rule [81] which is followed in current studies [16, 20, 21, 31, 82-88]. As demonstrated by a series of recent publications [17, 19, 22, 23, 31, 33, 36, 38, 41, 65, 84, 89-104] and summarized in two comprehensive review papers [80, 105], to develop a really useful predictor for a biological system, one needs to follow Chou's 5-step rule: (1) select or construct a valid benchmark dataset to train and test the predictor; (2) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm to conduct the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (5) establish a user-friendly web-server for the predictor that is accessible to the public. Papers presented for developing a new sequence-analyzing method or statistical predictor by observing the guidelines of Chou's 5-step rules have the following notable merits: (1) crystal clear in logic development, (2) completely transparent in operation, (3) easy to repeat the reported results by other investigators, (4) high potential in stimulating other sequence-analyzing methods, and (5) very convenient to be used by the majority of experimental scientists.

2. MATERIALS AND METHODS

In this section, the overall process adopted for the prediction of lysine methylation is described in detail, as shown in Fig. **2**. It involves three basic steps *i.e.* data collection, feature extraction, training neural network. In the first step, the benchmark datasets are collected from a renowned online database of proteins called UniProt [106] as well as dbPTM [107]. Sub-sequences which were most relevant to lysine methylation were extracted. After the extraction of the most relevant sequences, duplication in them was removed and carefully selected sequence data was used for training purpose. In the second step, feature extraction technique was applied to get a range of feature vectors (FV). In the end, the input matrix containing feature vectors (FVs) and an output matrix containing expected output were used to train the Neural Network (NN) *via* backpropagation technique. Furthermore, the trained model was used for the prediction of methylation sites. Afterwards, the trained model was validated on test dataset that will be explained under the validation part.

2.1. The Benchmark Datasets

The benchmark dataset collection is the first step according to the 5-step rule. To develop a statistical model, it is significant to initiate a consistent and standard dataset to train and test that model. The accuracy of the trained model would be fully unpredictable and meaningless if the standard dataset comprises of errors. The dataset was gathered from UniProt and dbPTM to predict lysine methylation sites. Universal Protein Resource (UniProt) is a broad, high-quality and freely available resource of protein sequence and annotation data whereas dbPTM is a unified resource for protein post-translational modifications (PTMs) data. DbPTM is a comprehensive database that combines experimentally verified PTMs from various databases and interprets potential PTMs for all UniProtKB protein entries. To collect positive samples, the database was downloaded from dbPTM in which a total of 226 positive samples were accumulated. From UniProt, 569 samples were collected, thus a total of 795 sequences were gathered.

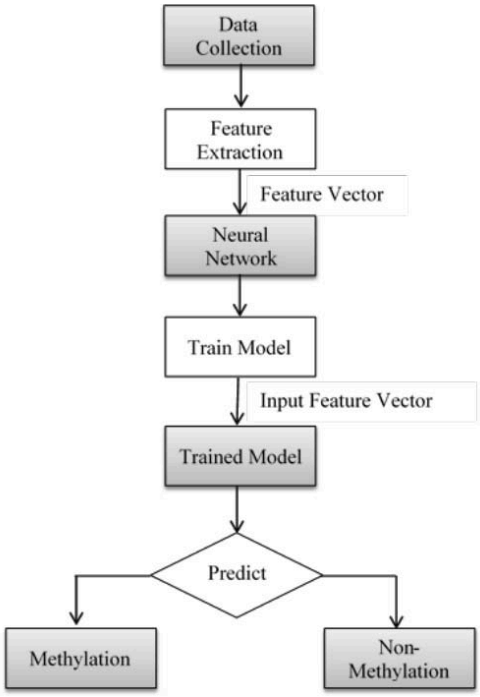

Fig. (2). Prediction model flow for iMethylK-PseAAC. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Moreover, for the negative dataset, a converse query (same query as for the positive dataset but initiating with a NOT) was produced. The sequences observed were those that had clear experimental annotations with field PTM/Processing about their Lysine Methylation sites. Furthermore, the sequences that we found were only those proteins which contained the term, modified residues (FT). To further find out the reliability of the dataset, only those proteins were acquired where the observation was dependent on experimental assertion. Accordingly, the sequences for negative methylation sites against raw data involved 135826 instances. After obtaining these sites, 2000 negative sites were randomly selected. The composed dataset was filtered by eliminating the redundant sequences using CD-HIT [108], with a threshold of 0.6 (60% similarity). After that, 670 positive instances and 984 negative sites of lysine methylation were left in the dataset. Taking into account Chou's scheme [81], a protein containing lysine site can be expressed as:

$$K_{\rho}(B) = M_{-\rho}M_{-(\rho-1)}\cdots M_{-2}M_{-1}\coprod M_{+1}M_{+2}\cdots M_{+(\rho-1)}M_{+\rho} \quad (1)$$

Amino acid code *K* is denoted by $\coprod$ in this equation, the character ρ is an integer, $M_{-\rho}$ represents ρ-th upstream

amino acid residue from the centre, $M_{+\rho}$ represents ρ +th downstream amino acid residue from the centre. $(2\rho+1)$ a tuple can be illustrated in 2 types:

$$K_v(\amalg) \varepsilon \begin{Bmatrix} K_v^+(\amalg) \\ K_v^-(\amalg) \end{Bmatrix} \quad (2)$$

In the following condition, $K_v^+(\amalg)$ holds if the centre is methyllysine site, if it is not then $K_v^-(\amalg)$holds. Set theory represents symbol 'ε' as "a member of".

Testing and training dataset is developed for the statistical prediction model. The model is trained using training dataset and then tested using testing dataset. The model is extensively illustrated [22], explaining that there is no compelling reason to isolate a benchmark dataset into two subsets if jackknife and cross-validation tests are used for testing the prediction model because the result acquired in this way is from a combination of many different independent dataset results. In this research paper, the ideal value of ρ for the test is 20, while the dataset has $(2\rho+1)$=41 residues. Considering all this, the dataset was minimized to

$$\mathbf{T} = \mathbf{T}^+ \cup \mathbf{T}^- \quad (3)$$

In the equation, T^+ holds 670 positive samples, T^- holds 984 negative samples and $\cup$ represents "union of two sets". In total, 670+984 = 1654 samples are included in the benchmark dataset (Supplementary information **S1**). Frequency plots for positive and negative sequences are presented in Figs. (**3** and **4**), respectively [109]. The significance of the length of neighbouring residues was based on examining and testing in order to get the best peak result.

2.2. Feature Extraction

For assistance in feature vector construction, Chou's computational model sample formation was implemented [110]. A feature is a numerical and computable property of protein represented as n-dimension by the vector. A feature vector represents multiple properties relevant to the protein sequence. To study the properties of the protein, the construction of the feature vector holds the primary position. An array of amino acids is utilized to develop a feature vector that increases the probability of site prediction in protein. Protein's performance is determined by the location of amino acid; a slight change in the location modifies protein qualities [111]. Feature vector sequences represented by the feature vector are broadly utilized in predicting different structural characteristics.

2.2.1. Site Vicinity Vector

Site vicinity vector is determined in terms of a sub-group of the protein sequence. α_q shows probable post-translational modification and its neighbouring amino acids are represented as:

$$\mathcal{P} = \{\alpha_1 \cdots \alpha_{q-2}, \alpha_{q-1}, \alpha_q, \alpha_{q+1}, \alpha_{q+2}, \cdots \alpha_n\} \quad (4)$$

It also comprises possible PTM sites with its neighbouring residues represented as:

$$[\alpha_{q-s} \cdots \alpha_{q-2}, \alpha_{q-1}, \alpha_q, , \alpha_{q+1}\alpha_{q+2}, \cdots \alpha_{q+s}] \quad (5)$$

In this equation, s is known as a smaller integer number and is ideally chosen *via* examining and experimentation. The site vicinity vector frames a segment of the comprehensive feature vector which is allocated remarkable numerical qualities replacing all residue positions. Only twenty amino acids are important for the extraction of feature vectors where each amino acid is allotted an exclusive integer.

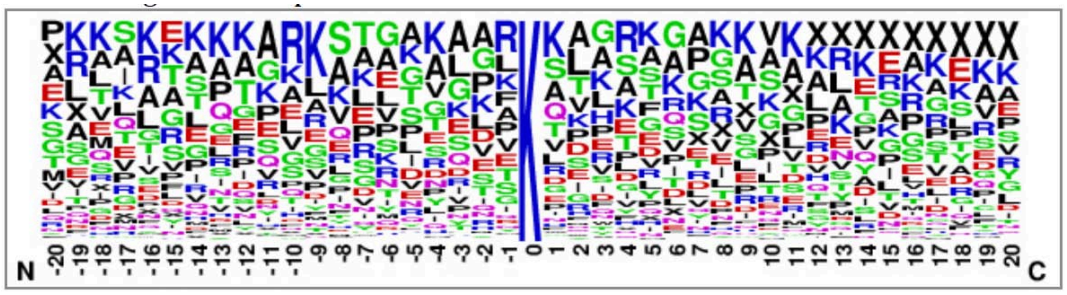

Fig. (3). Sequence diagram for (+ve) methyl-lysine sites. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

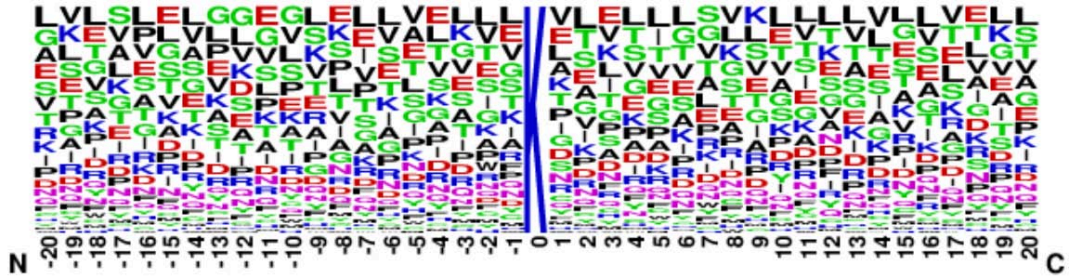

Fig. (4). Sequence diagram for (-ve) methyl-lysine sites. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

2.2.2. Statistical Moments

Statistical moments are the quantitative measures that are essentially used to represent an accumulation of data. A moment is a specific quantitative measure of the shape of a set of points. Researchers proved that statistical moments are suitable to make features from a known pattern. Various researchers have used these moments to capture the main features and describe the functionalities of a particular pattern [18, 77, 79]. The resolution to the proposed problem is pursued with the help of several moments such as raw moments, central moments, and Hahn moments along with the origin and centroid of the data as used in several studies [18, 77, 79]. In a recent study, it has been observed that discrete orthogonal moments produced better results than continuous orthogonal moments for discrete and quantized data. These orthogonal moments have the ability to transform the object illustrations with the lowest amount of loss of data.

A protein sequence is denoted as:

$$\mathcal{P} = \{a1, a2, a3, \dots\dots\dots, a\mathbb{k}\} \tag{6}$$

With a specific end goal to calculate two-dimensional moments, the one-dimensional design is reformed into a two-dimensional design by using a row-major scheme. The length of the two-dimensional matrix is calculated by taking the square root of the length of the protein.

$$n = \lceil \sqrt{k} \rceil \tag{7}$$

Where n is the measurement of the two-dimensional matrix and k is the length of the protein.

Furthermore, to adjust all the components of the protein sequence $\mathcal{P}$, a new matrix $\mathcal{P}'$ is designed along with $n * n$ dimensions.

$$P' = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nn} \end{bmatrix} \tag{8}$$

A function ϖ is used to convert the matrix $\mathcal{P}$ into $\mathcal{P}'$.

$$\varpi(a_u) = \omega_{ij} \tag{9}$$

Where $i = \frac{u}{v} + 1$ and $j = u \ mod \ v$ if $\mathcal{P}'$ is populated in a sequence substantial way.

The substance of two-dimensional matrix $\mathcal{P}'$ is considered to calculate the moments till order 3; the raw/crude moments are calculated from the given expression.

$$\mathbb{M}_{ij} = \sum_{\ell=1}^{v} \sum_{m=1}^{v} \ell^i m^j \, \omega_{\ell m} \tag{10}$$

Where, $i + j$ is the direction of the moment. The calculated moments are listed as:

$\mathbb{M}_{00}, \mathbb{M}_{01}, \mathbb{M}_{10}, \mathbb{M}_{11}, \mathbb{M}_{02}, \mathbb{M}_{20}, \mathbb{M}_{12}, \mathbb{M}_{21}, \mathbb{M}_{30}, \mathbb{M}_{03}$ up to direction/order 3.

The central moments use the centroid of the data as the reference point. These points can be represented as $\bar{y}$, $\bar{z}$, where,

$$\bar{y} = \mathbb{M}_{10}/\mathbb{M}_{00} \ and \ \bar{z} = \mathbb{M}_{01}/\mathbb{M}_{00} \tag{11}$$

It can be calculated after the calculation of raw moments by using the given equation,

$$\eta_{ij} = \sum_{\ell=1}^{v} \sum_{m=1}^{v} (\ell - \bar{y})^i \, (m - \bar{z})^j \omega_{\ell m} \tag{12}$$

Hahn moments can be computed after the transformation of 1D representation $\mathcal{P}$ into a 2D square matrix $\mathcal{P}'$. 2D Hahn moments require a square matrix as 2D input data. Hahn polynomial of direction/order n is known as:

$$\mathrm{H}_n^{a,b}(r, \mathbb{W}) = (\mathbb{W} + \mathbb{X} - 1)_n (\mathbb{W} - 1)_n \times \sum_{K=0}^{n} (-1)^K \frac{(-n)_K (-r)_K (2\mathbb{W} + a + b - n - 1)_K}{(N + b - 1)_K (\mathbb{W} - 1)_K} \frac{1}{K!} \tag{13}$$

The above equation utilizes the pochhammer representation summed up as,

$$(a)_k = a.(a+1) \dots\dots (a+k+1) \tag{14}$$

And can also be simplified *via* the Gamma Operator

$$(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{15}$$

Therefore, raw/crude values of orthogonal Hahn moments are generally mounted by utilizing a weight function and also a square root as given below:

$$\widetilde{H_n^{a,b}}(r, \mathbb{W}) = h_n^{a,b}(r, \mathbb{W}) \sqrt{\frac{p(r)}{d_n^2}}, \quad n = 0,1, \dots\dots, N-1 \tag{16}$$

While,

$$p(r) = \frac{\Gamma(a+r+b)(b+r+1)(a+b+r+1)_{\mathbb{W}}}{(a+b+2r+1)n!(\mathbb{W}-r-1)!} \tag{17}$$

Standardized Hahn moments for 2D discrete matrix are calculated by the given expression,

$$H_{ij} = \sum_{m=0}^{\mathbb{W}-1} \sum_{\ell=0}^{\mathbb{W}-1} \beta_{ij} \, \widetilde{h_i^{a,b}}(m, \mathbb{W}) \widetilde{h_j^{a,b}}(\ell, \mathbb{W}), m, n = 0,1, \dots\dots, N-1 \tag{18}$$

2D discrete orthogonal moments such as raw moments, central moments and Hahn moments are calculated up to direction/order 3. The whole procedure is carried out by following the method defined in various studies [13, 17, 19-23, 61, 97].

2.2.3. Position Relative and Reverse Position Relative Incidence Matrix

The initial phase in feature extraction is to calculate the matrix form of the input protein probe. For this reason, the length of the protein sequence is utilized to produce PRIM and RPRIM. These matrices are then utilized for the figuring of moments through which included vectors are shaped. A protein sequence S with the addition of N amino acid residues is characterised through PRIM as shown below:

$$S_{PRIM} = \begin{bmatrix} \mathrm{B}_{1\to1} & \mathrm{B}_{1\to2} & \cdots & \mathrm{B}_{1\to j} & \cdots & \mathrm{B}_{1\to20} \\ \mathrm{B}_{2\to1} & \mathrm{B}_{2\to2} & \cdots & \mathrm{B}_{2\to j} & \cdots & \mathrm{B}_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \mathrm{B}_{i\to1} & \mathrm{B}_{i\to2} & \cdots & \mathrm{B}_{i\to j} & \cdots & \mathrm{B}_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \mathrm{B}_{N\to1} & \mathrm{B}_{N\to2} & \cdots & \mathrm{B}_{N\to j} & \cdots & \mathrm{B}_{N\to20} \end{bmatrix} \tag{19}$$

RPRIM can also be represented as:

$$S_{RPRIM} = \begin{bmatrix} Z_{1\to1} & Z_{1\to2} & \cdots & Z_{1\to j} & \cdots & Z_{1\to20} \\ Z_{2\to1} & Z_{2\to2} & \cdots & Z_{2\to j} & \cdots & Z_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ Z_{i\to1} & Z_{i\to2} & \cdots & Z_{i\to j} & \cdots & Z_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ Z_{N\to1} & Z_{N\to2} & \cdots & Z_{N\to j} & \cdots & Z_{N\to20} \end{bmatrix} \quad (20)$$

In the given protein sequence, the indication of the gain of *ith* position residue is determined by $B_{i\to j}$ for PRIM and $Z_{i\to j}$ for RPRIM. In the genetic evolutionary procedure, this gain is replaced by amino acid form. The values of j = 1, 2. . . 20 are the presentation of the sequential order of 20 native amino acid residues. The method is further defined in other studies [13, 17, 20-22].

2.2.4. Frequency Matrix

Another matrix called a frequency matrix is designed which covers the information about the composition of protein structure. The main purpose of using this matrix is that it basically extracts the information of the sequence which has previously been mined into position relative incidence matrix (PRIM). The matrix is shown as:

$$\mathfrak{f} = \{\tau_1, \tau_2, \tau_3, \dots, \tau_{20}\} \quad (21)$$

Where τ_i shows the frequency of occurrence of *ith* native amino acid.

2.2.5. Accumulative Absolute Position Incidence Vector (AAPIV)

The frequency matrix was computed for extracting the compositional information but it did not provide the relative positions of residues. Therefore, a new matrix is formed named Accumulative Absolute Position Incidence matrix (AAPIV). AAPIV has a length of 20 elements where each element grasps the sum of all the ordinal values occurring in the primary sequence at their respective locations.

Consider AAPIV to be represented as:

$$\mathcal{K} = \{\mu_1, \mu_2, \mu_3, \dots, \mu_{20}\} \quad (22)$$

Therefore, *ith* element of AAPIV is calculated as given below:

$$\mu_i = \sum_{\mathcal{K}=1}^{n} \mathcal{P}_{\mathcal{K}} \quad (23)$$

Where $\mathcal{P}_{\mathcal{K}}$ represents the position of the occurrence of amino acid residues in the sequence. Further information regarding this protocol can be found in other studies [13, 17, 20-22].

2.2.6. Reverse Accumulative Absolute Position Incidence Vector (RAAPIV)

In order to extract deep and obscure information about the relative positioning of each amino acid residue, a reverse accumulative absolute position incidence vector (RAAPIV) is used. RAAPIV is developed by reversing the primary sequence and then producing AAPIV by using that reversed sequence.

Hence, RAAPIV having 20 elements is represented as:

$$Å = \{\eta_1, \eta_2, \eta_3, \dots \eta_{20}\} \quad (24)$$

RAAPIV is calculated on the same protocol as AAPIV.

2.3. Training of Operational Algorithm

The neural network is the most prevailing technique which is used to resolve decision problems. This network is composed of an enormous number of interrelated processing elements called neurons, which works correspondingly to solve a precise decision problem. Neural networks process the information in a similar manner as the human brain. As it takes the information from the environment and obtains experience, the neural network also embraces the same method. It takes many inputs as well as produce one output depending on the knowledge gained after each input throughout the training procedure. When the training procedure is over, the neural network apparently performs in a manner that categorises each known input with a suitable accuracy measure.

Throughout the overall learning procedure of the neural network, the basic goal is to decrease an error. This network modifies its weights during each iteration, in a way that the error is reduced between the preferred output and the real output which results in enhanced learning and improved accuracy. The neural network starts its training based on two values including the input values along with the initial weights that are given to the network as shown in Fig. **5**.

Neural networks are mostly referred to in terms of their depth, including the number of layers they have between input and output, called the hidden layers. They can also be labelled by the number of hidden nodes the model has or in terms of how many inputs and outputs each node has. Variations in the Neural Network (NN) model allow various forms of forward and backward propagation of information among iterations. In this study, a Multi-layered Back Propagation Neural Network (MBPNN) model has been used to tackle the research problem as used in a study [22]. The model is presented as given below in Fig. **6**.

The data sets were constructed containing positive and negative samples, and a feature vector (FV) is then constructed using the datasets for the prediction of Lysine Methylation sites consisting of a large number of coefficients. These two FVs are then merged to form an input matrix whereas each input vector is considered as both positive and negative samples in an additional output matrix. These two matrices are employed to train the Multi-Layer Neural Network. The input matrix iterates the input to the neural network while the output matrix is used to compute the errors through backpropagation methodology. To increase the prediction accuracy and reduce an error rate, gradient descent algorithm and adaptive learning rate were used.

2.3.1. Gradient Descent and Adaptive Learning

Backpropagation neural network uses the gradient descent algorithm. It makes several attempts to reduce its error along its gradient to increase the overall performance of the network. Gradient descent is a simple optimization technique that can be used to solve machine learning problems. It is an algorithm used to find the parameter values of the training function that minimizes the cost of that function. Gradient descent starts its working with an initial set of parameter values and iteratively transfers a set of parameter values which minimizes the objective function. Therefore, minimization is achieved by moving toward the opposite direction of the gradient function. To achieve successive results,

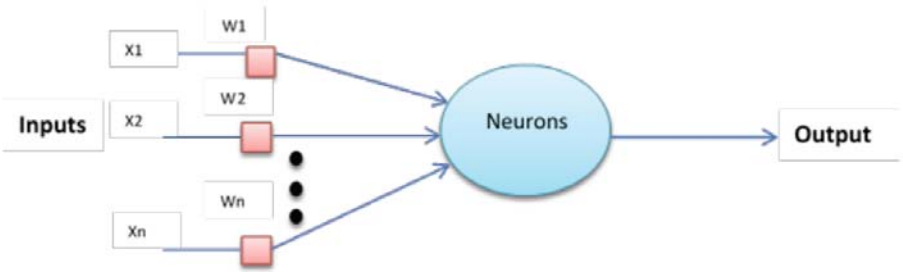

Fig. (5). Process of the neural network working. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

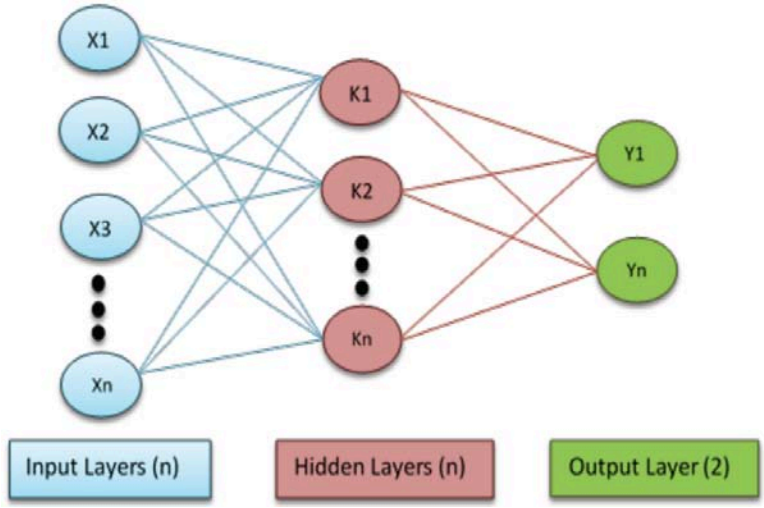

Fig. (6). Architecture of neural network for iMethylK-PseAAC. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

gradient function is computed *via* calculating the rate of change. As gradient descent is a way to minimize an objective function, objective function $J(\lambda)$ parameterized by model's variable $\lambda \in \mathbb{R}^d$ is assessed by updating the parameters in the opposite direction of the gradient of the opposite function $\nabla_\lambda J(\lambda)$ with respect to the parameters. On the basis of the above concept, these parameters are recomputed for each phase through the relation given below.

$$\lambda = \lambda - \gamma \nabla_\lambda J(\lambda) \quad (25)$$

Where Gamma $'\boldsymbol{\gamma}'$ is known as the learning rate**,** generally kept constant, and the performance of the algorithm is significantly determined by the learning rate. It usually concludes how quick the function is reduced. An appropriate learning rate may be hard to choose. It can be based on two conditions; if the learning rate is excessively small then it takes more time to achieve concurrence while if the learning rate is excessively large then the capacity of the function may get affected and never achieve the optimal point. In adaptive learning, the algorithm is permitted to make decisions and acclimate the learning process depending upon the information it already has from the existing data sets.

This algorithm differs from the estimation of the learning rate that is dependent on the execution of the calculation. Learning rate is different from the end goal as the size is limited in each rotation. λ_i and λ_{i+1} are considered as the two progressively assumed parameters. The assigned weights are re-calculated by applying the above two parameters and the concerning results, thus the errors are similarly calculated.

Subsequently, if the errors are more significant when linked with earlier epoch, at that point the learning rate is expanded, weights are disposed off and more recent estimation of λ_{i+1} is enumerated. Furthermore, with the enumeration being smaller, the learning rate is expanded. Hypothetically, the learning rate can change on every epoch, appropriately if $\lambda(\lambda_0, \lambda_1, \lambda_{2,\ldots\ldots})$ are the main parameters calculated for every epoch, and afterwards, they are calculated *via* the following condition.

$$\lambda_{m+1} = \lambda_m - \gamma_m \nabla J(\lambda_m) \quad (26)$$

Where, $'\gamma_m'$ Gamma is the learning rate which is utilized for m^{th} epoch. The algorithm confirms that the learning rate is difficult to achieve as the gradient function is minimized at every epoch. The following condition consistently fulfils at the time of the determination of the learning rate.

$$J(\lambda_0) \geq J(\lambda_1) \geq J(\lambda_2), \ldots \ldots \quad (27)$$

3. RESULTS AND DISCUSSION

The proposed model is employed for the prediction of lysine-methylated sites in protein molecules depending upon the specific amino acid representation of proteins. It plays a vital role in the reformation of protein molecules or protein folding. The prediction is based on the position variant feature extraction techniques. The overall process of validation tests and the results acquired from that validation are described in detail in the current section.

3.1. Estimated Accuracy

The objective evaluation of a newly developed predictor is a very important aspect, which helps to assess the success rate of that model [80]. However, for such objective evaluation, one needs to consider two important factors which are: (i) selection of accuracy metrics and (ii) the testing method employed to validate the model. Herein, firstly we formulate the metrics for objective evaluation and then employ various validation methods.

3.2. Formulation of Metrics

For objective evaluation, one needs to consider the metrics and method of evaluation. The most observed practice for the objective evaluation of the predictor is the use of accuracy metrics which are (1) Accuracy (Acc), which is used for the estimation of the overall accuracy of that prediction model, (2) Sensitivity (Sn), which is used for the estimation of positive sample prediction capability, (3) Specificity (Sp), which is used for the estimation of negative sample prediction capability, and (4) Mathews Correlation Coefficient (MCC), which is used for the estimation of prediction model stability. Initially, these measures have been introduced in a study [112], and a set of four intuitive equations has been derived in various studies [113, 114] for all these measures, which are:

$$\begin{cases} Sn = 1 - \dfrac{\mathcal{N}_{-}^{+}}{\mathcal{N}^{+}} & 0 \le \mathrm{Sn} \le 1 \\ Sp = 1 - \dfrac{\mathcal{N}_{+}^{-}}{\mathcal{N}^{-}} & 0 \le \mathrm{Sp} \le 1 \\ Acc = 1 - \dfrac{\mathcal{N}_{-}^{+} + \mathcal{N}_{+}^{-}}{\mathcal{N}^{+} + \mathcal{N}^{-}} & 0 \le \mathrm{Acc} \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{\mathcal{N}_{-}^{+}}{\mathcal{N}^{+}} + \dfrac{\mathcal{N}_{+}^{-}}{\mathcal{N}^{-}}\right)}{\sqrt{\left(1 + \dfrac{\mathcal{N}_{+}^{-} - \mathcal{N}_{-}^{+}}{\mathcal{N}^{+}}\right)\left(1 + \dfrac{\mathcal{N}_{-}^{+} - \mathcal{N}_{+}^{-}}{\mathcal{N}^{-}}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (28)$$

Where $\mathcal{N}^{-}$ represents the total number of non-methyllysine sites, correctly predicted as non-methyllysine sites by iMethylK-PseAAC. $\mathcal{N}_{+}^{-}$ represents the total number of non-methyllysine sites which are predicted incorrectly as methyllysine sites by iMethylK-PseAAC. Moreover, $\mathcal{N}^{+}$ is the total number of methyllysine sites which are correctly predicted as methyllysine sites by iMethylK-PseAAC and $\mathcal{N}_{-}^{+}$ is the total number of methyllysine sites which are predicted incorrectly as the non-protease by iMethylK-PseAAC. Thus, Eq. (28) explains the specificity, sensitivity, overall-accuracy, and stability in a more easy to understand and intuitive manner, particularly when we talk about MCC [115-117].

This set of perceptive metrics has been used by a number of modern publications [33-35, 38-41, 43, 91, 99, 113, 118-138] but only for binary labelled data. Multi-label prediction is a completely different problem, which has been more popular in computational biology [139-141] and biomedicine [142]. Thus, it requires a different kind of metrics [143]. For the multi-label systems (where a sample may simultaneously belong to several classes), the existence of which has become more frequent in system biology [84, 144-150], system medicine [151, 152] and biomedicine [47], a completely different set of metrics as defined in a study [153] is absolutely needed.

3.3. Self-consistency Testing

Self-consistency test is basically used to obtain the confusion matrix. It is one of the important tests used to substantiate the efficiency of the predictive model in which the training datasets were used for testing the model. The reason for conducting the self-consistency test is that we already know the actual true positive of benchmark dataset. The results of self-consistency are shown in Table **1**; it can be observed that iMethylK-PseAAC has 98.32% Acc, 98.76% Sp, 97.51% Sp, and 0.98 MCC.

The Receiver Operating Characteristics (ROC) is an additional essential tool used to explain the distribution of the experimental results [154]. It allows creating a curve and a complete sensitivity and specificity report. The curve is made by plotting the true positive rate (TPR) against the false positive rate (FPR) for a specific decision threshold. The TPR (sensitivity) depicts how many correct positive outcomes arise between all the positive sequences while FPR (specificity) expresses how many incorrect positive outcomes arise between all of the negative sequences accessible during the prediction. Moreover, Area under the Curve (AUC) is an extent of how well a parameter can differentiate between TPR and FPR. It represents the overall accuracy of the proposed system. In short, each prediction result or instance of a confusion matrix signifies one point in the ROC space. A graphical representation of ROC for the proposed model is shown in Fig. **7**.

3.4. 10-fold Cross-validation

Cross-validation is a procedure to estimate the predictive model by dividing the original sample into a training set to train the model, and a test set to evaluate it. It is useful in the situation when the accurate estimation of the predictive model is required. In a prediction problem, a model is typically a dataset of known data in which training is performed (training dataset), and a dataset of unknown data against which the model is tested (testing dataset).

Cross-validation is a technique to develop a possibility that the suggested method is smooth while an observable validation test set is not accessible. In k- fold cross-validation, the original sample is randomly split into k equal size subsamples. A single subsample obtained from k subsamples is used as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The procedure is repeated k times where$k = 10$, with each of the k subsamples used exactly once as the validation data as shown in Fig. **8**. The mean for all the values of k is used to produce a single estimation which is the result of the cross-validation.

Therefore, let $\mathcal{Z}$ be considered as the population of samples that contains equall positive and negative samples denoted as;

$$\mathcal{Z} = \{z_1, z_2, z_3 \ldots, z_n\}$$

Where $\mathcal{Z}_i$ is any random positive and negative sequence. Data set is split into k equivalent size subgroups $\mathcal{Z}_i$ essentially given as:

$$\bigcup_{i=1}^{k} \mathcal{Z}_i = \mathcal{Z}$$

Table 1. Results for self-consistency testing for iMethylK-PseAAC.

Predictor	Accuracy Metrics			
	***Acc* (%)**	***Sp* (%)**	***Sn* (%)**	***MCC***
iMethylK-PseAAC	96.7	97.7	95.8	0.934

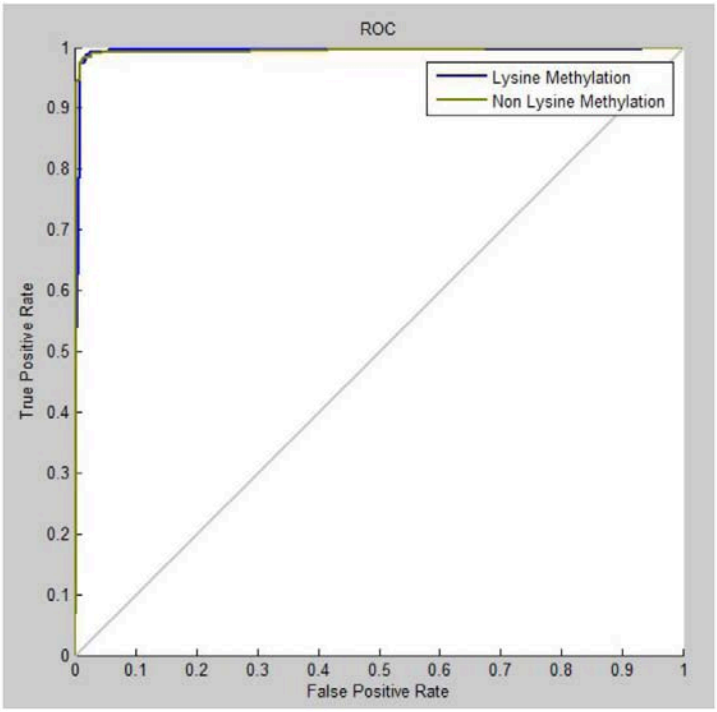

Fig. (7). An ROC graph for the proposed model. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Fig. (8). Graphical illustration of 10-fold cross-validation. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

And

$$\bigcup_{i=1}^{k} \mathcal{Z} = \phi$$

Similarly, the subsets are nominated arbitrarily such that the previous dimensions are parallel. *i.e.*

$$|\mathcal{Z}_i| \simeq |\mathcal{Z}_j|$$

Where $\mathcal{Z}_i$ and $\mathcal{Z}_j$ are some random sets, therefore in a distinct iteration the components of the set $\mathcal{Z}_i$ are excluded and the model is trained on the remaining dataset. After training, the trained model is used to test the excluded data and then the accuracy rate $\mathcal{R}_i$ is calculated. The whole cross-validation result $\mathcal{R}_a$ is calculated by just taking the average of the results of all the *k* iterations.

$$\mathcal{R}_a = \frac{\sum_{i=1}^{\kappa} \mathcal{R}_a}{\kappa}$$

The effects of cross-validation prove that the proposed model sufficiently performed well than the other predictors. The average of the cross-validation testing results is depicted in Table **2**.

3.5. Jackknife Testing

The effectiveness of cross-validation can be remarkable in the case of diversified and biased data. Jackknife testing is also executed by various researchers for validation [17, 19-23, 29-31, 33, 36, 42, 51, 59, 68, 82-88, 92, 96, 135, 148, 155-164]. Out of the frequently used methods, jackknife is the most common. There are few other techniques to test the predictor through randomly selected or portioned dataset. There are a lot of ways for partition, therefore, results can either be good or else through each partition. Due to the use of very small selection in this subsampling technique, it is inevitable for different selections to yield different results.

Table 2. Results for 10-fold cross-validation of iMethylK-PseAAC.

Folds	Sn (%)	Sp (%)	MCC	Acc (%)
K1	79.5	94.6	0.71	92.1
K2	81.3	92.7	0.67	91.2
K3	68.4	92.5	0.59	88.7
K4	83.3	94.6	0.74	92.9
K5	86.5	95.5	0.79	94.1
K6	85.7	91.9	0.66	91.2
K7	82.4	93.6	0.70	92.0
K8	73.8	94.9	0.69	91.2
K9	84.8	97.3	0.72	92.4
K10	73.7	93.5	0.65	90.3
Average	**79.88**	**94.11**	**0.692**	**91.61**

The jackknife can be utilized to evaluate the real predictive power of such models by anticipating the dependent variable estimations of every insight considering the insight as another observation. Keeping in mind the end goal, the predicted values of every new insight are acquired from the model based on the sample of insights minus the insights to be predicted. The jackknife, in this perspective, is a technique which is utilized to get a fair prediction, to assess the ability to produce unique results, (*i.e.*, an arbitrary effect) and to limit the hazard of over-fitting.

The main objective of jackknife is to evaluate a parameter of a population of interest from a random sample of data from this population. The parameter is indicated as θ, its measurement from a sample is denoted by T, and its jackknife measure is denoted by Ai for the ith iteration. The sample of N observations is a set represented as:

$$\mathrm{T} = \{\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_{3,} \dots, \mathrm{X}_n\}$$

The sample measure of the parameter is a factor of the observation in the sample. The dataset used to compute $\mathrm{A}_i$leaves out the ith element in the population using the dataset T_i, given as:

$$\mathrm{T}_i = \{\mathrm{X}_1, \mathrm{X}_{2,}\mathrm{X}_3, \dots, \mathrm{X}_{i-1,}\mathrm{X}_{i+1}, \dots, \mathrm{X}_n\}$$

The trained NN is simulated through the feature vector of all the samples in X_i. The accuracy of this permutation A_i is calculated by using the number of false positives and false negatives as well as the number of true positives and true negatives. The average of all the values of A_i is calculated A_*.

$$\mathrm{A}_* = \frac{1}{n}\sum_{i=1}^{n} \mathrm{A}_*$$

Where A_* signifies the whole accuracy of the predictors and n signifies the total amount of observations. The results for jackknife testing are shown in Table **3**.

3.6. Comparison with Existing Methodologies

In this section, we compare the proposed model with some existing methodologies as shown in Table **4**. These existing predictors have used the same classifier named SVM. The datasets used by previous researchers were imbalanced, outdated and the feature construction techniques did not extract important information which resulted in a low accuracy rate. In the proposed prediction system, the Neural Network is used to get high accuracy as some existing methods. Table **4** shows the existing methods *i.e.* iMethyl-PseACC [48], BPB-PPMS [71] and PMeS [73]. iMethyl-PseACC achieved an accuracy of 75% through independent testing and 70.74% using jackknife testing and BPB-PPMS achieved an accuracy of 91.19% through independent testing and 75.51% with 5-fold cross-validation. Similarly, PMeS achieved an accuracy rate of 89.16% by using 10-fold cross-validation and 85.87% through independent testing but the proposed model achieved an accuracy of 96.7% through self-consistency testing, 91.61% through 10-fold cross-validation and 93.42% with jackknife testing which means it is highly accurate by using the backpropagation methodology.

Similarly, comparative results of ROC for the proposed model and the existing models are shown in Fig. **9**. It is clearly shown in the figure that the Area under cure line in blue colour has almost 0.967 true positive rate which means that the accuracy of the proposed model is 96.7% and it is observed that the proposed predictor is highly accurate than the existing ones.

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for analyzing complicated relations therein, as indicated by many previous studies on a series of important biological topics [165-178], particularly in enzyme kinetics, protein folding rates [172, 179-181], and low-frequency internal motion [179-184].

Table 3. Jackknife validation results for iMethylK-PseAAC (Average of n-iterations).

Predictor	Accuracy Metrics			
	***Sn* (%)**	***Sp* (%)**	***Acc* (%)**	***MCC***
iMethylK-PseAAC	92.34	97.80	93.42	0.82

Table 4. Comparison between existing methodologies.

Predictors	Classifiers	Validation Methods	Sn%	Sp%	Acc%	MCC
iMethylK-PseAAC	Neural Network	Self-consistency testing	95.8	97.7	96.7	0.9345
		10-fold cross validation	79.88	94.11	91.61	0.692
		Jackknife testing	92.34	97.80	93.42	0.82
iMethyl-PseACC [48]	SVM	Independent testing	100.00	61.54	75.00	0.60
		Jackknife testing	68.58	72.99	70.74	0.42
BPB-PPMS [71]	SVM	Independent testing	71.43	91.51	91.19	-
		5-fold cross validation	70.05	77.08	75.51	0.3400
PMeS [73]	SVM	10-fold cross validation	84.38	93.94	89.16	0.786
		Independent testing	76.09	95.65	85.87	0.7315

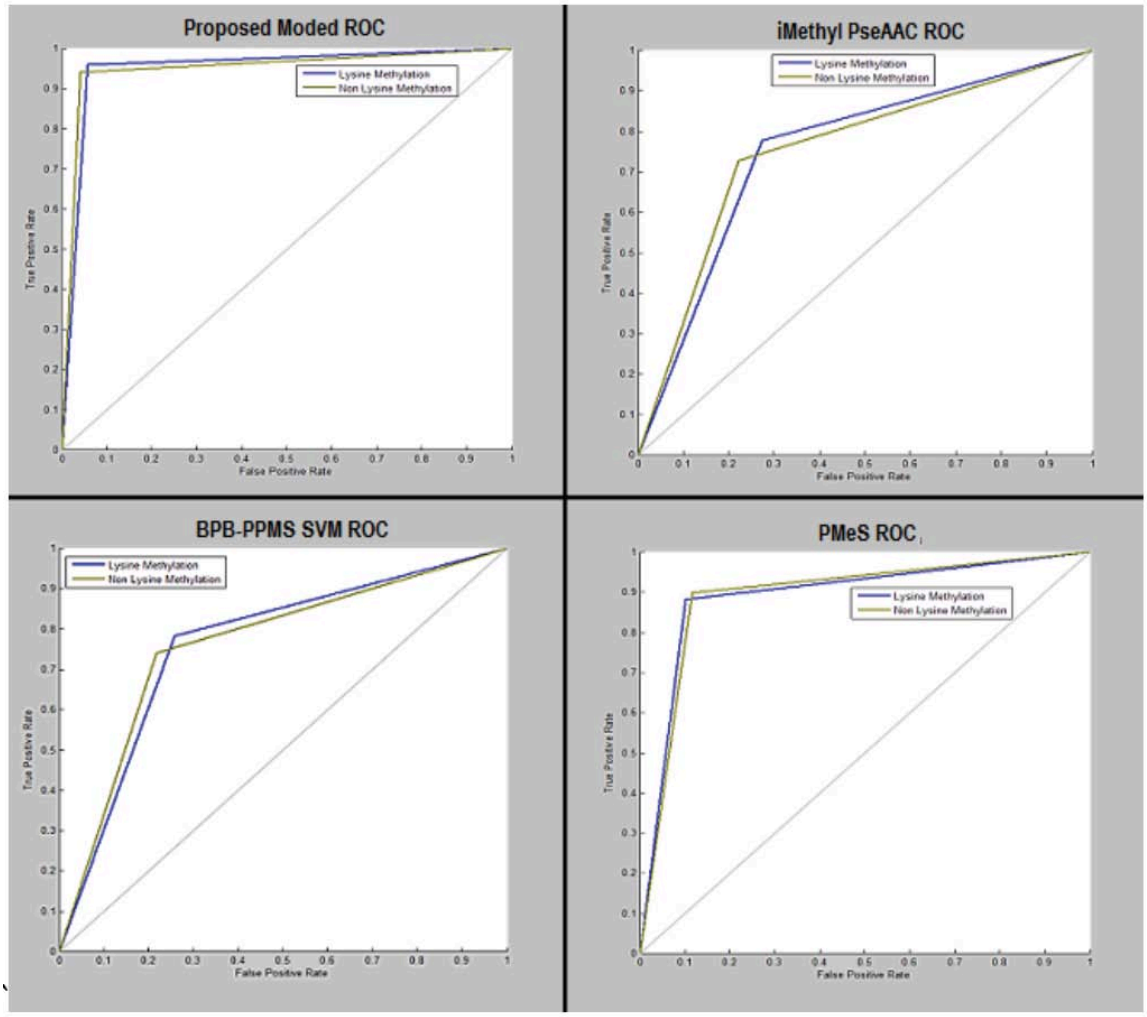

Fig. (9). A comparison of a Receiver Operating Characteristics (ROC) graph. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The importance of methylation has been highlighted in several existing research. Analysts have suggested different computational methodologies to recognize lysine methylation sites. Researchers have put their greatest determination towards the improvement of prediction accuracy and to distinguish lysine methylation sites. In this research, we focus on attaining greatest accuracy by overwhelming disadvantages within the existing ones. A few key highlights make the studied methodology more predictable as well as precise from the existing methodologies. The collected benchmark dataset is latest and balanced as experimental observations have been incorporated. Also, it is non-repetitive and has a defined length. Moreover, it is different in nature as it creates essential sequences from various entities. In particular, the feature extraction method is position and scale variant which is able to carefully extract profound ambiguous patterns. Furthermore, to gauge the performance of the prognostic model, complete 10-fold cross-validation and jackknife testing are implemented [36, 185]. Some existing approaches depicted previously have distinctive requisites in their methodologies. The datasets used by the previous researchers were imbalanced and outdated and the feature construction techniques do not mine the important information [48, 71, 73]. In this research, non-repetitive and up-to-date datasets of massive size have been employed and comprehensive features have been mined. After the mining of most relevant features, early experiments were directed with lesser feature vectors. These features were then extended by constant testing and experimentation to the point that most exact outcomes were accomplished. The feature which has been used for the proposed model helps in revealing profound ambiguous patterns, about position and composition having the most extreme significance. Moreover, different performance metrics have been processed and compared with some previous models. Finally, 10-fold cross-validation, as well as jackknife testing, have been used in order to verify and validate the accuracy of the proposed model.

3.7. Webserver

The final step of Chou's 5-steps rule is the development of user-friendly publicly available web-server for the ease of users and biologists as explained in recent publications by various authors [45, 116, 123, 126, 145, 146, 149, 151]. As pointed out in a study [183] and demonstrated in a series of recent publications [23, 44, 45, 65, 82-85, 89, 92, 93, 97, 98, 103, 133, 138, 144-150, 186], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have significantly increased the impacts of bioinformatics on medical science [63], driving medicinal chemistry into an unprecedented revolution [187]. Accordingly, in our future work, we shall strive to establish a web-server for the new method presented in this paper. However, the source code for iMethylK-PseAAC is available at: https://github.com/umtwaqar/iMethylK-PseAAC.

CONCLUSION

Methylation is one of the most significant Post-Translational Modification occuring on lysine residues which can cause dangerous reactions in the body. To appropriately obtain the information of lysine methylation sites is significant for studying and examining various human diseases. In this research, feature vectors are formed by using several feature extraction techniques including scale and position variant features and raw, central and Hahn moments. The proposed model could effectively identify the lysine methylation sites by using backpropagation methodology. The neural network is an effectual method for supervised and unsupervised learning problems. The prediction algorithm developed for lysine methylation also employs supervised learning. The results obtained from the trained neural network are authenticated by 10 fold cross-validation and jackknife testing. It proves that the model beats the existing methods such as iMethyl-PseACC, BPB-PPMS and PMeS. Also, the accuracy of the proposed model is proved with the help of accuracy metrics including sensitivity, specificity, Accuracy and Mathew's correlation coefficient which demonstrates that the accuracy of the model provides an effective and exact rate and time in comparison with the previous ones like iMethyl-PseACC, BPB-PPMS, PMeS.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this study are available within its supplementary materials S1.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Paik, W.K.; Kim, S. Enzymatic methylation of protein fractions from calf thymus nuclei. *Biochem. Biophys. Res. Commun.,* **1967**, *29*(1), 14-20. [http://dx.doi.org/10.1016/0006-291X(67)90533-5] [PMID: 6055181]

[2] Xu, Y.; Chou, K-C. Recent progress in predicting posttranslational

modification sites in proteins. *Curr. Top. Med. Chem.,* **2016**, *16*(6), 591-603.
[http://dx.doi.org/10.2174/1568026615666150819110421] [PMID: 26286211]

[3] Clarke, S. Protein methylation. *Curr. Opin. Cell Biol.,* **1993**, *5*(6), 977-983.
[http://dx.doi.org/10.1016/0955-0674(93)90080-A] [PMID: 8129951]

[4] Schubert, H.L.; Blumenthal, R.M.; Cheng, X. 1 Protein methyltransferases: Their distribution among the five structural classes of AdoMet-dependent methyltransferases. *The Enzymes*: Elsevier, **2006**, Vol. 24, pp. 3-28.

[5] Grewal, S.I.; Rice, J.C. Regulation of heterochromatin by histone methylation and small RNAs. *Curr. Opin. Cell Biol.,* **2004**, *16*(3), 230-238.
[http://dx.doi.org/10.1016/j.ceb.2004.04.002] [PMID: 15145346]

[6] Lee, D.Y.; Teyssier, C.; Strahl, B.D.; Stallcup, M.R. Role of protein methylation in regulation of transcription. *Endocr. Rev.,* **2005**, *26*(2), 147-170.
[http://dx.doi.org/10.1210/er.2004-0008] [PMID: 15479858]

[7] Chen, X.; Niroomand, F.; Liu, Z.; Zankl, A.; Katus, H.A.; Jahn, L.; Tiefenbacher, C.P. Expression of nitric oxide related enzymes in coronary heart disease. *Basic Res. Cardiol.,* **2006**, *101*(4), 346-353.
[http://dx.doi.org/10.1007/s00395-006-0592-5] [PMID: 16705470]

[8] Mastronardi, F.G.; Wood, D.D.; Mei, J.; Raijmakers, R.; Tseveleki, V.; Dosch, H-M.; Probert, L.; Casaccia-Bonnefil, P.; Moscarello, M.A. 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. *J. Neurosci.,* **2006**, *26*(44), 11387-11396.
[http://dx.doi.org/10.1523/JNEUROSCI.3349-06.2006] [PMID: 17079667]

[9] Shukla, A.; Chaurasia, P.; Bhaumik, S.R. Histone methylation and ubiquitination with their cross-talk and roles in gene expression and stability. *Cell. Mol. Life Sci.,* **2009**, *66*(8), 1419-1433.
[http://dx.doi.org/10.1007/s00018-008-8605-1] [PMID: 19370393]

[10] Varier, R.A.; Timmers, H.M. Histone lysine methylation and demethylation pathways in cancer. *Biochimica et Biophysica Acta (BBA)-. Rev. Can.,* **2011**, *1815*(1), 75-89.

[11] Predel, R.; Brandt, W.; Kellner, R.; Rapus, J.; Nachman, R.J.; Gäde, G. Post-translational modifications of the insect sulfakinins: sulfation, pyroglutamate-formation and O-methylation of glutamic acid. *Eur. J. Biochem.,* **1999**, *263*(2), 552-560.
[http://dx.doi.org/10.1046/j.1432-1327.1999.00532.x] [PMID: 10406966]

[12] Bannister, A.J.; Kouzarides, T. Reversing histone methylation. *Nature,* **2005**, *436*(7054), 1103-1106.
[http://dx.doi.org/10.1038/nature04048] [PMID: 16121170]

[13] Akmal, M.A.; Rasool, N.; Khan, Y.D. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PLoS One,* **2017**, *12*(8), e0181966.
[http://dx.doi.org/10.1371/journal.pone.0181966] [PMID: 28797096]

[14] Butt, A.H.; Khan, S.A.; Jamil, H.; Rasool, N.; Khan, Y.D. A prediction model for membrane proteins using moments based features. **2016**, *2016*, 1-7.
[http://dx.doi.org/10.1155/2016/8370132]

[15] Butt, A.H.; Rasool, N.; Khan, Y.D. A treatise to computational approaches towards prediction of membrane protein and its subtypes. *J. Membr. Biol.,* **2017**, *250*(1), 55-76.
[http://dx.doi.org/10.1007/s00232-016-9937-7] [PMID: 27866233]

[16] Butt, A.H.; Rasool, N.; Khan, Y.D. Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC. *Mol. Biol. Rep.,* **2018**, *45*(6), 2295-2306.
[http://dx.doi.org/10.1007/s11033-018-4391-5] [PMID: 30238411]

[17] Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K-C. SPalmitoylC-PseAAC: A sequence-based model developed *via* Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Anal. Biochem.,* **2019**, *568*, 14-23.
[http://dx.doi.org/10.1016/j.ab.2018.12.019] [PMID: 30593778]

[18] Khan, Y.D.; Ahmed, F.; Khan, S.A. Situation recognition using image moments and recurrent neural networks. *Neural Comput. Appl.,* **2014**, *24*(7-8), 1519-1529.
[http://dx.doi.org/10.1007/s00521-013-1372-4]

[19] Khan, Y.D.; Jamil, M.; Hussain, W.; Rasool, N.; Khan, S.A.; Chou, K-C. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *J. Theor. Biol.,* **2018**, *463*, 47-55.
[PMID: 30550863]

[20] Khan, Y.D.; Rasool, N.; Hussain, W.; Khan, S.A.; Chou, K-C. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Anal. Biochem.,* **2018**, *550*, 109-116.
[http://dx.doi.org/10.1016/j.ab.2018.04.021] [PMID: 29704476]

[21] Khan, Y.D.; Rasool, N.; Hussain, W.; Khan, S.A.; Chou, K-C. iPhosY-PseAAC: identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. *Mol. Biol. Rep.,* **2018**, *45*(6), 2501-2509.
[http://dx.doi.org/10.1007/s11033-018-4417-z] [PMID: 30311130]

[22] Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K-C. SPrenylC-PseAAC: A sequence-based model developed *via* Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *J. Theor. Biol.,* **2019**, *468*, 1-11.
[http://dx.doi.org/10.1016/j.jtbi.2019.02.007] [PMID: 30768975]

[23] Ghauri, A.W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.C. pNitro-Tyr-PseAAC: Predict nitrotyrosine sites in proteins by incorporating five features into Chou's general PseAAC. *Curr. Pharm. Des.,* **2018**, *24*(34), 4034-4043.
[http://dx.doi.org/10.2174/1381612825666181127101039] [PMID: 30479209]

[24] Ju, Z.; Cao, J-Z.; Gu, H. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou′s general PseAAC. *J. Theor. Biol.,* **2016**, *397*, 145-150.
[http://dx.doi.org/10.1016/j.jtbi.2016.02.020] [PMID: 26908349]

[25] Ju, Z.; He, J-J. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. *J. Mol. Graph. Model.,* **2017**, *77*, 200-204.
[http://dx.doi.org/10.1016/j.jmgm.2017.08.020] [PMID: 28886434]

[26] Liu, L-M.; Xu, Y.; Chou, K-C. iPGK-PseAAC: Identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem.,* **2017**, *13*(6), 552-559.
[http://dx.doi.org/10.2174/1573406413666170515120507] [PMID: 28521678]

[27] Qiu, W-R.; Jiang, S-Y.; Sun, B-Q.; Xiao, X.; Cheng, X.; Chou, K-C. iRNA-2methyl: Identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. *Med. Chem.,* **2017**, *13*(8), 734-743.
[http://dx.doi.org/10.2174/1573406413666170623082245] [PMID: 28641529]

[28] Chandra, A.; Sharma, A.; Dehzangi, A.; Ranganathan, S.; Jokhan, A.; Chou, K-C.; Tsunoda, T. PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. *Sci. Rep.,* **2018**, *8*(1), 17923.
[http://dx.doi.org/10.1038/s41598-018-36203-8] [PMID: 30560923]

[29] Wang, L.; Zhang, R.; Mu, Y. Fu-SulfPred: Identification of protein s-sulfenylation sites by fusing forests *via* Chou's general PseAAC. *J. Theor. Biol.,* **2019**, *461*, 51-58.
[http://dx.doi.org/10.1016/j.jtbi.2018.10.046] [PMID: 30365947]

[30] Akbar, S.; Hayat, M. iMethyl-STTNC: Identification of N^6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. *J. Theor. Biol.,* **2018**, *455*, 205-211.
[http://dx.doi.org/10.1016/j.jtbi.2018.07.018] [PMID: 30031793]

[31] Chen, W.; Ding, H.; Zhou, X.; Lin, H.; Chou, K-C. iRNA(m6A)-PseDNC: Identifying N^6-methyladenosine sites using pseudo dinucleotide composition. *Anal. Biochem.,* **2018**, *561-562*, 59-65.
[http://dx.doi.org/10.1016/j.ab.2018.09.002] [PMID: 30201554]

[32] Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K-C. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.,* **2015**, *490*, 26-33.
[http://dx.doi.org/10.1016/j.ab.2015.08.021] [PMID: 26314792]

[33] Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K-C. iRNA-3typeA: Identifying three types of modification at RNA's adenosine sites. *Mol. Ther. Nucleic Acids,* **2018**, *11*, 468-474.
[http://dx.doi.org/10.1016/j.omtn.2018.03.012] [PMID: 29858081]

[34] Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K-C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids,* **2016**, *5,* e332.
[PMID: 28427142]

[35] Feng, P.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K-C. iR-

NA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids,* **2017**, *7*, 155-163. [http://dx.doi.org/10.1016/j.omtn.2017.03.006] [PMID: 28624191]

[36] Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K-C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics,* **2018**, *111*(1), 96-102. [PMID: 29360500]

[37] Jia, C.; Lin, X.; Wang, Z. Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. *Int. J. Mol. Sci.,* **2014**, *15*(6), 10410-10423. [http://dx.doi.org/10.3390/ijms150610410] [PMID: 24918295]

[38] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K-C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.,* **2016**, *497*, 48-56. [http://dx.doi.org/10.1016/j.ab.2015.12.009] [PMID: 26723495]

[39] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K-C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.,* **2016**, *394*, 223-230. [http://dx.doi.org/10.1016/j.jtbi.2016.01.020] [PMID: 26807806]

[40] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K-C. iCar-PseCp: Identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget,* **2016**, *7*(23), 34558-34570. [http://dx.doi.org/10.18632/oncotarget.9148] [PMID: 27153555]

[41] Jia, J.; Zhang, L.; Liu, Z.; Xiao, X.; Chou, K-C. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics,* **2016**, *32*(20), 3133-3141. [http://dx.doi.org/10.1093/bioinformatics/btw387] [PMID: 27354696]

[42] Ju, Z.; Wang, S-Y. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. *Gene,* **2018**, *664*, 78-83. [http://dx.doi.org/10.1016/j.gene.2018.04.055] [PMID: 29694908]

[43] Liu, Z.; Xiao, X.; Yu, D-J.; Jia, J.; Qiu, W-R.; Chou, K-C. pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences *via* physical-chemical properties. *Anal. Biochem.,* **2016**, *497*, 60-67. [http://dx.doi.org/10.1016/j.ab.2015.12.017] [PMID: 26748145]

[44] Qiu, W.R.; Sun, B.Q.; Xiao, X.; Xu, D.; Chou, K.C. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC *via* grey system theory. *Mol. Inform.,* **2017**, *36*(5-6). [http://dx.doi.org/10.1002/minf.201600010] [PMID: 28488814]

[45] Qiu, W-R.; Jiang, S-Y.; Xu, Z-C.; Xiao, X.; Chou, K-C. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. *Oncotarget,* **2017**, *8*(25), 41178-41188. [http://dx.doi.org/10.18632/oncotarget.17104] [PMID: 28476023]

[46] Qiu, W-R.; Sun, B-Q.; Xiao, X.; Xu, Z-C.; Chou, K-C. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget,* **2016**, *7*(28), 44310-44321. [http://dx.doi.org/10.18632/oncotarget.10027] [PMID: 27322424]

[47] Qiu, W-R.; Sun, B-Q.; Xiao, X.; Xu, Z-C.; Chou, K-C. iPTM-mLys: Identifying multiple lysine PTM sites and their different types. *Bioinformatics,* **2016**, *32*(20), 3116-3123. [http://dx.doi.org/10.1093/bioinformatics/btw380] [PMID: 27334473]

[48] Qiu, W.-R.; Xiao, X.; Lin, W.-Z.; Chou, K.-C. iMethyl-PseAAC: identification of protein methylation sites *via* a pseudo amino acid composition approach. *BioMed Res. Int.,* **2014**, *2014*, 1-12.

[49] Qiu, W-R.; Xiao, X.; Lin, W-Z.; Chou, K-C. iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information *via* a gray system model. *J. Biomol. Struct. Dyn.,* **2015**, *33*(8), 1731-1742. [http://dx.doi.org/10.1080/07391102.2014.968875] [PMID: 25248923]

[50] Qiu, W-R.; Xiao, X.; Xu, Z-C.; Chou, K-C. iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget,* **2016**, *7*(32), 51270-51283. [http://dx.doi.org/10.18632/oncotarget.9987] [PMID: 27323404]

[51] Sabooh, M.F.; Iqbal, N.; Khan, M.; Khan, M.; Maqbool, H.F. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. *J. Theor. Biol.,* **2018**, *452*, 1-9. [http://dx.doi.org/10.1016/j.jtbi.2018.04.037] [PMID: 29727634]

[52] Xie, H-L.; Fu, L.; Nie, X-D. Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. *Protein Eng. Des. Sel.,* **2013**, *26*(11), 735-742. [http://dx.doi.org/10.1093/protein/gzt042] [PMID: 24048266]

[53] Xu, Y.; Ding, J.; Wu, L-Y.; Chou, K-C. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One,* **2013**, *8*(2), e55844. [http://dx.doi.org/10.1371/journal.pone.0055844] [PMID: 23409062]

[54] Xu, Y.; Shao, X-J.; Wu, L-Y.; Deng, N-Y.; Chou, K-C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *Peer J.,* **2013**, *1*, e171. [http://dx.doi.org/10.7717/peerj.171] [PMID: 24109555]

[55] Xu, Y.; Wang, Z.; Li, C.; Chou, K-C. iPreny-PseAAC: Identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med. Chem.,* **2017**, *13*(6), 544-551. [http://dx.doi.org/10.2174/1573406413666170419150052] [PMID: 28425870]

[56] Xu, Y.; Wen, X.; Shao, X-J.; Deng, N-Y.; Chou, K-C. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci.,* **2014**, *15*(5), 7594-7610. [http://dx.doi.org/10.3390/ijms15057594] [PMID: 24857907]

[57] Xu, Y.; Wen, X.; Wen, L-S.; Wu, L-Y.; Deng, N-Y.; Chou, K-C. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One,* **2014**, *9*(8), e105018. [http://dx.doi.org/10.1371/journal.pone.0105018] [PMID: 25121969]

[58] Zhang, J.; Zhao, X.; Sun, P.; Ma, Z. PSNO: Predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. *Int. J. Mol. Sci.,* **2014**, *15*(7), 11204-11219. [http://dx.doi.org/10.3390/ijms150711204] [PMID: 24968264]

[59] Ehsan, A.; Mahmood, K.; Khan, Y.D.; Khan, S.A.; Chou, K-C. A novel modeling in mathematical biology for classification of signal peptides. *Sci. Rep.,* **2018**, *8*(1), 1039. [http://dx.doi.org/10.1038/s41598-018-19491-y] [PMID: 29348418]

[60] Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K-C. SPalmitoylC-PseAAC: A sequence-based model developed *via* Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Anal. Biochem.,* **2018**, *568*, 14-23. [PMID: 30593778]

[61] Awais, M.; Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. Bioinformatics, iPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. **2019**, 1.

[62] Chen, Z.; Liu, X.; Li, F.; Li, C.; Marquez-Lago, T.; Leier, A.; Akutsu, T.; Webb, G.I.; Xu, D.; Smith, A.I.J.B.B. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform.,* **2018**. [http://dx.doi.org/10.1093/bib/bby089]

[63] Chou, K-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.,* **2015**, *11*(3), 218-234. [http://dx.doi.org/10.2174/1573406411666141229162834] [PMID: 25548930]

[64] Li, F.; Zhang, Y.; Purcell, A.W.; Webb, G.I.; Chou, K.-C.; Lithgow, T.; Li, C.; Song, J. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinform.,* **2019**, *20*(1), 112. [http://dx.doi.org/10.1186/s12859-019-2700-1]

[65] Qiu, W-R.; Sun, B-Q.; Xiao, X.; Xu, Z-C.; Jia, J-H.; Chou, K-C. iKCR-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. *Genomics,* **2017**, *110*(5), 239-246.

[PMID: 29107015]
[66] Wang, L.; Zhang, R.; Mu, Y. J. J. o. t. b. Fu-SulfPred: Identification of protein S-sulfenylation sites by fusing forests *via* Chou's general PseAAC. **2019**, *461*, 51-58.
[67] Xie, H.-L.; Fu, L.; Nie, X.-D. J. P. E. Design; Selection, using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. **2013**, *26*(11), 735-742.
[68] Zhang, Y.; Xie, R.; Wang, J.; Leier, A.; Marquez-Lago, T.T.; Akutsu, T.; Webb, G.I.; Chou, K-C.; Song, J. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.,* **2018**, *5*.
[http://dx.doi.org/10.1093/bib/bby079] [PMID: 30351377]
[69] Li, B-Q.; Hu, L-L.; Chen, L.; Feng, K-Y.; Cai, Y-D.; Chou, K-C. Prediction of protein domain with mRMR feature selection and analysis. *PLoS One,* **2012**, *7*(6), e39308.
[http://dx.doi.org/10.1371/journal.pone.0039308] [PMID: 22720092]
[70] Sun, Y-M.; Liao, W-L.; Huang, H-D.; Liu, B-J.; Chang, C-W.; Horng, J-T.; Wu, L-C. In: A human DNA methylation site predictor based on SVM. *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering,* **2009**, pp. 22-29.
[http://dx.doi.org/10.1109/BIBE.2009.22]
[71] Shao, J.; Xu, D.; Tsai, S-N.; Wang, Y.; Ngai, S-M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One,* **2009**, *4*(3), e4920.
[http://dx.doi.org/10.1371/journal.pone.0004920] [PMID: 19290060]
[72] Hu, L.L.; Li, Z.; Wang, K.; Niu, S.; Shi, X.H.; Cai, Y.D.; Li, H.P. Prediction and analysis of protein methylarginine and methyllysine based on multisequence features. *Biopolymers,* **2011**, *95*(11), 763-771.
[PMID: 21544797]
[73] Shi, S-P.; Qiu, J-D.; Sun, X-Y.; Suo, S-B.; Huang, S-Y.; Liang, R-P. PMeS: Prediction of methylation sites based on enhanced feature encoding scheme. *PLoS One,* **2012**, *7*(6), e38772.
[http://dx.doi.org/10.1371/journal.pone.0038772] [PMID: 22719939]
[74] Valavanis, I.; Sifakis, E.G.; Georgiadis, P.; Kyrtopoulos, S.; Chatziioannou, A.A. A composite framework for the statistical analysis of epidemiological DNA methylation data with the Infinium Human Methylation 450K BeadChip. *IEEE J. Biomed. Health Inform.,* **2014**, *18*(3), 817-823.
[http://dx.doi.org/10.1109/JBHI.2014.2298351] [PMID: 24808224]
[75] Li, Z.; Chen, L.; Lai, Y.; Dai, Z.; Zou, X. The prediction of methylation states in human DNA sequences based on hexanucleotide composition and feature selection. *Anal. Methods,* **2014**, *6*(6), 1897-1904.
[http://dx.doi.org/10.1039/c3ay41962b]
[76] Karagod, V.V.; Sinha, K. In: A novel machine learning framework for phenotype prediction based on genome-wide DNA methylation data. *2017 International Joint Conference on Neural Networks (IJCNN),* **2017**, pp. 1657-1664.
[http://dx.doi.org/10.1109/IJCNN.2017.7966050]
[77] Khan, Y.D.; Ahmad, F.; Anwar, M.W. A neuro-cognitive approach for iris recognition using back propagation. *World Appl. Sci. J.,* **2012**, *16*(5), 678-685.
[78] Khan, Y.D.; Khan, N.S.; Farooq, S.; Abid, A.; Khan, S.A.; Ahmad, F.; Mahmood, M.K. An efficient algorithm for recognition of human actions. *Sci. World J.,* **2014**, *2014*, 875879.
[http://dx.doi.org/10.1155/2014/875879]
[79] Khan, Y. D.; Khan, S. A.; Ahmad, F.; Islam, S. Iris recognition using image moments and k-means algorithm. *Sci. World J.,* **2014**, *2014*, 1-9.
[http://dx.doi.org/10.1155/2014/723595]
[80] Chou, K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247.
[http://dx.doi.org/10.1016/j.jtbi.2010.12.024] [PMID: 21168420]
[81] Chou, K-C. Using subsite coupling to predict signal peptides. *Protein Eng.,* **2001**, *14*(2), 75-79.
[http://dx.doi.org/10.1093/protein/14.2.75] [PMID: 11297664]
[82] Cheng, X.; Lin, W-Z.; Xiao, X.; Chou, K-C.; Hancock, J. pLoc_bal-mAnimal: Predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics,* **2018**, *1*, 9.
[PMID: 30010789]
[83] Cheng, X.; Xiao, X.; Chou, K-C. pLoc_bal-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC. *J. Theor. Biol.,* **2018**, *458*, 92-102.
[http://dx.doi.org/10.1016/j.jtbi.2018.09.005] [PMID: 30201434]
[84] Xiao, X.; Cheng, X.; Chen, G.; Mao, Q.; Chou, K-C. pLoc_bal-mGpos: Predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. *Genomics,* **2018**, *114*(4), 886-892.
[PMID: 29842950]
[85] Chou, K-C.; Cheng, X.; Xiao, X. pLoc_bal-mHum: Predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. *Genomics,* **2018**, S0888-7543(18)30276-3.
[http://dx.doi.org/10.1016/j.ygeno.2018.08.007] [PMID: 30179658]
[86] Sankari, E.S.; Manimegalai, D. Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC. *J. Theor. Biol.,* **2018**, *455*, 319-328.
[http://dx.doi.org/10.1016/j.jtbi.2018.07.032] [PMID: 30056084]
[87] Contreras-Torres, E. Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC. *J. Theor. Biol.,* **2018**, *454*, 139-145.
[http://dx.doi.org/10.1016/j.jtbi.2018.05.033] [PMID: 29870696]
[88] Javed, F.; Hayat, M. Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC. *Genomics,* **2018**, S0888-7543(18)30519-6.
[http://dx.doi.org/10.1016/j.ygeno.2018.09.004] [PMID: 30196077]
[89] Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K-C. iRNA-AI: Identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget,* **2017**, *8*(3), 4208-4217.
[http://dx.doi.org/10.18632/oncotarget.13758] [PMID: 27926534]
[90] Chen, W.; Feng, P.-M.; Deng, E.-Z.; Lin, H.; Chou, K.-C. iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem*., **2014**, *462*, 76-83.
[91] Chen, W.; Feng, P.-M.; Lin, H.; Chou, K.-C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. **2013**, *41*(6), e68.
[92] Cheng, X.; Xiao, X.; Chou, K-C. pLoc_bal-mPlant: Predict subcellular localization of plant proteins by general PseAAC and balancing training dataset. *Curr. Pharm. Des.,* **2018**, *24*(34), 4013-4022.
[http://dx.doi.org/10.2174/1381612824666181119145030] [PMID: 30451108]
[93] Chou, K.; Cheng, X.; Xiao, X. pLoc_bal-mEuk: Predict subcellular localization of eukaryotic proteins by general PseAAC and quasi-balancing training dataset. *Med. Chem.,* **2018**, *15*(5):472-485.
[94] Ding, H.; Deng, E.-Z.; Yuan, L.-F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.-C. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *Biomed. Res. Int*., **2014**, *2014*, 286419.
[95] Feng, P-M.; Chen, W.; Lin, H.; Chou, K-C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.,* **2013**, *442*(1), 118-125.
[http://dx.doi.org/10.1016/j.ab.2013.05.024] [PMID: 23756733]
[96] Jia, J.; Li, X.; Qiu, W.; Xiao, X.; Chou, K-C. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. *J. Theor. Biol.,* **2019**, *460*, 195-203.
[http://dx.doi.org/10.1016/j.jtbi.2018.10.021] [PMID: 30312687]
[97] Khan, Y.D.; Batool, A.; Rasool, N.; Khan, S.A.; Chou, K-C. Prediction of nitrosocysteine sites using position and composition variant features. **2019**, *16*(4), 283-293.
[98] Li, J.-X.; Wang, S.-Q.; Du, Q.-S.; Wei, H.; Li, X.-M.; Meng, J.-Z.; Wang, Q.-Y.; Xie, N.-Z.; Huang, R.-B.; Chou, K.-C. Simulated protein thermal detection (SPTD) for enzyme thermostability study and an application example for pullulanase from *Bacillus deramificans*. **2018**, *24*(34), 4023-4033.
[99] Lin, H.; Deng, E-Z.; Ding, H.; Chen, W.; Chou, K-C. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.,* **2014**, *42*(21), 12961-12972.
[http://dx.doi.org/10.1093/nar/gku1019] [PMID: 25361964]
[100] Liu, B.; Fang, L.; Long, R.; Lan, X.; Chou, K.-C. J. B. iEnhancer-2L: A two-layer predictor for identifying enhancers and their

strength by pseudo k-tuple nucleotide composition. **2015**, *32*(3), 362-369.

[101] Liu, B.; Fang, L.; Wang, S.; Wang, X.; Li, H.; Chou, K.-C. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J. Theor. Biol*., **2015**, *385*, 153-9. [http://dx.doi.org/10.1016/j.jtbi.2015.08.025]

[102] Liu, Z.; Xiao, X.; Qiu, W.-R.; Chou, K.-C. iDNA-Methyl: Identifying DNA methylation sites *via* pseudo trinucleotide composition. *Anal. Biochem.* **2015**, *474*, 69-77.

[103] Lu, Y.; Wang, S.; Wang, J.; Zhou, G.; Zhang, Q.; Zhou, X.; Niu, B.; Chen, Q.; Chou, K-C. An epidemic avian influenza prediction model based on google trends. **2019**, *16*(4), 303-310.

[104] Xiao, X.; Min, J.-L.; Lin, W.-Z.; Liu, Z.; Cheng, X.; Chou, K.-C. Dynamics, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking *via* benchmark dataset optimization approach. *J. Biomol. Struct. Dyn.,* **2015**, *33*(10), 2221-33.

[105] Chou, K.J.C. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.,* **2019**. [http://dx.doi.org/10.2174/0929867326666190507082559]

[106] Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.,* **2004**, *45*(Database issue), D158-D169. [http://dx.doi.org/10.1093/nar/gkh131]

[107] Huang, K-Y.; Su, M-G.; Kao, H-J.; Hsieh, Y-C.; Jhong, J-H.; Cheng, K-H.; Huang, H-D.; Lee, T-Y. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res.,* **2016**, *44*(D1), D435-D446. [http://dx.doi.org/10.1093/nar/gkv1240] [PMID: 26578568]

[108] Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics,* **2012**, *28*(23), 3150-3152. [http://dx.doi.org/10.1093/bioinformatics/bts565] [PMID: 23060610]

[109] Crooks, G.E.; Hon, G.; Chandonia, J-M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.,* **2004**, *14*(6), 1188-1190. [http://dx.doi.org/10.1101/gr.849004] [PMID: 15173120]

[110] Chen, W.; Xing, P.; Zou, Q.; Detecting, N. Detecting N^6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep.,* **2017**, *7*, 40242. [http://dx.doi.org/10.1038/srep40242] [PMID: 28079126]

[111] Liu, B.; Wu, H.; Zhang, D.; Wang, X.; Chou, K-C. Pse-Analysis: A python package for DNA/RNA and protein/ peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget,* **2017**, *8*(8), 13338-13343. [http://dx.doi.org/10.18632/oncotarget.14524] [PMID: 28076851]

[112] Chou, K.-C. Prediction of signal peptides using scaled window. *Peptides,* **2001**, *22*(12), 1973-1979.

[113] Feng, P.-M.; Ding, H.; Chen, W.; Lin, H. Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math Methods Med.,* **2013**, *2013*, 530696. [http://dx.doi.org/10.1155/2013/530696]

[114] Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *Peer J.,* **2013**, *1,* e171. [http://dx.doi.org/10.7717/peerj.171] [PMID: 24109555]

[115] Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K-C. Using deformation energy to analyze nucleosome positioning in genomes. *Genomics,* **2016**, *107*(2-3), 69-75. [http://dx.doi.org/10.1016/j.ygeno.2015.12.005] [PMID: 26724497]

[116] Qiu, W.R.; Sun, B.Q.; Xiao, X.; Xu, D.; Chou, K.C. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC *via* grey system theory. *Mol. Inform.,* **2017**, *36*(5-6), 1600010. [http://dx.doi.org/10.1002/minf.201600010] [PMID: 28488814]

[117] Xiao, X.; Ye, H-X.; Liu, Z.; Jia, J-H.; Chou, K-C. iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget,* **2016**, *7*(23), 34180-34189. [http://dx.doi.org/10.18632/oncotarget.9057] [PMID: 27147572]

[118] Lin, H.; Deng, E.Z.; Ding, H.; Chen, W.; Chou, K.C. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.,* **2014**, *42*(21), 12961-12972. [http://dx.doi.org/10.1093/nar/gku1019] [PMID: 25361964]

[119] Xu, Y.; Wen, X.; Wen, L.S.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One,* **2014**, *9*(8), e105018. [http://dx.doi.org/10.1371/journal.pone.0105018] [PMID: 25121969]

[120] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.,* **2016**, *394*, 223-230. [http://dx.doi.org/10.1016/j.jtbi.2016.01.020] [PMID: 26807806]

[121] Zhang, C.J.; Tang, H.; Li, W.C.; Lin, H.; Chen, W.; Chou, K.C. iOri-Human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget,* **2016**, *7*(43), 69783-69793. [http://dx.doi.org/10.18632/oncotarget.11975] [PMID: 27626500]

[122] Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget,* **2016**, *7*(13), 16895-16909. [http://dx.doi.org/10.18632/oncotarget.7815] [PMID: 26942877]

[123] Liu, B.; Yang, F.; Chou, K.C. 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Ther. Nucleic Acids,* **2017**, *7*, 267-277. [http://dx.doi.org/10.1016/j.omtn.2017.04.008] [PMID: 28624202]

[124] Liu, B.; Wang, S.; Long, R.; Chou, K.C. iRSpot-EL: Identify recombination spots with an ensemble learning approach. *Bioinformatics,* **2017**, *33*(1), 35-41. [http://dx.doi.org/10.1093/bioinformatics/btw539] [PMID: 27531102]

[125] Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K.C. iRNA-AI: Identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget,* **2017**, *8*(3), 4208-4217. [http://dx.doi.org/10.18632/oncotarget.13758] [PMID: 27926534]

[126] Feng, P.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K.C. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids,* **2017**, *7*, 155-163. [http://dx.doi.org/10.1016/j.omtn.2017.03.006] [PMID: 28624191]

[127] Liu, B.; Yang, F.; Huang, D.S.; Chou, K.C. iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics,* **2018**, *34*(1), 33-40. [http://dx.doi.org/10.1093/bioinformatics/btx579] [PMID: 28968797]

[128] Ehsan, A.; Mahmood, K.; Khan, Y.D.; Khan, S.A.; Chou, K.C. A novel modeling in mathematical biology for classification of signal peptides. *Sci. Rep.,* **2018**, *8*(1), 1039. [http://dx.doi.org/10.1038/s41598-018-19491-y] [PMID: 29348418]

[129] Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics,* **2018**, *111*(1), 96-102. [http://dx.doi.org/10.1016/j.ygeno.2018.01.005] [PMID: 29360500]

[130] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *J. Theor. Biol*., **2015**, *377*, 47-56.

[131] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. J. M. iPPBS-Opt: A sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. Molecules, **2016**, *21*(1), 95.

[132] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. Dynamics, identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. *J. Biomol. Structure Dynamics,* **2016**, *34*(9), 1946-1961.

[133] Liu, B.; Wang, S.; Long, R.; Chou, K-C. iRSpot-EL: Identify recombination spots with an ensemble learning approach. *Bioinformatics,* **2017**, *33*(1), 35-41. [http://dx.doi.org/10.1093/bioinformatics/btw539] [PMID: 27531102]

[134] Qiu, W.-R.; Xiao, X.; Chou, K.-C. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.,* **2014**, *15*(2), 1746-1766.

[135] Song, J.; Wang, Y.; Li, F.; Akutsu, T.; Rawlings, N.D.; Webb, G.I.; Chou, K-C. iProt-Sub: A comprehensive package for accurately

mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.,* **2018**, *20*(2), 638-658. [PMID: 29897410]

[136] Xiao, X.; Ye, H.-X.; Liu, Z.; Jia, J.-H.; Chou, K.-C. iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget*, **2016**, *7*(23), 34180.

[137] Yang, H.; Qiu, W.-R.; Liu, G.; Guo, F.-B.; Chen, W.; Chou, K.-C.; Lin, H.J. iRSpot-Pse6NC: Identifying recombination spots in *Saccharomyces* cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.,* **2018**, *14*(8), 883.

[138] Liu, B.; Yang, F.; Chou, K-C. 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Ther. Nucleic Acids,* **2017**, *7*, 267-277. [http://dx.doi.org/10.1016/j.omtn.2017.04.008] [PMID: 28624202]

[139] Chou, K-C.; Wu, Z-C.; Xiao, X. iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.,* **2012**, *8*(2), 629-641. [http://dx.doi.org/10.1039/C1MB05420A] [PMID: 22134333]

[140] Lin, W-Z.; Fang, J-A.; Xiao, X.; Chou, K-C. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.,* **2013**, *9*(4), 634-644. [http://dx.doi.org/10.1039/c3mb25466f] [PMID: 23370050]

[141] Xiao, X.; Wu, Z-C.; Chou, K-C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.,* **2011**, *284*(1), 42-51. [http://dx.doi.org/10.1016/j.jtbi.2011.06.005] [PMID: 21684290]

[142] Xiao, X.; Wang, P.; Lin, W-Z.; Jia, J-H.; Chou, K-C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.,* **2013**, *436*(2), 168-177. [http://dx.doi.org/10.1016/j.ab.2013.01.019] [PMID: 23395824]

[143] Chou, K-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.,* **2013**, *9*(6), 1092-1100. [http://dx.doi.org/10.1039/c3mb25555g] [PMID: 23536215]

[144] Cheng, X.; Xiao, X.; Chou, K-C. pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. *Genomics,* **2017**, *110*(1), 50-58. [PMID: 28818512]

[145] Cheng, X.; Xiao, X.; Chou, K-C. pLoc-mPlant: Predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.,* **2017**, *13*(9), 1722-1727. [http://dx.doi.org/10.1039/C7MB00267J] [PMID: 28702580]

[146] Cheng, X.; Xiao, X.; Chou, K-C. pLoc-mVirus: Predict subcellular localization of multi-location virus proteins *via* incorporating the optimal GO information into general PseAAC. *Gene,* **2017**, *628*, 315-321. [http://dx.doi.org/10.1016/j.gene.2017.07.036] [PMID: 28728979]

[147] Cheng, X.; Xiao, X.; Chou, K-C. pLoc-mHum: Predict subcellular localization of multi-location human proteins *via* general PseAAC to winnow out the crucial GO information. *Bioinformatics,* **2018**, *34*(9), 1448-1456. [http://dx.doi.org/10.1093/bioinformatics/btx711] [PMID: 29106451]

[148] Cheng, X.; Xiao, X.; Chou, K-C. pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning *via* general PseAAC. *Genomics,* **2017**, *110*(4), 231-239. [http://dx.doi.org/10.1016/j.ygeno.2017.10.002] [PMID: 28989035]

[149] Cheng, X.; Zhao, S-G.; Lin, W-Z.; Xiao, X.; Chou, K-C. pLoc-mAnimal: Predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics,* **2017**, *33*(22), 3524-3531. [http://dx.doi.org/10.1093/bioinformatics/btx476] [PMID: 29036535]

[150] Xiao, X.; Cheng, X.; Su, S.; Mao, Q.; Chou, K-C. pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. *Nat. Sci.,* **2017**, *9*(9), 330. [http://dx.doi.org/10.4236/ns.2017.99032]

[151] Cheng, X.; Zhao, S-G.; Xiao, X.; Chou, K-C. iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics,* **2017**, *33*(3), 341-346. [http://dx.doi.org/10.1093/bioinformatics/btx387] [PMID: 28172617]

[152] Cheng, X.; Zhao, S.-G.; Xiao, X.; Chou, K.-C. iATC-mHyb: A hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget*, **2017**, 8(35), 58494.

[153] Chou, K.-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol Biosyst.,* **2013**, *9*(6), 1092-100. [http://dx.doi.org/10.1039/c3mb25555g]

[154] Goksuluk, D.; Korkmaz, S.; Zararsiz, G.; Karaagaoglu, A.E. easyROC: An interactive web-tool for ROC curve analysis using R language environment. *R. J.,* **2016**, *8*(2), 213-230. [http://dx.doi.org/10.32614/RJ-2016-042]

[155] Xiao, X.; Xu, Z-C.; Qiu, W-R.; Wang, P.; Ge, H-T.; Chou, K-C. iPSW(2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features *via* pseudo K-tuple nucleotide composition. *Genomics,* **2018**, S0888-7543(18)30613-X. [http://dx.doi.org/10.1016/j.ygeno.2018.12.001] [PMID: 30529532]

[156] Wang, J.; Li, J.; Yang, B.; Xie, R.; Marquez-Lago, T.T.; Leier, A.; Hayashida, M.; Akutsu, T.; Zhang, Y.; Chou, K-C. Bastion3: A two-layer ensemble predictor of type III secreted effectors. *Bioinformatics,* **2018**, *35*(12), 2017-2028. [PMID: 30388198]

[157] Srivastava, A.; Kumar, R.; Kumar, M. BlaPred: Predicting and classifying β-lactamase using a 3-tier prediction system *via* Chou's general PseAAC. *J. Theor. Biol.,* **2018**, *457*, 29-36. [http://dx.doi.org/10.1016/j.jtbi.2018.08.030] [PMID: 30138632]

[158] Song, J.; Li, F.; Takemoto, K.; Haffari, G.; Akutsu, T.; Chou, K-C.; Webb, G.I. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *J. Theor. Biol.,* **2018**, *443*, 125-137. [http://dx.doi.org/10.1016/j.jtbi.2018.01.023] [PMID: 29408627]

[159] Rasool, N.; Iftikhar, S.; Amir, A.; Hussain, W. Structural and quantum mechanical computations to elucidate the altered binding mechanism of metal and drug with pyrazinamidase from Mycobacterium tuberculosis due to mutagenicity. *J. Mol. Graph. Model.,* **2018**, *80*, 126-131. [http://dx.doi.org/10.1016/j.jmgm.2017.12.011] [PMID: 29331879]

[160] Mei, J.; Zhao, J. Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. *Sci. Rep.,* **2018**, *8*(1), 2359. [http://dx.doi.org/10.1038/s41598-018-20819-x] [PMID: 29402983]

[161] Li, F.; Wang, Y.; Li, C.; Marquez-Lago, T.T.; Leier, A.; Rawlings, N.D.; Haffari, G.; Revote, J.; Akutsu, T.; Chou, K-C.; Purcell, A.W.; Pike, R.N.; Webb, G.I.; Ian Smith, A.; Lithgow, T.; Daly, R.J.; Whisstock, J.C.; Song, J. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: A comprehensive revisit and benchmarking of existing methods. *Brief. Bioinform.,* **2018**. [http://dx.doi.org/10.1093/bib/bby077] [PMID: 30184176]

[162] Li, F.; Li, C.; Marquez-Lago, T.T.; Leier, A.; Akutsu, T.; Purcell, A.W.; Ian Smith, A.; Lithgow, T.; Daly, R.J.; Song, J.; Chou, K.C. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics,* **2018**, *34*(24), 4223-4231. [http://dx.doi.org/10.1093/bioinformatics/bty522] [PMID: 29947803]

[163] Muthu Krishnan, S. Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. *J. Theor. Biol.,* **2018**, *445*, 62-74. [http://dx.doi.org/10.1016/j.jtbi.2018.02.008] [PMID: 29476832]

[164] Arif, M.; Hayat, M.; Jan, Z. iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition. *J. Theor. Biol.,* **2018**, *442*, 11-21. [http://dx.doi.org/10.1016/j.jtbi.2018.01.008] [PMID: 29337263]

[165] Althaus, I.W.; Chou, J.; Gonzales, A.; Deibel, M.; Chou, K.; Kezdy, F.; Romero, D.; Aristoff, P.; Tarpley, W.; Reusser, F. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J. Biol. Chem.,* **1993**, *268*(9), 6119-6124.

[166] Althaus, I.W.; Chou, J.J.; Gonzales, A.J.; Deibel, M.R.; Kuo-Chen, C.; Kezdy, F.J.; Romero, D.L.; Thomas, R.C.; Aristoff, P.A.; Tarpley, W.G. Kinetic studies with the non-nucleoside human im-

munodeficiency virus type-1 reverse transcriptase inhibitor U-90152E. *Biochem. Pharmacol.,* **1994**, *47*(11), 2017-2028. [http://dx.doi.org/10.1016/0006-2952(94)90077-9]

[167] Althaus, I. W.; Gonzales, A.; Chou, J.; Romero, D.; Deibel, M.; Chou, K.-C.; Kezdy, F.; Resnick, L.; Busso, M.; So, A. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *J. Biol. Chem.,* **1993**, *268*(20), 14875-14880.

[168] Chou, K.; Forsen, S.; Zhou, G. Schematic rules for deriving apparent rate constants. **1980**, *16*(4), 109-113.

[169] Chou, K.-C.; Forsén, S. Graphical rules for enzyme-catalysed rate laws. *Biochem. J.,* **1980**, *187*(3), 829-35. [http://dx.doi.org/10.1042/bj1870829]

[170] Chou, K.-C.; Lin, W.-Z.; Xiao, X. Wenxiang: A web-server for drawing wenxiang diagrams. *Nat. Sci.,* **2011**, *03*(10). [http://dx.doi.org/10.4236/ns.2011.310111]

[171] Chou, K.-C. Graphic rules in steady and non-steady state enzyme kinetics. *J. Biol. Chem.,* **1989**, *264*(20), 12074-12079.

[172] Chou, K.-C. Applications of graph theory to enzyme kinetics and protein folding kinetics: Steady and non-steady-state systems. *Biophys. Chem.,* **1990**, *35*(1), 1-24.

[173] Chou, K.-C. Graphic rule for drug metabolism systems. *Curr. Drug Metab.,* **2010**, *11*(4), 369-378. [http://dx.doi.org/10.2174/138920010791514261]

[174] Chou, K. Graph theory of enzyme kinetics. *J. Physics Chem.,* **1979**, *60*, 1375-1378.

[175] Kuochen, C.; Carter, R.; Forsen, S. A new graphical-method for deriving rate-equations for complicated mechanisms. *J. Pre-Proof,* **1981**, *18*(2), 82-86.

[176] Kuo-Chen, C.; Forsen, S. Graphical rules of steady-state reaction systems. *Can. J. Chem.,* **1981**, *59*(4), 737-755. [http://dx.doi.org/10.1139/v81-107]

[177] Zhou, G.; Deng, M. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. *Biochem. J.,* **1984**, *222*(1), 169-76. [http://dx.doi.org/10.1042/bj2220169]

[178] Zhou, G.-P. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. *J. Theor. Biol.,* **2011**, *284*(1), 142-148. [http://dx.doi.org/10.1016/j.jtbi.2011.06.006]

[179] Chou, K.-c.; Forsén, S. Diffusion-controlled effects in reversible enzymatic fast reaction systems-critical spherical shell and proximity rate constant. *Biophys. Chem.,* **1980**, *12*(3-4), 255-263. [http://dx.doi.org/10.1016/0301-4622(80)80002-0]

[180] Chou, K.-C.; Li, T.-t.; Forsén, S. The critical spherical shell in enzymatic fast reaction systems. *Biophys. Chem.,* **1980**, *12*(3-4), 265-269. [http://dx.doi.org/10.1016/0301-4622(80)80003-2]

[181] Shen, H.-B.; Song, J.-N.; Chou, K.-C. Engineering, Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *J. Biomed. Sci. Eng*., **2009**, *2*, 136-143.

[182] Chou, K.; Chen, N.; Forsen, S. The biological functions of low-frequency phonons. 2. Cooperative effects. *Biophys. Chem.,* **1981**, *18*(3), 126-132.

[183] Chou, K.-C.; Shen, H.-B. Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.,* **2009**, *1*(2), 63-92 [http://dx.doi.org/10.4236/ns.2009.12011]

[184] Chou, K.-C. Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.,* **1988**, *30*(1), 3-48. [http://dx.doi.org/10.1016/0301-4622(88)85002-6]

[185] Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics,* **2017**, *33*(22), 3518-3523. [http://dx.doi.org/10.1093/bioinformatics/btx479] [PMID: 28961687]

[186] Xiao, X.; Cheng, X.; Chen, G.; Mao, Q.; Chou, K. pLoc_balmVirus: Predict subcellular localization of multi-label virus proteins by PseAAC and IHTS treatment to balance training dataset. *Med. Chem.,* **2018**, *15*(5), 496-509.

[187] Chou, K-C. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem.,* **2017**, *17*(21), 2337-2358. [http://dx.doi.org/10.2174/1568026617666170414145508] [PMID: 28413951]