Send Orders for Reprints to reprints@benthamscience.net

RESEARCH ARTICLE

iSulfoTyr-PseAAC: Identify Tyrosine Sulfation Sites by Incorporating Statistical Moments *via* Chou's 5-steps Rule and Pseudo Components

Omar Barukab[1], Yaser Daanial Khan[2], Sher Afzal Khan[1,4,*] and Kuo-Chen Chou[3]

[1]*Department of Information Technology, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, P.O. Box 344, Rabigh, 21911, Saudi Arabia;* [2]*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;* [3]*Gordon Life Science Institute, Boston, MA 02478, USA;* [4]*Department of Computer Sciences, Abdul Wali Khan University, Mardan, Pakistan*

ARTICLE HISTORY

Received: May 15, 2019
Revised: August 04, 2019
Accepted: August 06, 2019

DOI:
10.2174/1389202920666190819091609

Abstract: ***Background*:** The amino acid residues, in protein, undergo post-translation modification (PTM) during protein synthesis, a process of chemical and physical change in an amino acid that in turn alters behavioral properties of proteins. Tyrosine sulfation is a ubiquitous posttranslational modification which is known to be associated with regulation of various biological functions and pathological processes. Thus its identification is necessary to understand its mechanism. Experimental determination through site-directed mutagenesis and high throughput mass spectrometry is a costly and time taking process, thus, the reliable computational model is required for identification of sulfotyrosine sites.

***Methodology*:** In this paper, we present a computational model for the prediction of the sulfotyrosine sites named iSulfoTyr-PseAAC in which feature vectors are constructed using statistical moments of protein amino acid sequences and various position/composition relative features. These features are incorporated into PseAAC. The model is validated by jackknife, cross-validation, self-consistency and independent testing.

***Results*:** Accuracy determined through validation was 93.93% for jackknife test, 95.16% for cross-validation, 94.3% for self-consistency and 94.3% for independent testing.

***Conclusion*:** The proposed model has better performance as compared to the existing predictors, however, the accuracy can be improved further, in future, due to increasing number of sulfotyrosine sites in proteins.

Keywords: Sulfation, sulfotyrosine, statistical moments, PseAAC, 5-step rule, pseudo components.

1. INTRODUCTION

Proteins are the diverse macromolecules in living organisms and have an important role in all biological development of organisms [1]. Proteins, as an enzyme, boost chemical reaction within a cell and produce movement, broadcast nerve force and increase muscle growth. These proteins are comprised of amino acid residues, joined by a peptide bond to make a polypeptide chain in protein. The amino acid residues, in protein, undergo post-translation modification (PTM) during protein synthesis, a process of chemical and physical change in an amino acid that in turn alters behavioural properties of proteins [2, 3]. The process of protein synthesis starts from the nucleus where ribonucleic acid (RNA) copies code for specific proteins from Deoxyribonucleic acid (DNA) then messenger ribonucleic acid(mRNA) takes the copy to protein-making factory namely ribosome in the cytoplasm. The ribosome with transfer ribonucleic acid (tRNA) continues to add a correct sequence of amino acid till it receives ending codon from mRNA thus making the protein ready to function. PTM could occur during or after protein synthesis to enhance proteomics range, control cell action and to use the same proteins for various cell functions [4, 5].

Tyrosine sulfation is a ubiquitious posttranslation modificaiton which is known to be associated with the regulation of various biological functions including protien-protein interactions, transportation modulation, and the proteolysis [6, 7]. Besides all this, the tyrosine sulfaction is linked with various pathalogical processes includng HIV infection, atherosciclerosis, and numerous lung diseases [4, 5, 8]. This depicts the dire need of idnetifying the mechanism of tyrosine sulfation which cannot be understood without the identification of tyrosine sulfation sites [6, 9, 10]. Thus, identification of tyrosine sulfation sites is of great importance. Although, the sites can be identified through various experimetnal techniques including site directed mutagenesis and high throughput mass spectrometery, however, all these techniques are laborius, time taking and costly. Therfore, the identification of sulfotyrosine sites through computaitonal predictors is one of the most optimal approaches and for this purpose, various researchers have

*Address correspondence to this author at the Department of Information Technology, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, P.O. Box 344, Rabigh, 21911, Saudi Arabia; and Department of Computer Sciences, Abdul Wali Khan University, Mardan, Pakistan; E-mail: sher.afzal@awkum.edu.pk

propsed different methods previously for the identification of sulfotyrosine site. Also, the use of computaitonal predictors for the identification of sulfotyrosine sites can help process large scale proteomic data as well. Computational predictors using the neural network and statistical moments for feature extraction has been developed and used previously. In the last few years, many studies have been reported by the previous investigators in the field of bioinformatics and computational biology, which help in identifying the function and characteristics of proteins [3, 10-25]. Besides these, various papers have been reported targeting the prediction of PTM [3, 11-16, 18-62].

Yu and coworkers [63] used position specific scoring matrix (PSSM) to predict the sulfotyrosine sites in proteins. Later on, Sulfinator [64] named predictor was proposed by Monigatti and coworkers to identify sulfotyrosine sites using hidden markov models and sequence alignment information. A further improvement in sulfotyrosine predictors was observed with the development of SulfoSite [65], by incorporating acessible surface area and prositional weighted matrix for the prediction of tyrosine sulfation sites. Niu and coworkers [66] proposed another predictor for sulfotyrosine sites based on sequence and amino acid level information. In 2012, PredSulSite [67] was prposed by Huang *et al.* which incorported various features such as secondary strucutre inforamtion, physiochemcial characteristics and residue postion information. Various models were trained while SVM outperforemd the counterparts. Later on, in 2014, another SVM based method named SulfoTyrP [68] was propsoed by Jia *et al.* which is supposed to be the most accurate method for prediction of sulfotyrosine to date. Although, these predictors have been propsoed for sulfotyrosine sites, still there are limitations in the accuracy of prediction.

Herein, we propose a computational model named iSulfoTyr-PseAAC for the prediction of Sulfotyrosine sites in proteins. The dataset used in this model is experimentally verified and updated. The feature vectors are constructed using statistical moments of protein amino acid sequences and various position/composition relative features. These features are incorporated into PseAAC [69]. The whole process is carried out by the aid of Chou's 5-step rule [70] which are followed by current studies [12, 27, 28, 52, 71-77]. As demonstrated by a series of recent publications [12, 14, 17, 18, 20, 23, 49, 57, 59, 73, 78-94] and summarized in two comprehensive review papers [69, 95], to develop a really useful predictor for a biological system, one needs to follow Chou's 5-step rule to go through the following five steps: (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) easily to repeat the reported results by other investigators, (4) with 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

This section elaborates the first three phases of Chou's 5-step rule. Fig. (**1**) explains that at first stage raw data with the standard format is collected from online protein database known as UniProt. Raw data undergoes the process of filtration at the second stage. The filtration process removes duplicated data and extracts sequences which are most suitable for sulfotyrosine. After the process of filtration, features are extracted of selected sequences. At the last stage filtered data are used for training purpose then the trained neural network is tested with different dataset.

2.1. Dataset Collection

The data used for the prediction of sulfotyrosine sites was taken from the UniProt Protein database. The UniProt database is verified and contains complete features of all proteins. Dataset was downloaded in the XML format, which was processed to extract sequences along accession number. For purposed technique Data of two types, *i.e.* positive and negative type was gathered from the UniProt. Preprocessing was performed on both sets of data to remove any duplication. The data have only alphabetic sequences. The positive dataset contained all the sequences which have experimental evidence of sulfotyrosine sites. The positive dataset contained those protein sequences which were explained with the field PTM/Processing. Dataset quality was enhanced by removing proteins which were not reviewed. On both sides of tyrosine (Y), 20 amino acid residues were selected.

Taking into account Chou's scheme [70], a protein containing tyrosine site can be expressed as:

$$\mathrm{K}_{\rho}(\mathrm{B}) = \mathrm{M}_{-\rho}\mathrm{M}_{-(\rho-1)}\cdots\mathrm{M}_{-2}\mathrm{M}_{-1}Y\mathrm{M}_{+1}\mathrm{M}_{+2}\cdots\mathrm{M}_{+(\rho-1)}\mathrm{M}_{+\rho} \quad (1)$$

Amino acid code *Y* is the targeted tyrosine residue in this equation, the character ρ is an integer, $\mathrm{M}_{-\rho}$ represent ρ-th upstream amino acid residue from the centre, $\mathrm{M}_{+\rho}$ represents ρ +th downstream amino acid residue from the centre. $(2\rho+1)$ a tuple can be illustrated in 2 types:

$$\mathrm{K}_{\nu}(Y)\varepsilon\begin{Bmatrix}\mathrm{K}_{\nu}^{+}(Y)\\ \mathrm{K}_{\nu}^{-}(Y)\end{Bmatrix} \quad (2)$$

The following condition $\mathrm{K}_{\nu}^{+}(Y)$ holds if the centre is sulfotyrosine site, it is not true than $\mathrm{K}_{\nu}^{-}(Y)$ holds. Set theory represents ε symbol as "a member of".

Testing and training dataset is developed for the statistical prediction model. The model is trained using training dataset then tested using testing dataset. The matter is extensively illustrated in [32], 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 prediction model because result acquired in the way is from a combination of many different independent dataset results. In this research paper, the ideal value of ρ for test

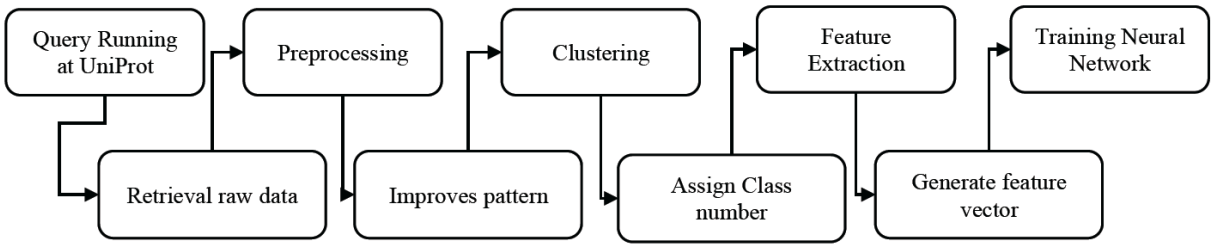

Fig. (1). Detailed step for proposed methodology. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

is 20, meanwhile, the dataset has $(2\rho+1)$=41 residues. Considering all, the dataset was minimized to

$$\mathrm{T} = \mathrm{T}^{+} \cup \mathrm{T}^{-} \tag{3}$$

In the equation T^{+} hold 200 positive sample, T^{-} holds 420 negative sample and $\cup$ represents "union of two set". In total 200+420 = 620 samples are included in benchmark dataset (Supplementary information **S1**).

2.2. Feature Vector Construction

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms (such as "Optimization" algorithm [96], "Covariance Discriminant" or "CD" algorithm [97, 98], "Nearest Neighbor" or "NN" algorithm [99], and "Support Vector Machine" or "SVM" algorithm [99, 100] can only handle vectors as elaborated in a comprehensive review [56]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [97] or PseAAC [101] was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics [102-112] as well as a long list of references cited in [113]. Because it has been widely and increasingly used, four powerful open access soft-wares, called 'PseAAC' [114], 'PseAAC-Builder' [115], 'propy' [116], and 'PseAAC-General' [117], were established: the former three are for generating various modes of Chou's special PseAAC [118]; while the 4th one for those of Chou's general PseAAC, including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs. 9, 10 of [69]), "Gene Ontology" mode (see Eqs. 11, 12 of [69]), and "Sequential Evolution" or "PSSM" mode (see Eqs. 13, 14 of [69]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [119] was developed for generating various feature vectors for DNA/RNA sequences [120-122] that have proved very useful as well. Particularly, recently a very powerful web-server called 'Pse-in-One' [123] and its updated version 'Pse-in-One2.0' [124] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

For the help in feature vector construction, Chou's computational model sample formation was implemented. 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 protein sequence. For studying the properties of the protein, construction of feature vector holds the primary position. An array of the amino acids is utilized to develop a feature vector that increases the probability of site prediction in protein. Proteins' performance is determined by amino acid location and little change in location modifies protein qualities. Feature vector sequences represented by feature vector is broadly utilized in predicting different structural characteristic [27, 28, 49, 53, 57, 85, 87, 125-127].

2.2.1. Site Vicinity Vector

Many elements make some sites in protein sensitive to post-translational modification. Most elements are environmental while neighbouring resides in peptide chain makes sites more sensitive to modification [35]. Supposing ζ_x to be PTM site, then neighbouring resides is represented as:

$$T = \{\zeta_1 \ldots \zeta_{x-2}, \zeta_{x-1}, \zeta_x, \zeta_{x+1}, \zeta_{x+2}, \zeta_{x+3}, \ldots\ldots \zeta_n\} \tag{4}$$

Substructure in a primary sequence which contains possible sites and its neighbour help in making site vicinity vector such as,

$$\zeta_{x-r} \ldots \zeta_{x-2}, \zeta_{x-1}, \zeta_x, \zeta_{x+1}, \zeta_{x+2}, \ldots\ldots \zeta_{x+r} \tag{5}$$

In this equation, ***r*** is an integer chosen through testing and experiments. In feature vector, site vicinity vector form sections that are awarded various numerical value replacing every residue position. Only 20 amino acids are important for protein synthesis and for calculating feature vector, every amino acid is given special integral value. If the values are changed, sections are allocated regularly and it doesn't make a difference which number is assigned to which amino acid.

2.2.2. Statistical Moments

The numerical quantity that describes various characteristics or distribution of data is called Statistical moments. These moments explains the shape of data's histogram and provides data which is enough for making frequency distribution function. It helps to quantify the symmetry of data in a set by the use of variation and skewness. Mathematicians and analysts have shaped different moments in the light of certain outstanding polynomials. Raw, central and Hahn moments are utilized to illustrate their polynomial tasks. A raw moment is the mean of all number in a set with

order k, before taking mean each number is raised to the k^{th} power. The first raw moment is the mean of all addition, second is average of squared number, while third the average of the cubed number. Change and unevenness made by the composed dataset are calculated by these moments [126-129].

The central moments are also used for the same purpose. A central moment is dependent at the average of the difference between numbers from their mean. The second central moment is achieved by the squared differences before averaging, while the third central moment is achieved by cubing difference before averaging. Hahn moments is used widely for feature extraction. These moments are dependent at Hahn polynomials [129]. It is used as an input in a neural network for providing an asymmetric grouping of feature selection beside classification.

Merely 20 amino acids are present, for calculating moments every amino acid is allocated exclusive numerical value. Since the values are distinctive, numerical values are allocated again and again, so any value can be allocated to any amino acid. The 1-dimensional grouping of the amino acid is changed into 2-dimensional form.

Suppose **S** stand for series of protein and sequence is given as:

$$S = \{\beta_1, \beta_2, \beta_3, \ldots, \beta_{m-1}, \beta_m\} \tag{6}$$

m residue exists in the primary sequence of the protein, where β_i is the i^{th} amino acid residue, also let,

$$z = \left\lceil \sqrt{m} \right\rceil$$

All amino acid component of protein S are held by matrix S' created with $\mathrm{m} \times \mathrm{m}$ dimensions.

$$\mathrm{S}' = \begin{bmatrix} \kappa_{11} & \kappa_{12} & \cdots & \kappa_{1n} \\ \kappa_{21} & \kappa_{22} & \cdots & \kappa_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{m1} & \kappa_{m2} & \cdots & \kappa_{mm} \end{bmatrix} \tag{7}$$

The 2-dimensional matrix S' corresponds to the matrix S. The matrix S is converted to S' by using V as the mapping function.

$$v(\beta_x) = \alpha_{pq} \tag{8}$$

Where $p = \frac{c}{d} + 1$ and $q = c \bmod d$ if S' is populated in row-major order.

Moments till 3 degrees are calculated using a 2D matrix S', the following equation is used for calculating raw moments.

$$Z_{mn} = \sum_{x=1}^{l} \sum_{y=1}^{l} x^m y^n \alpha_{xy} \tag{9}$$

Where m+n denotes the order of moments. Moments till level three are calculated as $\mathbf{Z_{00}}$, $\mathbf{Z_{01}}$, $\mathbf{Z_{02}}$, $\mathbf{Z_{10}}$, $\mathbf{Z_{11}}$, $\mathbf{Z_{12}}$, $\mathbf{Z_{20}}$, $\mathbf{Z_{21}}$, $\mathbf{Z_{30}}$ and $\mathbf{Z_{03}}$.

The data centre is similar to the centre of gravity. Data is fairly distributed at the data's central point with reference to average weight. It is calculated after calculation of raw moments. It is known as an argument $(\overline{v}, \overline{w})$where,

$$\overline{v} = \frac{Z_{10}}{Z_{00}} \quad \text{and} \quad \overline{w} = \frac{Z_{01}}{Z_{00}} \tag{10}$$

Central moments are calculated with the help of centroid. Central moments lies at data central point where centroid acts as data's centre of gravity. Following equation is used to calculate central moments.

$$\mathrm{B}_{st} = \sum_{k=1}^{m} \sum_{l=1}^{m} (k - \overline{v})^s \ (l - \overline{w})^t \alpha_{kl} \tag{11}$$

In order to calculate Hahn moment, 1-dimensional interpretation S was converted to a square matrix interpretation S′. Two-dimensional input data is needed by two dimensional Hahn moments. The Hahn polynomial of order n is given as:

$$\omega_m^{a,b}(p,M) = (M+b-1)_m (M-1)_m \times \sum_{l=0}^{m} (-1)^l \frac{(-m)_l (-p)_l (2M+a+b-m-1)_l}{(M+b-1)_l (M-1)_l} \frac{1}{l!} \tag{12}$$

The above expression uses the Pochhammer symbol generalized as:

$$(b)_l = b.(b+1)\ldots(b+l-1) \tag{13}$$

And is simplified using the Gamma operator.

$$(b)_l = \frac{\Delta(b+l)}{\Delta(b)} \tag{14}$$

The raw values of Hahn moments are usually scaled using a weighting function and a square norm is given as:

$$\widetilde{\beta}_m^{a,b}(p,M) = \beta_m^{a,b}(p,M) \sqrt{\frac{o(p)}{c_m^2}} \qquad m = 0,1,\ldots M-1 \tag{15}$$

While,

$$o(p) = \frac{\varphi(a+p+b)\varphi(b+p+1)(a+b+p+1)_M}{(a+b+2p+1)m!(M-p-1)!} \tag{16}$$

The orthogonal normalized Hahn for the two-dimensional discrete data are computed using the following equation:

$$G_{ef} = \sum_{a=0}^{M-1} \sum_{b=0}^{M-1} \alpha_{ab} \widetilde{J}_t^{c,d}(a,M) \widetilde{J}_s^{u,v}(b,M), \qquad m,n = 0,1,\ldots M-1 \tag{17}$$

The central moments and the Hahn moments are computed up to order 3.

2.2.3. Position Relative Incidence Matrix

Informational series is the root of a mathematical model that predict that role of proteins. Location of amino acid plays a key role in determining the physical properties of the protein. It is also important to minimize placement of amino acid in the polypeptide chain. Position relative incidence matrix (PRIM) extracts location information of amino acid in the polypeptide chain. The matrix of PRIM is made with 20x20 dimensions as given below.

$$Z_{PRIM} = \begin{bmatrix} Q_{1\to1} & Q_{1\to2} & Q_{1\to3} & Q_{1\to b} & \cdots & Q_{1\to20} \\ Q_{2\to1} & Q_{2\to2} & Q_{2\to3} & Q_{2\to b} & \cdots & Q_{2\to20} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ Q_{d\to1} & Q_{d\to2} & Q_{d\to3} & Q_{d\to b} & \cdots & Q_{d\to20} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ Q_{U\to1} & Q_{U\to2} & Q_{U\to3} & Q_{U\to b} & \cdots & Q_{U\to20} \end{bmatrix} \quad (18)$$

An item $Q_{d\to b}$ holds the total of b^{th} residue against the first occurrence of d^{th} residue. Prim makes 400 coefficient which is a large number. For reducing the coefficient more, moments.

2.2.4. Reverse Position Relative Incidence Matrix

Machine learning algorithm accuracy mostly depends on the perfection of data's feature extraction and the algorithm is able to change itself for understanding data's unclear pattern. The relative positioning of amino acid in the polypeptide chain is extracted by PRIM matrix. Similar workflow at the reverse primary sequence is followed by Reverse Position Relative Incident Matrix (RPRIM). Addition of RPRIM reveals more hidden pattern and uncertainties among proteins in the polypeptide sequence. Similar to PRIM, RPRIM also has 400 elements with 20x20 dimension. RPRIM matrix is represented as:

$$Q_{RPRIM} = \begin{bmatrix} R_{1\to1} & R_{1\to2} & \cdots & R_{1\to k} & \cdots & R_{1\to20} \\ R_{2\to1} & R_{2\to2} & \cdots & R_{2\to k} & \cdots & R_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ R_{t\to1} & R_{t\to2} & \cdots & R_{t\to k} & \cdots & R_{t\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ R_{z\to1} & R_{z\to2} & \cdots & R_{z\to k} & \cdots & R_{z\to20} \end{bmatrix} \quad (19)$$

The dimension of RPRIM matrix is minimized by calculating raw, central and Hahn moments.

2.2.5. Frequency Matrix

The amino acid sequence makes the native shape of the protein and their number of occurrence is calculated by the frequency matrix. Frequency matrix has a vital role in protein alignment. The amino acid series information is retrieved by PRIM and frequency matrix does not hold series information. The frequency matrix is calculated by the given formula:

$$\xi = \{\tau_1, \tau_2, \tau_3, \tau_4, \ldots, \tau_{20}\} \quad (20)$$

In this formula τ_i represents the frequency of i^{th} native amino acid.

2.2.6. Accumulative Absolute Position Incidence Vector

Amount of Amino acid residue in the polypeptide chain is represented by a frequency matrix and it also gives information relevant to protein formation. The frequency matrix lacks information relevant to the position of amino acid residues in the polypeptide chain and this deficit is accommodated by Accumulative Absolute Position Incidence Vector (AAPIV). AAPIV represent relevant positioning of amino acid residues in the polypeptide chain. A vector containing 20 elements is made where every element has a numerical ordered value that represents relevant residue in the primary sequence. Primary sequence showing the occurrence of specific residue in the structure is represented as:

$$\upsilon^k_{r^1} \cdots \upsilon^k_{r^2} \cdots \upsilon^k_{r^3} \cdots \upsilon^k_{r^n} \quad (21)$$

It shows that residue υ^k located at a position $r^1, r^2, r^3, \cdots r^n$

Let AAPIV be represented as:

$$T = \{\nu_1, \nu_2, \nu_3, \nu_4, \cdots, \nu_{20}\} \quad (22)$$

Therefore the i[th] element of AAPIV is calculated as:

$$\nu_i = \sum_{t=1}^{n} st \quad (23)$$

2.2.7. Reverse Accumulative Absolute Position Incidence Vector

As prior discussion, feature extraction is efficient in detecting an ambiguous pattern. Reverse accumulative absolute position incidence vector (RAAPIV) performs the same task, it is made from reversed AAPIV string. RAAPIV contain 20 elements is shown as:

$$\delta = \{o_1, o_2, o_3, o_4, o_5, \cdots, o_{20}\} \quad (24)$$

Specific residue in the Reversed sequence is shown as:

$$\omega^k_{m1} \cdots \omega^k_{m2} \cdots \omega^k_{m3} \cdots \omega^k_{mn} \quad (25)$$

In the sequence above residue ω^k occur in reverse sequence and $m_1, m_2, m_3, \ldots, m_n$ are their ordered location. The value of any element is calculated as:

$$\ell_i = \sum_{m=1}^{n} t_m \quad (26)$$

2.3. Neural Network

The neural network is one of the most important tools for solving the problem discussed in this paper, it simulates processing information as shown in Fig. (**2**). Neural network explains the basic shape of each residue in a given protein. For training the network, negative and positive samples are made that are used to calculate feature vector which represents 2-dimensional protein structures by using raw, central and Hahn moments.

2.3.1. Gradient Descent and Adaptive Learning

Different algorithms with different characteristic and performance are available to train the neural network. Among all, Gradient Decent algorithm performs the best. It is an iterative minimization method that finds out best set of weight which is used for making a prediction during neural network training. The main objective of algorithms is to find weights that reduce the error of the model on the training dataset. The training process is started by randomly guessing set of weight, the weight set whose loss function has more steps down value is selected. The process is repeated following a negative gradient until a satisfied lowest point is found

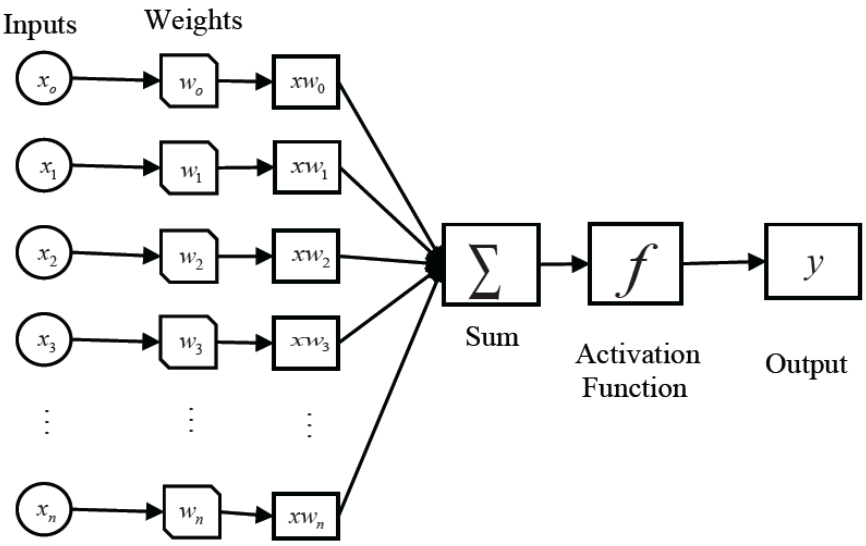

Fig. (2). Architecture of the artificial neural network for the iSulfoTyr-PseAAC.

and then the gradient of the loss function is calculated against all parameters. A gradient is a multidimensional vector containing the slope of loss function along every axis [126, 127].

The weight **W** is updated with the help of learning rate **R**, objective function **F(W)** and its gradient **£F(W)**. The central goal of the algorithm is to find the ideal weight **W** by minimizing **F(W)**. Depending on this algorithm, the parameters are iteratively computer at every stage by given equation.

$$W=W\text{-}R.\ £F(W) \quad (27)$$

Algorithm execution depends at learning rate **R** and it is mostly kept constant. It defines the time for function minimization and small learning rate requires more time to reach an optimal point whereas high learning rate may lead function to never reach the optimal point, thus, learning rate should have the ideal value to reach the optimal point. Mostly the starting process starts with a higher learning rate which slowly decreases as training proceeds. The learning rate may change at each layer which reduces the chance of gradient vanish. Weights stop to change at the first layer. Considering W_i and W_{i+1} calculated sequentially parameters. Using this parameter weight, output and expected error are calculated. Comparing with the previous iteration if the error is greater than the learning rate is decreased or if the error is smaller than the learning rate is increased, weights are excluded and new weight W_{i+1} is calculated. Weight calculation at each iteration is represented as (W_1, W2, W3, W4 …). The following equation is used to calculate weight for the successive epoch.

$$W_{t+1}=W_t\text{-}R_t.\ £L\ (W_t) \quad (28)$$

In the equation, R_t is used for t^{th} epoch. The adaptive algorithm guarantees normalization of learning rate while minimizing function at each epoch. Following condition is fulfilled before choosing the learning rate.

$$L\ (W_0) \geq L\ (W_1) \geq L\ (W_2)\ .\ . \quad (29)$$

3. RESULTS AND DISCUSSION

3.1. Accuracy Estimation

The objective evaluation of a newly developed predictor is a very important aspect, which helps to assess the success rate of that model [69]. 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 will formulate the metrics for objective evaluation, then we will employ various validation methods.

3.2. Formulation of Metrics

For objective evaluation, one needs to consider the metrics of evaluation 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 perdition 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. Either the set of traditional metrics copied from math books or the intuitive metrics derived from the Chou's symbols [70, 130, 131] are valid only for the single-label systems (where each sample only belongs to one class). Initially, these measures have been introduced in [132], and a set of four intuitive equation have been derived in [133, 134] for all these measures, which are:

$$\begin{cases} Sn = 1 - \dfrac{\mathcal{N}_-^+}{\mathcal{N}^+} & 0 \leq \text{Sn} \leq 1 \\ Sp = 1 - \dfrac{\mathcal{N}_+^-}{\mathcal{N}^-} & 0 \leq \text{Sp} \leq 1 \\ Acc = 1 - \dfrac{\mathcal{N}_-^+ + \mathcal{N}_+^-}{\mathcal{N}^+ + \mathcal{N}^-} & 0 \leq \text{Acc} \leq 1 \\ MCC = \dfrac{1 - \left(\dfrac{\mathcal{N}_-^+}{\mathcal{N}^+} + \dfrac{\mathcal{N}_+^-}{\mathcal{N}^-}\right)}{\sqrt{\left(1 + \dfrac{\mathcal{N}_+^- - \mathcal{N}_-^+}{\mathcal{N}^+}\right)\left(1 + \dfrac{\mathcal{N}_-^+ - \mathcal{N}_+^-}{\mathcal{N}^-}\right)}} & -1 \leq \text{MCC} \leq 1 \end{cases} \quad (30)$$

Where $\mathcal{N}^-$ represents the total number of non-sulfotyrosine sites, correctly predicted as non-sulfotyrosine sites by iSulfoTyr-PseAAC. $\mathcal{N}_+^-$ represents the total number non-sulfotyrosine sites which are predicted incorrectly as sulfotyrosine sites by iSulfoTyr-PseAAC. Moreover, $\mathcal{N}^+$ is the total number of sulfotyrosine sites which are correctly predicted as sulfotyrosine sites by iSulfoTyr-PseAAC and $\mathcal{N}_-^+$ is the total number of sulfotyrosine sites which are predicted incorrectly as the non- sulfotyrosine sites by iSulfoTyr-PseAAC. Thus, Eq. (30) gives the explanation of specificity, sensitivity, overall-accuracy, and stability more easy to understand and intuitive, particularly when we talk about MCC [135-137].

This set of perceptive metrics have been used by a number of modern publications [14-16, 20-23, 30, 80, 89, 133, 138-158], but only for binary labelled data. Multi-label prediction is a completely different problem, which has been more popular in computational biology [159-161] and biomedicine [162]. Thus, it requires a different kind of metrics [163]. For the multi-label systems (where a sample may simultaneously belong to several classes), whose existence has become more frequent in system biology [73, 164-170], system medicine [171, 172] and biomedicine [35], a completely different set of metrics as defined in [173] is absolutely needed.

3.3. Self-consistency Testing

To test the proposed prediction model accuracy, self-consistency testing was performed in which same training and testing datasets were used through using which the model was built. There is a reason for doing the self-consistency test and that is, 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 the proposed model has the 99.23% Acc, 99.10% Sp, 99.75% Sp, and 0.99 MCC.

3.4. Validation of Model

In general, prediction models are trained using experimentally proven dataset for prediction but some of the time we don't have experimentally proven datasets for model prediction testing. Interestingly, if somehow we have the experimentally proven dataset, it might be possible that data is not suitable or not sufficient for model testing against the prediction accuracy. To check the score four metrics of Eq. (30), what kind of testing method should be used to check the accuracy reliability of prediction model? Normally, a prediction model can be tested using Leave-one-out (jackknife), k-folds (Subsampling) and independent test [174].

3.4.1. Jackknife Testing

In jackknife testing, every time model is trained on N – 1, where N is a total number of instances of benchmark dataset and testing is done by the rest of the 1 instance of benchmark dataset. Each time data for training and testing is selected randomly and the model is trained and tested according to that datasets.

In jackknife validation of prediction model, training and testing both datasets are open and every sample of the benchmark dataset is used for training and testing, it's very exhaustive because of huge turn in and out of data samples and it excludes the memory effects. Its validation always gives different output for given benchmark dataset instances. The arbitrariness problem caused by independent test and subsampling completely avoided by using jackknife. Using jackknife, perdition model validation gives 97.07% accuracy as shown in Table **2**. It has been widely used to validate the prediction model by investigators [78, 145, 175-184].

3.4.2. K-fold Cross-Validation

Cross-validation is one of the best available methods to validate model prediction, cross-validation is the best option to choose and to give the validation that the proposed model is predicting true Sulfotyrosine sites.

Using cross-validation, the benchmark dataset is distributed into total k number of unique folds, where k is the number in which the benchmark dataset is divided, for now, k=10. In each round of validation, a different subset of data is selected randomly for validation across the rest of the data, by this, each part of the dataset is used for training and testing both. At the end of last round of cross-validation, the cumulated accuracy for k=10 is calculated by adding the accuracy of each validation round and dividing it by 10 and it's 94.26% in this study as shown in Table **3**.

This shows that the accuracy of the proposed method is higher than the other previously proposed methods for sulfotyrosine site prediction, as shown in Fig. (**3**).

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

Predictor	Accuracy Metrics			
	Acc (%)	Sp (%)	Sn (%)	MCC
iSulfoTyr-PseAAC	99.23	99.10	99.75	0.99

Table 2. Results for jackknife testing of iSulfoTyr-PseAAC (Average of n-iterations).

Predictor	Accuracy Metrics			
	Acc (%)	Sp (%)	Sn (%)	MCC
iSulfoTyr-PseAAC	97.07	97.39	96.96	0.92

Table 3. Results for 10-fold cross-validation of iSulfoTyr-PseAAC (Average of 10-folds).

Predictor	Accuracy Metrics			
	Acc (%)	Sp (%)	Sn (%)	MCC
iSulfoTyr-PseAAC	94.26	94.55	94.16	0.86

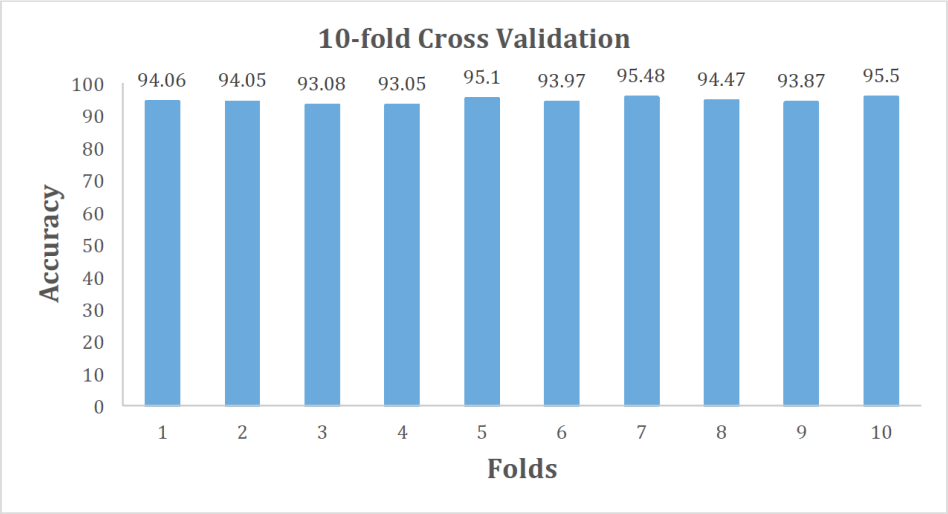

Fig. (3). 10-fold cross validation of iSulfoTyr-PseAAC. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein, as indicated by many previous studies on a series of important biological topics, [185-198], particularly in enzyme kinetics, protein folding rates [192, 199-201], and low-frequency internal motion [199-204] (Table **4**).

3.5. Comparative Analysis

In a comparative analysis of iSulfoTyr-PseAAC, the results of iSulfoTyr-PseAAC for the metrics of Eq. (30) are compared with already existing methods. For this purpose, an independent dataset of 80 positive and 80 negative samples was used.

Numerous imperative highlights make the proposed approach dignified and detailed from previous methods. First of all standard and balanced dataset has been included, which is experimentally verified and is of discrete nature. Secondly, the data is non repetitive, precise and complete in scope. Moreover, performance evaluation of proposed model is performed with 10 fold cross validation. The proposed model uses artificial neural networks which carefully handle dependence.

iSulfoTyr-PseAAC applies a novel approach and uses the compositional and positional features of primary sequences of protein to perform the prediction of Sulfotyrosine sites. In first, it uses PseAAC and cut the sequence by modified residue from 20 downstream and upstream, then calculate the AAPIV, RAAPIV, PRIM, RPIRM, and statistical moments, using the compositional and positional features of primary sequences of protein, iSulfoTyr-PseAAC outperforms its counterparts.

Table 4. Comparison with existing models.

Predictor	Accuracy Metrics				Number of Proteases			
	Acc (%)	*Sp (%)*	*Sn (%)*	*MCC*	$\mathcal{N}^{+}$	$\mathcal{N}^{+}_{-}$	$\mathcal{N}^{-}_{+}$	$\mathcal{N}^{-}$
iSulfoTyr-PseAAC	85.63	88.75	82.50	0.71	71	9	14	6
Sulfinator [64]	68.13	71.25	65.00	0.36	57	23	28	52
SulfoSite [65]	73.13	76.25	70.00	0.46	61	19	24	56
PredSulSite [67]	76.88	81.25	72.50	0.54	65	15	22	58
SulfoTyrP [68]	80.00	83.75	76.25	0.60	67	13	19	61

4. WEB SERVER

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 [33, 136, 143, 146, 165, 166, 169, 171]. As pointed out in [203] and demonstrated in a series of recent publications [18, 31, 33, 59, 71-74, 78, 81, 82, 87, 88, 93, 153, 158, 164-170, 205], 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 [56], driving medicinal chemistry into an unprecedented revolution [206]. Accordingly, in our future work, we shall strive to establish a web-server for the new method presented in this paper.

CONCLUSION

In this study, using Chou's 5-step rule we have developed a model for sulfotyrosine sites prediction based on ANN. Due to its strong biological importance, the finding of sulfotyrosine sites positions is a primary and essential task. The aim of the study is to develop an efficient and more accurate sulfotyrosine sites predictor and enhance it in usage. By implementing the PseAAC we have used many positional and compositional features of proteins samples. After model development, the prediction model was tested and validated against various exhaustive validation methods and techniques *i.e.* self-consistency, cross-validation, and jackknife. The self-consistency validation gives the 99.23% accuracy, for cross-validation the accuracy is 94.26% and jackknife gives 97.07% accuracy. The prediction models give overall 97.07% accuracy, sensitivity value 96.96% and specificity 97.39%. Using the above-mentioned accuracy and other values it concludes, the proposed model iSulfoTyr-PseAAC for prediction of sulfotyrosine site has the great ability to predict these sites in given proteins. In computational ways, the proposed model still can be improved as the number of protein sequences is rapidly growing, day to day.

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

Not applicable.

FUNDING

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under grant no. KEP-11-611-39.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

The authors acknowledge with thanks DSR for technical and financial support.

SUPPLEMENTARY MATERIAL

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

REFERENCES

[1] Whitford, D. *Proteins: structure and function.* John Wiley and Sons: **2013**.

[2] Lazure, C.; Seidah, N.G.; Pélaprat, D.; Chrétien, M. Proteases and posttranslational processing of prohormones: A review. *Can. J. Biochem. Cell Biol.*, **1983**, *61*(7), 501-515. [http://dx.doi.org/10.1139/o83-066] [PMID: 6354396]

[3] 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]

[4] Farzan, M.; Babcock, G.J.; Vasilieva, N.; Wright, P.L.; Kiprilov, E.; Mirzabekov, T.; Choe, H. The role of post-translational modifications of the CXCR4 amino terminus in stromal-derived factor 1 α association and HIV-1 entry. *J. Biol. Chem.*, **2002**, *277*(33), 29484-29489. [http://dx.doi.org/10.1074/jbc.M203361200] [PMID: 12034737]

[5] Huttner, W.B. Protein tyrosine sulfation. *Trends Biochem. Sci.*, **1987**, *12*, 361-363. [http://dx.doi.org/10.1016/0968-0004(87)90166-6]

[6] Moore, K.L. The biology and enzymology of protein tyrosine O-sulfation. *J. Biol. Chem.*, **2003**, *278*(27), 24243-24246. [http://dx.doi.org/10.1074/jbc.R300008200] [PMID: 12730193]

[7] Yu, Y.; Hoffhines, A.J.; Moore, K.L.; Leary, J.A. Determination of the sites of tyrosine O-sulfation in peptides and proteins. *Nat. Methods*, **2007**, *4*(7), 583-588. [http://dx.doi.org/10.1038/nmeth1056] [PMID: 17558413]

[8] Zhang, Y.; Jiang, H.; Go, E.P.; Desaire, H. Distinguishing phosphorylation and sulfation in carbohydrates and glycoproteins using ion-pairing and mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **2006**, *17*(9), 1282-1288. [http://dx.doi.org/10.1016/j.jasms.2006.05.013] [PMID: 16820302]

[9] Kehoe, J.W.; Bertozzi, C.R. Tyrosine sulfation: A modulator of extracellular protein-protein interactions. *Chem. Biol.*, **2000**, *7*(3), R57-R61. [http://dx.doi.org/10.1016/S1074-5521(00)00093-4] [PMID: 10712936]

[10] Önnerfjord, P.; Heathfield, T.F.; Heinegård, D. Identification of tyrosine sulfation in extracellular leucine-rich repeat proteins using mass spectrometry. *J. Biol. Chem.*, **2004**, *279*(1), 26-33. [http://dx.doi.org/10.1074/jbc.M308689200] [PMID: 14551184]

[11] 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]

[12] 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]

[13] 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]

[14] 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]

[15] Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K-C. iRNA-PseU: Iden-

tifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids*, **2016**, *5,* e332.
[PMID: 28427142]

[16] 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]

[17] 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]

[18] 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]

[19] 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]

[20] 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]

[21] 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]

[22] 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]

[23] 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]

[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] 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]

[27] 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]

[28] 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]

[29] 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]

[30] 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]

[31] 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]

[32] 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]

[33] 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]

[34] 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]

[35] 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]

[36] 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*.

[37] 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]

[38] 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]

[39] 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]

[40] 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]

[41] 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]

[42] 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. *PeerJ*, **2013**, *1,* e171.
[http://dx.doi.org/10.7717/peerj.171] [PMID: 24109555]

[43] 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]

[44] 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]

[45] 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]

[46] 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]

[47] 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]

[48] 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]

[49] 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]

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

[51] 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]

[52] 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]

[53] Awais, M.; Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. 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. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **2019**.

[54] 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]

[55] Chen, Z.; Liu, X.; Li, F.; Li, C.; Marquez-Lago, T.; Leier, A.; Akutsu, T.; Webb, G. I.; Xu, D.; Smith, A. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform.,* **2018**, 30285084.
[http://dx.doi.org/10.1093/bib/bby089]

[56] 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]

[57] 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]

[58] 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. **2019**, *20*(1), 112.
[http://dx.doi.org/10.1186/s12859-019-2700-1]

[59] 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]

[60] Wang, L.; Zhang, R.; Mu, Y. Fu-SulfPred: Identification of protein s-sulfenylation sites by fusing forests *via* Chou's general PseAAC. **2019**, *461*, 51-58.

[61] Xie, H.-L.; Fu, L.; Nie, X.-D. J.; Design, P.E. 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.

[62] 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]

[63] Yu, K.M.; Liu, J.; Moy, R.; Lin, H.C.; Nicholas, H.B., Jr; Rosenquist, G.L. Prediction of tyrosine sulfation in seven-transmembrane peptide receptors. *Endocrine*, **2002**, *19*(3), 333-338.
[http://dx.doi.org/10.1385/ENDO:19:3:333] [PMID: 12624435]

[64] Monigatti, F.; Gasteiger, E.; Bairoch, A.; Jung, E. The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*, **2002**, *18*(5), 769-770.
[http://dx.doi.org/10.1093/bioinformatics/18.5.769] [PMID: 12050077]

[65] Chang, W.C.; Lee, T.Y.; Shien, D.M.; Hsu, J.B.K.; Horng, J.T.; Hsu, P.C.; Wang, T.Y.; Huang, H.D.; Pan, R.L. Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J. Comput. Chem.*, **2009**, *30*(15), 2526-2537.
[http://dx.doi.org/10.1002/jcc.21258] [PMID: 19373826]

[66] Niu, S.; Huang, T.; Feng, K.; Cai, Y.; Li, Y. Prediction of tyrosine sulfation with mRMR feature selection and analysis. *J. Proteome Res.*, **2010**, *9*(12), 6490-6497.
[http://dx.doi.org/10.1021/pr1007152] [PMID: 20973568]

[67] Huang, S-Y.; Shi, S-P.; Qiu, J-D.; Sun, X-Y.; Suo, S-B.; Liang, R-P. PredSulSite: Prediction of protein tyrosine sulfation sites with multiple features and analysis. *Anal. Biochem.*, **2012**, *428*(1), 16-23.
[http://dx.doi.org/10.1016/j.ab.2012.06.003] [PMID: 22691961]

[68] Jia, C.; Zhang, Y.; Wang, Z. SulfoTyrP: A high accuracy predictor of protein sulfotyrosine sites. *Match Commun. Math. Comput. Chem*, **2014**, *71*, 227-240.

[69] 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]

[70] 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]

[71] 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]

[72] 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]

[73] 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**, *111*(4), 886-892.
[PMID: 29842950]

[74] 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]

[75] 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]

[76] 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]

[77] 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]

[78] 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]

[79] 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.

[80] Chen, W.; Feng, P.-M.; Lin, H.; Chou, K.-C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **2013**, *41*(6), e68.

[81] 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]

[82] 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.

[83] Ding, H.; Deng, E.-Z.; Yuan, L.-F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.-C. A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. **2014**, *2014.* 1-10.

[84] 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]

[85] 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]

[86] 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]

[87] Khan, Y.D.; Batool, A.; Rasool, N.; Khan, S.A.; Chou, K-C. Prediction of nitrosocysteine sites using position and composition variant features. *Lett. Organic Chem.,* **2019**, *16*(4), 283-293.

[88] 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.

[89] 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]

[90] Liu, B.; Fang, L.; Long, R.; Lan, X.; Chou, K.-C. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, **2015**, *32*(3), 362-369.

[91] 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. **2015**, *385*, 153-159. [http://dx.doi.org/10.1016/j.jtbi.2015.08.025]

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

[93] 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. *Lett. Organic Chem.,* **2019**, *16*(4), 303-310.

[94] 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. **2015**, *33*(10), 2221-2233.

[95] Chou, K.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] [PMID: 31060481]

[96] Zhang, C.T.; Chou, K.C. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.,* **1992**, *1*(3), 401-408. [http://dx.doi.org/10.1002/pro.5560010312]

[97] Chou, K.C.; Cai, Y.D. Prediction and classification of protein subcellular location-sequenceorder effect and pseudo amino acid composition. *J. Cell Biochem.,* **2003**, *90*(6), 1250-1260.

[98] Chou, K.-C.; Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res*., **2002**, *1*(5), 429-433. [http://dx.doi.org/10.1021/pr025527k]

[99] Hu, L.; Huang, T.; Shi, X.; Lu, W.-C.; Cai, Y.-D.; Chou, K.-C. Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. **2011**, *6*(1), e14556. [http://dx.doi.org/10.1371/journal.pone.0014556]

[100] Cai, Y.-D.; Feng, K.-Y.; Lu, W.-C.; Chou, K.-C. Using LogitBoost classifier to predict protein structural classes. **2006**, *238*(1), 172-176. [http://dx.doi.org/10.1016/j.jtbi.2005.05.034]

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

[102] Ahmad, J.; Hayat, M. MFSC: Multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. *J. Theor. Biol.,* **2019**, *463*, 99-109.

[103] Akbar, S.; Hayat, M. iMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. *J. Theor. Biol.,* **2018**, *455*, 205-211.

[104] Behbahani, M.; Mohabatkar, H.; Nosrati, M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. *J. Theor. Biol.,* **2016**, *411*, 1-5. [http://dx.doi.org/10.1016/j.jtbi.2016.09.001]

[105] 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]

[106] Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.; Sattar, A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou' s general PseAAC. *J. Theor. Biol.,* **2015**, *364*, 284-294.

[107] Ju, Z.; He, J.-J. Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. *J. Mol. Graph Model*., **2017**, *76*, 356-363.

[108] Kabir, M.; Hayat, M. iRSpot-GAEnsC: Identifing recombination spots *via* ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. *Mol. Genet. Genomics,* **2016**, *291*(1), 285-296.

[109] Meher, P. K.; Sahu, T. K.; Saini, V.; Rao, A. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.,* **2017**, *7*, 42362. [http://dx.doi.org/10.1038/srep42362]

[110] Tahir, M.; Hayat, M.; Khan, S. iNuc-ext-PseTNC: An efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition*., Mol. Genet. Genomics.* **2019**, *294*(1), 199-210.

[111] Yu, B.; Li, S.; Qiu, W.-Y.; Chen, C.; Chen, R.-X.; Wang, L.; Wang, M.-H.; Zhang, Y. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. *Oncotarget*, **2017**, *8*(64), 107640.

[112] Zhang, S.; Liang, Y. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. *J. Theor. Biol.,* **2018**, *457*, 163-169. [http://dx.doi.org/10.1016/j.jtbi.2018.08.042]

[113] Chou, K.-C. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem*., **2017**, *457*, 163-169. [http://dx.doi.org/10.2174/1568026617666170414145508]

[114] Shen, H.-B.; Chou, K.-C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.,* **2008**, *373*(2), 386-388. [http://dx.doi.org/10.1016/j.ab.2007.10.012]

[115] Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem*., **2012**, *425*(2), 117-119. [http://dx.doi.org/10.1016/j.ab.2012.03.015]

[116] Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z. J. B. Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics,* **2013**, *29*(7),

960-962.
[117] Du, P.; Gu, S.; Jiao, Y. PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci*., **2014**, *15*(3), 3495-3506.
[http://dx.doi.org/10.3390/ijms15033495]
[118] Chou, K.-C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics,* **2009**, *6*(4), 262-274.
[http://dx.doi.org/10.2174/157016409789973707]
[119] Chen, W.; Lei, T-Y.; Jin, D-C.; Lin, H.; Chou, K-C. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.*, **2014**, *456*, 53-60.
[http://dx.doi.org/10.1016/j.ab.2014.04.001] [PMID: 24732113]
[120] Chen, W.; Lin, H.; Chou, K.-C. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Mol. Biosyst*., **2015**, *11*(10), 2620-2634.
[http://dx.doi.org/10.1039/C5MB00155B]
[121] 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]
[122] Tahir, M.; Tayara, H.; Chong, K. iRNA-PseKNC (2methyl): Identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components. *J. Theoretical Biol.,* **2019**, *465*, 1-6.
[123] Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*., **2015**, *43*(W1), W65-W71.
[124] Liu, B.; Wu, H.; Chou, K.-C. J. N. S. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*., **2017**, *9*(04), 67.
[125] 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]
[126] 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.
[127] 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]
[128] 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. *The Sci. World J.*, **2014**, *2014,* 1-11.
[http://dx.doi.org/10.1155/2014/875879]
[129] Khan, Y.D.; Khan, S.A.; Ahmad, F.; Islam, S. Iris recognition using image moments and k-means algorithm. *The Sci. World J.*, **2014**, *2014,* 1-9.
[http://dx.doi.org/10.1155/2014/723595]
[130] Chou, K.-C. Prediction of signal peptides using scaled window. *Peptides,* **2001**, *22*(12), 1973-1979.
[http://dx.doi.org/10.1016/S0196-9781(01)00540-X]
[131] Chou, K.C. Bioinformatics, Prediction of protein signal sequences and their cleavage sites. *Proteins*, **2001**, *42*(1), 136-139.
[132] Chou, K.-C. *Prediction of signal peptides using scaled window. Peptides.* **2001**, *22*(12), 1973-1979.
[133] 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]
[134] 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]
[135] 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]
[136] 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]
[137] 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]
[138] 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]
[139] 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]
[140] 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]
[141] 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]
[142] 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]
[143] 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]
[144] 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]
[145] 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]
[146] 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]
[147] 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]
[148] 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]
[149] 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]
[150] 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.
[151] 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.
[152] 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.

[153] 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]

[154] 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.

[155] 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]

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

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

[158] 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]

[159] 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]

[160] 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]

[161] 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]

[162] 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. [177]. [http://dx.doi.org/10.1016/j.ab.2013.01.019] [PMID: 23395824]

[163] 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]

[164] 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]

[165] 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]

[166] 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]

[167] 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]

[168] 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]

[169] 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]

[170] 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*(09), 330. [http://dx.doi.org/10.4236/ns.2017.99032]

[171] 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]

[172] 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*(5), 58494-346.

[173] 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]

[174] Chou, K-C.; Zhang, C-T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*(4), 275-349. [http://dx.doi.org/10.3109/10409239509083488] [PMID: 7587280]

[175] Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.; Sattar, A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.*, **2015**, *364*, 284-294. [http://dx.doi.org/10.1016/j.jtbi.2014.09.029] [PMID: 25264267]

[176] Dou, Y.; Yao, B.; Zhang, C.; Phospho, S.V.M. PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids*, **2014**, *46*(6), 1459-1469. [http://dx.doi.org/10.1007/s00726-014-1711-5] [PMID: 24623121]

[177] Feng, K-Y.; Cai, Y-D.; Chou, K-C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.*, **2005**, *334*(1), 213-217. [http://dx.doi.org/10.1016/j.bbrc.2005.06.075] [PMID: 15993842]

[178] Kumar, R.; Srivastava, A.; Kumari, B.; Kumar, M. Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **2015**, *365*, 96-103. [http://dx.doi.org/10.1016/j.jtbi.2014.10.008] [PMID: 25454009]

[179] Mondal, S.; Pai, P.P. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.*, **2014**, *356*, 30-35. [http://dx.doi.org/10.1016/j.jtbi.2014.04.006] [PMID: 24732262]

[180] Nanni, L.; Brahnam, S.; Lumini, A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2014**, *360*, 109-116. [http://dx.doi.org/10.1016/j.jtbi.2014.07.003] [PMID: 25026218]

[181] 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. [http://dx.doi.org/10.3390/ijms15021746] [PMID: 24469313]

[182] Shen, H-B.; Yang, J.; Chou, K-C. Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **2007**, *33*(1), 57-67. [http://dx.doi.org/10.1007/s00726-006-0478-8] [PMID: 17235453]

[183] Wu, Z-C.; Xiao, X.; Chou, K-C. iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.*, **2011**, *7*(12), 3287-3297. [http://dx.doi.org/10.1039/c1mb05232b] [PMID: 21984117]

[184] Zhou, G.P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins*, **2003**, *50*(1), 44-48. [http://dx.doi.org/10.1002/prot.10251] [PMID: 12471598]

[185] 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.

[186] 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. Kinetic studies with the non-nucleoside human immunodeficiency 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]

[187] 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.

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

[189] Chou, K.-C.; Forsén, S. Graphical rules for enzyme-catalysed rate laws. *Biochem. J.*, **1980**, *187*(3), 829-835. [http://dx.doi.org/10.1042/bj1870829]

[190] Chou, K.-C.; Lin, W.-Z.; Xiao, X. Wenxiang: A web-server for drawing wenxiang diagrams. *Nat. Sci*., **2011**, *3*(10), 862. [http://dx.doi.org/10.4236/ns.2011.310111]

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

[192] 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.

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

[194] Chou, K. Graph theory of enzyme kinetics. *Scientia Sinica*, **1979**, *22*, 341-358.

[195] Chen, K.-C.; Carter, R.E.; Forsen, S. A new graphical-method for deriving rate-equations for complicated mechanisms. *Chemica Scripta,* **1981**, *18*(2), 82-86.

[196] 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]

[197] Zhou, G.; Deng, M. J. B. J. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. *Biochem. J.,* **1984**, *222*(1), 169-176. [http://dx.doi.org/10.1042/bj2220169]

[198] 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]

[199] 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]

[200] 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]

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

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

[203] Chou, K.-C.; Shen, H.-B. Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci*., **2009**, *1*(02), 63. [http://dx.doi.org/10.4236/ns.2009.12011]

[204] 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]

[205] 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.

[206] 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]