RESEARCH ARTICLE

iCrotoK-PseAAC: Identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule

Sharaf Jameel Malebary[1], **Muhammad Safi ur Rehman**[2], **Yaser Daanial Khan**[2]*

1 Department of Information Technology, King Abdul Aziz University, Rabigh, Kingdom of Saudi Arabia, **2** Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan

* yaser.khan@umt.edu.pk

Abstract

Among different post-translational modifications (PTMs), one of the most important one is the lysine crotonylation in proteins. Its importance cannot be undermined related to different diseases and essential biological practice. The key step for finding the hidden mechanisms of crotonylation along with their occurrence sites is to completely apprehend the mechanism behind this biological process. In previously reported studies, researchers have used different techniques, like position weighted matrix (PWM), support vector machine (SVM), k nearest neighbors (KNN), and many others. However, the maximum prediction accuracy achieved was not such high. To address this, herein, we propose an improved predictor for lysine crotonylation sites named iCrotoK-PseAAC, in which we have incorporated various position and composition relative features along with statistical moments into PseAAC. The results of self-consistency testing were 100% accurate, while the 10-fold cross validation gave 99.0% accuracy. Based on the validation and comparison of model, it is concluded that the iCrotoK-PseAAC is more accurate than the previously proposed models.

OPEN ACCESS

Citation: Malebary SJ, Rehman MSu, Khan YD (2019) iCrotoK-PseAAC: Identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. PLoS ONE 14(11): e0223993. https://doi.org/10.1371/journal.pone.0223993

Editor: Frederique Lisacek, Swiss Institute of Bioinformatics, SWITZERLAND

Received: July 8, 2019

Accepted: October 2, 2019

Published: November 21, 2019

Data Availability Statement: All relevant data are within the manuscript within its Supporting Information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

In all the living organism, the cells contain the chromosomes having the stored information, which deals with the normal body functions. Specifically, chromosomes contain DNA (Deoxyribonucleic acid), the polymer of deoxyribonucleotides [1]. Almost 80–90% of the DNA is considered as junk, whose exact functioning is not identified yet; but for the rest, it has been discovered. Although, the complexity of the Genomic sequences within the chromosomes follows highly patterned and efficiently organized methodology, the main function of the DNA is to replicate and encode the body proteins [2]. These body proteins are formed by unwinding of the DNA by semi-conservative methods. Both strands of DNA encode its signals to the Messenger RNA (mRNA) by the process of transcription. It is followed by the translation, in which protein sequence is transmitted to the transfer ribonucleic acid (tRNA). At the end, the tRNA terminates the process by joining amino acids over the ribosome (organelle) and forming protein exactly according to the gene sequence. After this translation, post-translational modification can take place, which can either activate or inactivate the function of that protein [3].

Crotonylation is a reversible post-translational modification process, which usually takes place over lysine [4] residues of a protein. It holds immense importance, with respect to its various effects on the body's metabolism, genetic expressions and multiple diseases i.e. carcinomas and malignancies etc. [5]. Lysine is an essential amino acid with a basic side chain and a product of putrefaction in the gut. The chemical structure of the lysine contains a carboxylic group, amino group, hydrogen and a basic R group, attached to the central carbon of the amino acid backbone [6].

Up till now, a couple of computational based techniques have been proposed for the prediction of lysine crotonylation sites in proteins. According to a theory that both crotonylated peptides and non-crotonylated peptides are produced by particular mechanisms, Huang and Zeng [7] proposed an automated predictor, named CrotPred, for the prediction of histone crotonylation sites in proteins. Later on, Qiu et al. [8] provided a mechanism for the identification of lysine crotonylation sites, which utilized position weight amino acid composition, to identify CrotoK sites using a support vector machine (SVM) algorithm. Notwithstanding, the prescient accuracies of CrotPred and Qiu's technique achieved just 79.41% and 71.69%, individually. The prediction accuracy of the over two techniques is as yet not much high. It ought to likewise be noticed that the over two techniques [8] did not provide good accuracy results. Later on, Ju et al. proposed CKSAAP_CrotSite to identify lysine crotonylation sites with an accuracy of 98.11% [9] and Qiu et al. proposed iKcr-PseEns to identify lysine crotonylation sites in histone proteins with an accuracy of 94.49% [10].

The efficiency of the previous research processes lacks the relative positioning and composition information which holds immense importance, with two basic goals; to make the relevant theoretical study and to give scientist an easier layout for research purpose. In order to achieve these goals, a 5-steps rule [11–18] should be employed, as used by the researchers in the past [13, 19–22]. These steps include (i) collection of a standardized dataset, (ii) mathematical formulation of features and association with biological target classes, (iii) training of predictor via operational algorithm, (iv) objective evaluation and validation of results, and (v) implementation of webserver for proposed predictor. Although the credibility of the methods has been improved successively still its upper limit could touch the satisfactory value. We, herein, propose a predictor named iCrotoK-PseAAC which aims to identify CrotoK sites with improved efficiency than previously reported methods, using PseAAC and Chou's 5-steps rule.

2. Materials and methods

In the following section, the primary three portions of the 5-steps rule are being discussed. The methodology is being presented in the flowchart in Fig 1.

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein as shown by the eight master pieces of pioneering papers from the then Chairman of Nobel Prize Committee Sture Forsen [23, 24], and many follow-up papers [25–29]. They are very useful for in-depth investigation into the topic of the current paper, and we will use them in our future efforts.

2.1. Benchmark dataset

Chou's peptide formulation [30] was used for an accurate elaboration of samples, which had generally utilized in computational biology and bioinformatics; for example, in predicting the signal peptide cleavage sites [31], nitrotyrosine sites [32], methylation sites [33], hydroxylysine and hydroxyproline sites [34], lysine ubiquitination sites [35], protein-protein binding sites

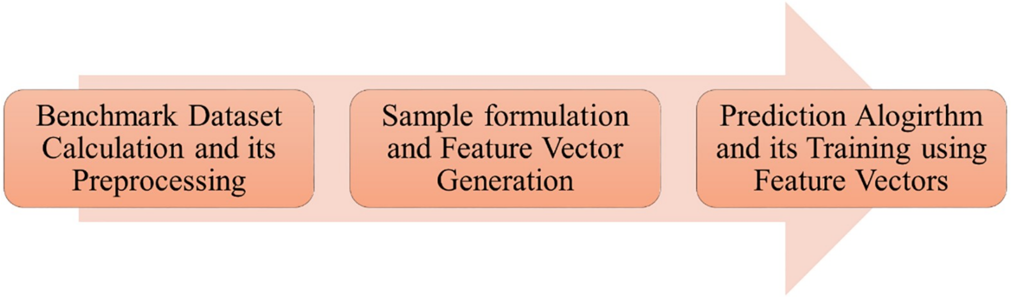
Benchmark Dataset Calculation and its Preprocessing
Sample formulation and Feature Vector Generation
Prediction Alogirthm and its Training using Feature Vectors

Fig 1. Three steps of methodology.

https://doi.org/10.1371/journal.pone.0223993.g001

[36], phosphorylation sites [37, 38], lysine succinylation sites [39], and, numerous lysine PTM sites [40].

Collection of benchmark dataset was performed from Universal Resource of Protein (UniProt). A query was applied to database using advanced search options and the sequences with PTM/Processing annotations were chosen. In modified residue field, 'crotonyllysine' phrase for positive samples was used. For negative data, a converse query was used. Thus, by these searches, 2665 positive and 868 negative datasets were filtered out. Redundancy and duplication from dataset was removed by using CD-HIT webserver [41] with a threshold of 0.6. The reduced dataset was used to extract sites from each sequence of positive and negative set, and a subsequence of length 41 was extracted for all positive and negative sites. Experiment was repeated for other lengths, i.e. 21 and 61, however, best results were achieved for 41. Before proceeding to the next step, we have removed the unnecessary information like some of the special characters and notations (B, J, O, U, X and Z) and spaces from the dataset of positive and negative samples of protein. Thus, a total of 378 positive and 500 negative samples were achieved.

Stated by Chou's scheme [30], the CrotoK site-having peptide sample can be usually defined as

$$\boldsymbol{P}_{\xi}(\mathbb{K}) = R_{-\xi}R_{-(\xi-1)}\cdots R_{-2}R_{-1}\mathbb{K}R_{+2}\cdots R_{+(\xi-1)}R_{+\xi} \tag{1}$$

Importance of amino acid $\mathbb{K}$ is shown by emphasizing the symbol using double strike, the subscript ξ is a number ranging [-ξ, +ξ] which include both positive and negative numbers excluding 0, $R_{-\xi}$ gives the -ξ-th upstream amino acid between the range, the $R_{+\xi}$ the +ξ-th gives the downstream amino acid other than the $R_{-\xi}$, etc. The (2ξ + 1)-tuple peptide sample $\mathbf{P}_{\xi}(\mathbb{K})$ can be more arranged in further two classes as shown in Eq (2).

$$\boldsymbol{P}_{\xi}(\mathbb{K}) \in \begin{cases} \boldsymbol{P}_{\xi}^{+}(\mathbb{K}),\ for\ the\ CrotoK \\ \boldsymbol{P}_{\xi}^{-}(\mathbb{K}),\ for\ non-CrotoK \end{cases} \tag{2}$$

Where $\mathbf{P}_{\xi}^{+}(\mathbb{K})$ indicate the positive CrotoK sites with $\mathbb{K}$ at mid, while $\mathbf{P}_{\xi}^{-}(\mathbb{K})$ indicate the non-CrotoK sites, and the mark ∈ means "a member of" in a group of sets.

Data-collection for benchmarks was based on two types of the dataset, one was used for training which we called training dataset and the other dataset was used for testing which we called testing dataset. There is no compelling reason to isolate a benchmark dataset into two datasets if you have used either of jackknife or subsampling using k-fold cross validation (where k = 5 or k = 10) because the results were generated using different folds [13, 21]. In this research, the value of ξ was 20 which was then put in the equation (2ξ+1) which equals to 41 characters in the subsequence using Eq (2). As needs are, the benchmark dataset for this investigation can be accumulated into $\mathbb{S}$ as shown in Eq (3).

Where $\mathbb{S}^{+}$ have 378 positive samples, $\mathbb{S}^{-}$ have 500 negative samples while ∪ is denoting "union" in a group of hypotheses. As per user's convenience, the 378 + 500 = 878 example sequences are given in S1 File [42].

$$\mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \tag{3}$$

2.2. Sample formulation

Due to the recent advancement in the biological computation because of the increase in the biological sequence, one of the popular problems is to compute the biological sequences using a discrete model or vector by the keeping the actual sequence intact.

Since all the conventional machine learning algorithms deal with the vectors and not directly with the sequence samples [43], so pseudo amino acid composition or PseAAC [44] was then created to solve such problems. With the creation of PseAAC or Chou's PseAAC [45, 46], the field of biomedicine and drug development industries started to use it [47, 48] and almost every area of study in the field of computational proteomics [49–51] took interest in it. Due to its rapid and sudden popularity, many open source software [45, 52] were developed to generate the various sequences of the PseAAC. Due to the increasing popularity of the PseAAC, several webservers [53] were created to generate different DNA/RNA sequence and has proven itself to be very useful in the field of computational genomes [54]. Pse-in-One [55] was developed to generate any number of desired feature vectors according to the user's focuses for both DNA/RNA and protein/peptide sequences.

As indicated by Eq (3) and formulated in PseAAC, the sample sequences of the peptide in S1 File are extracted as suggested in [30]

$$P_{\text{ş}=7}(\mathbb{K}) = [\breve{\mathsf{A}}_1 \breve{\mathsf{A}}_2 \cdots \breve{\mathsf{A}}_{\Omega}]^{T} \tag{4}$$

Where $\breve{A}_1\breve{A}_2\cdots\breve{A}_{\Omega}$ will be characterized by how to extricate valuable features from samples of a peptide sequence, and T is the transpose administrator.

The UniProt database provides number of ways to extract the required data. During this study the data was extracted by parsing XML files generated by UniProt as a result of query that search for all proteins containing the required modified residue. Different values of **Ω** were tried in order to find the most optimal one. Extensive trials of experimentation and probing showed that **Ω** = 20 yielded best results. Hence the benchmark data was formed using 20 residues at either terminals of the crotonylation site. Subsequently, the benchmark dataset contains peptide sequences of length 41 as given in S1 File can be rearranged to the accompanying simpler to deal with frame and their reduction is shown in Eq (5).

$$\boldsymbol{P} = R_1, R_2, R_3 \cdots R_{19}, R_{20}, R_{21} \cdots R_{39}, R_{40}, R_{41} \tag{5}$$

Where R_{21} is the modified residue K and R_i for *(i = 1, 2, . . . 41; i ≠ 21)* can be any of the 20 native amino acids or the dummy code X as discussed before. For further clarification, 20 native amino acids were given numerical values ranging from 1 to 20 with respect to the alphabetic order of their letter code and remaining 21 dummy amino acids were given the numeric values ranging from 21 to 41.

2.2.1. Site vicinity vector (SVV). There are various remaining sites which are similar to Post-Translational Modification (PTM) in the chain of polypeptide and there exist many characteristics that channel us towards the modification. Other than the previously discussed characteristics, the other neighboring residues of a site are also important in which PTM is observed [56]. The sequence which contains the potential PTM residue site from the primary sequence of the protein is known as Site Vicinity Vector (SVV).

Following is the primary sequence having the potential PTM site with its surrounding residues and in Eq (6) it is denoted as

$$P = [\mu_1 \ldots \mu_{x-2}, \mu_{x-1}, \mu_x, \mu_{x+1}, \mu_{x+2}, \ldots..\mu_n] \tag{6}$$

Sub-sequence of the primary sequence is SVV which is represented as

$$S = [\mu_{\mathrm{x-k}} \ldots \mu_{\mathrm{x-2}}, \mu_{\mathrm{x-1}}, \mu_{\mathrm{x}}, \mu_{\mathrm{x+1}}, \mu_{\mathrm{x+2}}, \ldots..\mu_{\mathrm{x+k}}] \tag{7}$$

After thorough experiments, the result shows that the minimum constant value best suited for k is 20 in our specific scenario. Each characteristic in SVV shows the specific amino acid already known, out of given 20 amino acids in which a unique numerical value are assigned to each amino acid ranging from 1 to 20.

2.2.2. Calculation of statistical moments (SM). Statistical moment approach is utilized to consider the quantitative explanation for benchmark dataset of protein sample sequence to define the dimensions and components of Eq (6). To keep the crucial information safe regarding the protein samples' sequence order, as it is necessary that the benchmark dataset of protein samples have the amino acid residues to be in a specific order. Different types of factors were considered while gathering the information of protein samples using different variables such as order of moments in which some of the moments for data evaluation and some for peculiarity of data and some for orientation from enormous amount of gathered protein samples and some of them were elaborated by the mathematicians using distributions and polynomial functions [13, 21, 37, 38, 56–65].

We have calculated Hahn moments by using Hahn polynomial, raw moments by using probability distribution of benchmarked dataset of protein samples and central moments using mean, variance and asymmetry of information and these are computed for iCrotoK -PseAAC predications. Their results are scale variant [63], whereas the central moment is both scale variant and vicinity variant [63, 65] and the results computed in raw moments of mean, variance and asymmetry are then further used in the calculations of the central moment.

Moments of the scale variants are not used, and each method describes the qualified values [66, 67] of the data in 2 dimension (2D) square matrix P (in which rows and columns are same and denoted as m) with dimension *m x m* as written in Eq (8) and accommodate all of the protein samples residues in our proposed methodology.

$$\begin{bmatrix} P_{11} & P_{12}\ldots & P_{1m} \\ P_{12} & P_{22}\ldots & P_{2m} \\ \vdots & \vdots & \vdots \\ P_{m1} & P_{m2}\ldots & P_{mm} \end{bmatrix} \tag{8}$$

Now the generated square matrix ***P*** from Eq (8) is passed to the function *ω* which transformed that ***P*** matrix into ***P'***. This can be mathematically written as $P' = \omega(P)$) with the 3rd degree of statistical moments (means rows and columns should not exceed 3 from the moments matrix P) and then the newly generated matrix ***P'*** is used by Hahn moment, which speeds up our calculation. The property of ***P'*** being a square matrix made Hahn moment orthogonal in which the inverse propriety was preserved and could be converted back to ***P*** using the inverse function of discrete Hahn moment as shown in Eq (9).

$$H_n^{u,z}(r,M) = (M+Z-1)_n(M-1)_n \times \sum\nolimits_{i=0}^{n} (-1)^i \frac{(-n)_i(-r)_i(2M+u+z-n-1)_i}{(M+z-1)_i(M-1)_i} \times \frac{1}{\bar{i}!} \quad (9)$$

The pochhammer symbol and the gamma operator is applied in (9) are described in [56, 67]. The normalized orthogonal Hahn moments calculation was performed using Eq (10)

$$h_{rs} = \sum\nolimits_{a=1}^{M-1} \sum\nolimits_{b=1}^{M-1} \partial_{rs} H_r^{\widetilde{u,z}}(b,M) H_s^{\widetilde{u,z}}(a,M), n = 0,\ 1,\ 2,\ 3\ldots M-1 \quad (10)$$

Highly imported information is conserved by central moments which is related to the variance, asymmetry and means of the sample proteins and its calculation is carried out using Eq (11).

$$\Im_{rs} = \sum\nolimits_{a=1}^{k} \sum\nolimits_{b=1}^{k} (a-\bar{x})^r (a-\bar{y})^s \partial_{ab} \quad (11)$$

In the end, benchmark dataset of samples protein's information is gathered using the calculation of raw moments through distribution probability as shown in Eq (12)

$$M_{rs} = \sum\nolimits_{a=1}^{k} \sum\nolimits_{b=1}^{k} a^r b^s \partial_{ab} \quad (12)$$

Raw moment's degree is *r+s* and M_{00}, M_{01}, M_{02}, M_{10}, M_{11}, M_{20}, M_{12}, M_{21}, M_{03} *and* M_{30} of 3rd degree is mediated here.

2.2.3. Calculations of position relative incidence matrix (PRIM). The primary sequence plays a vital role in calculating the hidden patterns from the sequence of the protein. The positional information of the protein benchmarked dataset and the residues in it play a key role in central mathematical paradigm of the model and 20x20 matrix is formed by using quantization of relative positional information of the residues in the amino acid and we called it H_{PRIM}.

$$\mathrm{H}_{\mathrm{PRIM}} = \begin{bmatrix} \mathrm{H}_{1\to1} & \mathrm{H}_{1\to2}\cdots & \mathrm{H}_{1\to\mathrm{j}}\cdots & \mathrm{H}_{1\to1} \\ \mathrm{H}_{2\to1} & \mathrm{H}_{2\to2}\cdots & \mathrm{H}_{2\to\mathrm{j}}\cdots & \mathrm{H}_{2\to20} \\ \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to1}} & \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to2}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to\mathrm{j}}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to20}} \\ \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to1}} & \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to2}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to\mathrm{j}}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to20}} \end{bmatrix} \quad (13)$$

In Eq (13) *i* represents the specific row number and *j* represents the specific column which contains the sum of the first appearance of the i^{th} residue with respect to the j^{th} relative position and it generates 400 coefficients which then reduced to 30 generated coefficients by using the statistical moments [56].

2.2.4. Calculations of reverse position relative incidence matrix (RPRIM). The H_{RPRIM} is calculated using the reverse of protein samples instead of using the actual ones which we have used in H_{PRIM}, so that we can be able to find out the ambiguities in the samples of proteins, the matrix is shown in Eq (14).

$$\mathrm{H}_{\mathrm{RPRIM}} = \begin{bmatrix} \mathrm{H}_{1\to1} & \mathrm{H}_{1\to2}\cdots & \mathrm{H}_{1\to\mathrm{j}}\cdots & \mathrm{H}_{1\to1} \\ \mathrm{H}_{2\to1} & \mathrm{H}_{2\to2}\cdots & \mathrm{H}_{2\to\mathrm{j}}\cdots & \mathrm{H}_{2\to20} \\ \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to1}} & \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to2}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to\mathrm{j}}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{i}\to20}} \\ \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to1}} & \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to2}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to\mathrm{j}}}\cdots & \mathrm{H}_{\substack{\vdots\\ \mathrm{N}\to20}} \end{bmatrix} \quad (14)$$

The H_{RPRIM} gives the same amount of dataset but generates the different values and generates the same number of coefficients as H_{PRIM} in Eq (14) does and which can be reduced the same way we reduced H_{PRIM} [13, 21, 56].

2.2.5. Determination of frequency vector (FV). The vector containing the frequencies of all the residues of the protein in the benchmark dataset and we called it frequency vector which has the property to sustain the compositional and distributional information regarding the samples of the protein sequence. FV is shown in Eq (15).

$$FV = [f_1, f_2, f_3, f_4, f_5, .., f_{20}] \tag{15}$$

Where each f_i holds the frequency by the alphabetic order of each residue of amino acid in the sequence.

2.2.6. Determination of accumulative absolute position incidence vector (AAPIV). The cumulative frequency distribution of the residues of amino acids in protein polypeptide chain related to the composition of the protein is known as accumulative absolute position incidence vector (AAPIV) and it only holds the information of absolute position thus the information regarding the relative position is lost. A vector of 20 elements is then formed (as shown in Eq (18) which formed based on primary structure given in Eq (16) where each element in this vector represents the sum of ordinal values of its occurrences of respective residue at $p_1, p_2, \ldots, p_n$ locations within the primary structure as shown in (17).

$$\alpha^i_{p1},\ \ldots,\ \alpha^i_{p2},\ \ldots,\ \alpha^i_{pn} \tag{16}$$

$$\mu_i = \sum\nolimits_{k=1}^{n} p_k \tag{17}$$

$$AAPIV = \{\mu_1.\mu_2, \ldots, \mu_{20}\} \tag{18}$$

2.2.7. Determination of reverse accumulative absolute position incidence vector (RAAPIV). Similarly, 20 elements of reverse accumulated absolute position incidence vector (RAAPIV) is formed (as shown in Eq (20)) by simply reversing the primary structure of the p^{th} residue (as shown in Eq (19)).

$$\alpha^i_{pn},\ \ldots,\ \alpha^i_{p(n-1)},\ \ldots,\ \alpha^i_{p1} \tag{19}$$

$$RAAPIV = \{\mu'_1.\mu'_2, \ldots, \mu'_{20}\} \tag{20}$$

2.3. Prediction model

Billions of neurons are present in the human brain, which works by taking in the information, processing it and then passing further for its functioning. Each time, patterns are being learned and information is extracted which is used to act against situations and things with no specificity. ANN (Artificial Neural Network) is a system, based on technique like the human brain. Information is being learned from various situations and patterns. Later on, it uses interconnect neurons to works in problem-solving. A number of neurons take the input and previous information, along with patterns are used. There are two working modes of ANN, the first one is the training mode, in which ANN is being trained on provided data and information is extracted from that set of data and learn. The second one is using or working, it's on working mode; like neurons provide input and best match for output is extracted using available information [68, 69] and taught patterns as shown in Fig 2.

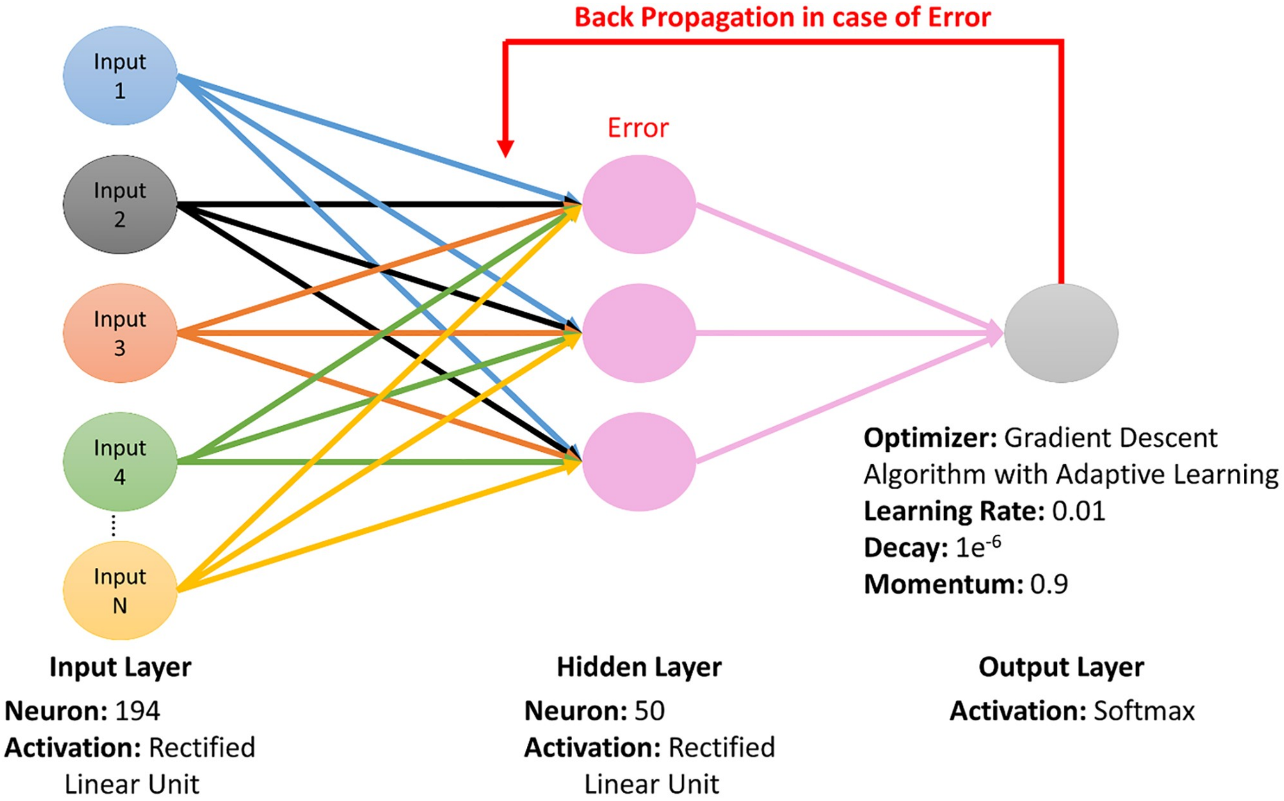
Back Propagation in case of Error
Error
Input 1
Input 2
Input 3
Input 4
Input N
Optimizer: Gradient Descent Algorithm with Adaptive Learning
Learning Rate: 0.01
Decay: 1e-6
Momentum: 0.9
Input Layer
Neuron: 194
Activation: Rectified Linear Unit
Hidden Layer
Neuron: 50
Activation: Rectified Linear Unit
Output Layer
Activation: Softmax

Fig 2. The architecture of ANN for the proposed prediction model.

https://doi.org/10.1371/journal.pone.0223993.g002

For this prediction model, ANN is implemented with back propagation methods, to reduce the error. For the specific and standard dataset, feature extraction was conducted and feature vector consisted of central, Hahn, and raw moments of sequence matrix, RPRIM and PRIM; FV, SVV, RAAPIV and AAPIV. The information linked to the position of proteins and representation was protected and saved in the form of final Feature Vector (FV) and these were total of 194. IFM (Input Feature Matrix) was made using all these FV and each row in IFM shows the FV related to every sample of the specific benchmark dataset. For each sample output, OM (the Output Matrix) was produced using all these output labels of input from IFM. ANN was being trained using both OM and IFM, OM to reduce the error by back propagation method and IFM as input [70, 71] as shown in Fig 2.

The neural Network was trained and tested using the Python version 3.4 along with web application framework Flask.

3. Results and discussion

3.1. Estimated accuracy

During the production of any new model, objective evaluation of its rate is one of the most important objectives. In order to justify the evaluation of the model, two questions must be catered. 1) What type of metrics should we use to review the quality of prediction model? 2) What are the various test methods that can be used for score metrics?

3.2. Formulation of metrics

Following are the four different evaluation metrics, used to evaluate the precision of the prediction model. (1) Acc to estimate the prediction model's overall accuracy, (2) S_n for the sensitivity of the prediction model, (3) S_p for the specificity of the prediction model, (4) MCC for the stability of the prediction model to access the accuracy and quality of the prediction model, various conventional metrics which have been used frequently in literature, are no more beneficial, because of lack in insightfulness and difficulty faced by the biologist in understanding it. Particularly for the Mathew's Correlation Coefficient (MCC), which is highly significant in illustrating the stability of the prediction model. Conveniently, various symbols have been an introduction to study protein signal peptides by Chou [72], A set of four intuitive equation were derived [13, 21, 73] as follows in Eq (21).

$$\begin{cases} Sensitivity\,(S_n) = 1 - \dfrac{N_-^+}{N^+} & 0\ \le S_n \le 1 \\ Specificity\,(S_p) = 1 - \dfrac{N_+^-}{N^-} & 0 \le S_p \le 1 \\ Accuracy\,(Acc) = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le Acc \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le MCC \le 1 \end{cases} \tag{21}$$

N_+^- illustrates the number of the incorrectly predicted non-crotonyllysine sites as crotonyllysine sites. N^- represents the number of the correctly predicted non- crotonyllysine sites. Furthermore, N^+ presents the number of the correctly predicted crotonyllysine sites. N_-^+ represents the number of the incorrectly predicted crotonyllysine sites as non- crotonyllysine sites.

From Eq (21), it can be seen that when $N_-^+ = 0$ it means that $S_n = 1$, because not a single crotonyllysine site is predicted as the non- crotonyllysine sites. If $N_-^+ = N^+$, then $S_n = 0$, as all the crotonyllysine sites are incorrectly predicted as non- crotonyllysine sites. Additionally, if $N_+^- = 0$, then the specificity is Sp = 1, which means none of the single non- crotonyllysine site is incorrectly predicted as a crotonyllysine site; while if we have $N_+^- = N^-$, then the specificity is Sp = 0 as all the non- crotonyllysine sites. If $N_-^+ = N_+^- = 0$, it shows that not a single non-crotonyllysine site in the negative dataset and crotonyllysine site in the positive dataset incorrectly predicted, and it provides us with the Acc = 1 and MCC = 1; if we have $N_+^- = N^-$, and $N_-^+ = N^+$, it shows that all the non-crotonyllysine sites in the negative dataset and all the crotonyllysine sites in the positive dataset are incorrectly predicted, and it provides us with the Acc = 0 and MCC = –1. Whereas, if $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ then it will provide us with the Acc = 0.5 and MCC = 0, leaving us with doubt, whether it is non- crotonyllysine site or crotonyllysine site. Overall, Eq (21) illustrates the explanation of sensitivity, specificity, overall accuracy in relation to MCC [74, 75].

These perceptive metrics have been adapted and reported by various modern publications (see, e.g. [76–79]. In Eq (21) the set of defined notations can only successfully function for binary labelled data, like whether the predicting site is crotonyllysine or non- crotonyllysine. In case of multi-label prediction, the problem is completely different, which becomes more general in biomedicine [80] and computational biology [81], so required a different type of metrics [82].

Table 1. Self-consistency testing results for iCrotoK-PseAAC.

Predictor	Accuracy Metrics			
	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
iCrotoK-PseAAC	100	100	100	1

https://doi.org/10.1371/journal.pone.0223993.t001

3.3. Self-consistency testing

On iCrotoK-PseAAC, self-consistency test was implemented which is to use the same benchmark dataset to first train and then test the proposed model [73]. When true positive (TP) value is already familiar, then such a data is generally used. In the following Table 1, results are shown, which provides the predicted and actual classification compiled by the proposed computational model. It describes the overall presentation of the proposed systems.

3.4. Testing via 10-fold cross-validation

Authentic datasets are not available always for validation of model, thus, in that case, cross-validation is chosen to develop the exception that the proposed model is predicting accurately [73].

During the cross-validation, k is considered as constant when the disjoint k-fold dataset is achieved after breaking up. Testing is executed k-times for each and every partition after training and accuracy is computed for each of the reiterations. In the end, the average or mean of all the accuracies is recorded as cross-validation. For both positive and negative data samples, the same technique was correlated. To make the subsets for $k = 10$, a casual selection was conducted as compared to the other methods of validation. Table 2 reports the 10-fold cross-validations results for iCrotoK-PseAAC.

3.5. Comparative analysis

Different techniques have reported different results but our methodology has improved the accuracy, specificity but reduced the sensitivity and also improved the Mathew's Co-relation Coefficient. The results were compared with iKcr-PseEns [10] and CKSAAP_CrotSite [9] predictor, the most recent method for predicting crotonyllysine sites. In Table 3 the results of all the four matrices *Sn*, *Sp*, *Acc*, and *MCC* are depicted, signifying that anticipated predictor. It is

Table 2. 10-fold cross-validation results for iCrotoK-PseAAC (average of 10 folds).

Predictor	Accuracy Metrics			
	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
iCrotoK-PseAAC	99.17	99.53	99.40	0.98

https://doi.org/10.1371/journal.pone.0223993.t002

Table 3. Comparative analysis of methods.

Predictor	Accuracy Metrics			
	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
iKcr-PseEns [10]	94.49	95.27	90.53	0.81
CKSAAP_CrotSite [9]	98.11	99.17	92.45	0.9283
iCrotoK-PseAAC	99.17	99.53	99.40	0.98

https://doi.org/10.1371/journal.pone.0223993.t003

observed that the proposed predictor performs better in terms of accuracy, specificity, and sensitivity as compared to the previously reported methods using independent data set.

4. Webserver

Development of a webserver which is user-friendly; is the last 5-step rule. Publicly accessible and user-friendly webserver provides the future directions for making prediction methods and computational tools which will be more useful practically as demonstrated in the recent publications [13, 19, 21, 22, 73, 83, 84]. Specifically, these useful and practical webservers have an increasing effect on medical sciences, leading medicinal chemistry into an unequalled revolution [48], thus, efforts will be made to construct a webserver for the prediction model reported in this paper.

5. Conclusion

Among different post-translational modifications (PTMs), one of the most important ones is lysine crotonylation in proteins. The initial and crucial step is to identify crotonylation occurrence in protein along with their sites to fully understand the mechanism behind these biological processes. The rigid and imbalanced nature of dataset of protein-peptide makes it difficult to understand and time-consuming which affects the precision of the prediction model. We weren't able to get efficient results of sensitivity yet, like accuracy, specificity and MCC. So, the essential demands of computational methods to predict the sites of Crotonylation are highly justified. Our proposed model, iCrotoK-PseAAC which used statistical moments and position relativity to increase the accuracy of the site prediction of lysine Crotonylation. Our predictor model has used SVV, SM, FV, PRIM, RPRIM, AAPIV and RAAPIV (as we discussed above) to compute the accuracy, sensitivity, specificity and MCC. The results of independent dataset testing were 99% accuracy, 89.1% sensitivity, 99.4% specificity and 0.98 MCC. Since our major emphasis was on increasing the accuracy of sites prediction of lysine Crotonylation, which we have achieved.

Supporting information

S1 File. Dataset containing 378 positive sample and 500 negative samples.
(XLSX)

Author Contributions

Conceptualization: Sharaf Jameel Malebary.

Data curation: Muhammad Safi ur Rehman.

Investigation: Muhammad Safi ur Rehman.

Methodology: Muhammad Safi ur Rehman.

Project administration: Sharaf Jameel Malebary.

Resources: Muhammad Safi ur Rehman.

Supervision: Yaser Daanial Khan.

Validation: Yaser Daanial Khan.

Visualization: Sharaf Jameel Malebary.

Writing – original draft: Sharaf Jameel Malebary.

Writing – review & editing: Yaser Daanial Khan.

References

1. Chatterjea M, Shinde R. Textbook of medical biochemistry: 2011.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. 2001; 291(5507):1304–51.
3. Chou K-C. Progresses in predicting post-translational modification. International Journal of Peptide Research and Therapeutics. 2019:1–16.
4. Li S, Li H, Li M, Shyr Y, Xie L, Li YJP, et al. Improved prediction of lysine acetylation by support vector machines. 2009; 16(8):977–83.
5. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. JNar 2002; 30(1):207–10.
6. Glozak M, Sengupta N, Zhang X, Seto EJG. Leaders in Pharmaceutical Business Intelligence (LPBI) Group. 2005; 363(19):15–23.
7. Huang G, Zeng W. A discrete hidden Markov model for detecting histone crotonyllysine sites. JMCMCC 2016; 75:717–30.
8. Qiu W-R, Sun B-Q, Tang H, Huang J, Lin H. Identify and analysis crotonylation sites in histone by using support vector machines. JAiim 2017; 83:75–81.
9. 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. JJoMG, Modelling 2017; 77:200–4.
10. 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.
11. Awais M, Hussain W, Khan YD, Rasool N, Khan SA, 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 transactions on computational biology and bioinformatics. 2019.
12. Ehsan A, Mahmood MK, Khan YD, Barukab OM, Khan SA, Chou K-C. iHyd-PseAAC (EPSV): Identifying Hydroxylation Sites in Proteins by Extracting Enhanced Position and Sequence Variant Feature via Chou's 5-Step Rule and General Pseudo Amino Acid Composition. Current Genomics. 2019; 20(2):124–33. https://doi.org/10.2174/1389202920666190325162307 PMID: 31555063
13. Hussain W, Khan YD, Rasool N, Khan SA, 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. Analytical biochemistry. 2019; 568:14–23. https://doi.org/10.1016/j.ab.2018.12.019 PMID: 30593778
14. Kabir M, Ahmad S, Iqbal M, Hayat M. iNR-2L: A two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families. Genomics. 2019.
15. Le NQK, Yapp EKY, Ou Y-Y, Yeh H-Y. iMotor-CNN: Identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. Analytical biochemistry. 2019; 575:17–26. https://doi.org/10.1016/j.ab.2019.03.017 PMID: 30930199
16. Tahir M, Tayara H, Chong KT. iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. Chemometrics and Intelligent Laboratory Systems. 2019; 189:96–101.
17. He B, Kang J, Ru B, Ding H, Zhou P, Huang J. SABinder: a web service for predicting streptavidin-binding peptides. BioMed research international. 2016; 2016.
18. Kang J, Fang Y, Yao P, Li N, Tang Q, Huang J. NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition. Interdisciplinary Sciences: Computational Life Sciences. 2019; 11(1):108–14.
19. Zhang M, Li F, Marquez-Lago TT, Leier A, Fan C, Kwoh CK, et al. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. 2019.
20. Wang L, Zhang R, Mu Y. Fu-SulfPred: Identification of Protein S-sulfenylation Sites by Fusing Forests via Chou's General PseAAC. Journal of theoretical biology. 2019; 461:51–8. https://doi.org/10.1016/j.jtbi.2018.10.046 PMID: 30365947
21. Hussain W, Khan YD, Rasool N, Khan SA, 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. Journal of theoretical biology. 2019.
22. He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics. 2019; 35(4):593–601. https://doi.org/10.1093/bioinformatics/bty668 PMID: 30052767.

23. Chou K, Forsen S, Zhou G. 3 SCHEMATIC RULES FOR DERIVING APPARENT RATE CONSTANTS. Chemica Scripta. 1980; 16(4):109–13.

24. Li T, Chou K. The flow of substrate molecules in fast enzyme-catalyzed reaction systems. Chemica Scripta. 1980; 16(5):192–6.

25. Lian P, Wei D-Q, Wang J-F, Chou K-C. An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes. PLoS One. 2011; 6(4):e18587. https://doi.org/10.1371/journal.pone.0018587 PMID: 21525996

26. Zhou G-P. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. Journal of Theoretical Biology. 2011; 284(1):142–8. https://doi.org/10.1016/j.jtbi.2011.06.006 PMID: 21718705

27. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. Journal of Biomolecular Structure and Dynamics. 2016; 34(9):1946–61. https://doi.org/10.1080/07391102.2015.1095116 PMID: 26375780

28. Andraos J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws—new methods based on directed graphs. Canadian Journal of Chemistry. 2008; 86(4):342–57.

29. Liu H, Wang M, Chou K-C. Low-frequency Fourier spectrum for predicting membrane protein types. Biochemical and biophysical research communications. 2005; 336(3):737–9. https://doi.org/10.1016/j.bbrc.2005.08.160 PMID: 16140260

30. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology. 2011; 273(1):236–47. https://doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

31. Shen H-B, Chou K-C, Signal-3L: A 3-layer approach for predicting signal peptides. JB communications br. 2007; 363(2):297–303.

32. 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. https://doi.org/10.1371/journal.pone.0105018 PMID: 25121969

33. 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 research international. 2014; 2014.

34. 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. International journal of molecular sciences. 2014; 15(5):7594–610. https://doi.org/10.3390/ijms15057594 PMID: 24857907

35. 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. Journal of Biomolecular Structure and Dynamics. 2015; 33(8):1731–42. https://doi.org/10.1080/07391102.2014.968875 PMID: 25248923

36. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. JJoBS, Dynamics 2016; 34(9):1946–61.

37. Khan YD, Rasool N, Hussain W, Khan SA, Chou K-C. iPhosY-PseAAC: Identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. Molecular Biology Reports. 2018:1–9.

38. Khan YD, Rasool N, Hussain W, Khan SA, Chou K-C. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. Analytical biochemistry. 2018; 550:109–16. https://doi.org/10.1016/j.ab.2018.04.021 PMID: 29704476

39. 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. Analytical biochemistry. 2016; 497:48–56. https://doi.org/10.1016/j.ab.2015.12.009 PMID: 26723495

40. 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):3116–23. https://doi.org/10.1093/bioinformatics/btw380 PMID: 27334473

41. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23):3150–2. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

42. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004; 14(6):1188–90. https://doi.org/10.1101/gr.849004 PMID: 15173120

43. Chou K-C. Impacts of bioinformatics to medicinal chemistry. Medicinal chemistry. 2015; 11(3):218–34. https://doi.org/10.2174/1573406411666141229162834 PMID: 25548930

44. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. JPS, Function, Bioinformatics 2001; 43(3):246–55.

45. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. JB 2013; 29(7):960–2.

46. Lin S-X, Lapointe J. Theoretical and experimental biology in one. JJBSE 2013; 6(4).

47. Zhong W-Z, Zhou S-F. Molecular science for drug development and biomedicine. Multidisciplinary Digital Publishing Institute; 2014.

48. Zhou G-P, Zhong W-Z. Perspectives in Medicinal Chemistry. JCtimc 2016; 16(4):381.

49. Ali F, Hayat M. Classification of membrane protein types using Voting Feature Interval in combination with Chou′s Pseudo Amino Acid Composition. JJotb 2015; 384:78–83.

50. Hajisharifi Z, Piryaiee M, Beigi MM, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou′s pseudo amino acid composition and investigating their mutagenicity via Ames test. JJoTB 2014; 341:34–40.

51. Kabir M, Hayat M. iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. JMg, genomics 2016; 291(1):285–96.

52. 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. JIjoms 2014; 15(3):3495–506.

53. Liu B, Liu F, Fang L, Wang X, Chou K-C. repRNA: a web server for generating various feature vectors of RNA sequences. JMG, Genomics 2016; 291(1):473–81.

54. Chen W, Lin H, Chou K-C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. JMB 2015; 11(10):2620–34.

55. 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. JNar 2015; 43(W1):W65–W71.

56. Akmal MA, Rasool N, Khan YD. Prediction of N-linked glycosylation sites using position relative features and statistical moments. PloS one. 2017; 12(8):e0181966. https://doi.org/10.1371/journal.pone.0181966 PMID: 28797096

57. Butt A, Mahmood MK, Khan YD. ANEXPOSITIONANALYSIS OF FACIAL EXPRESSION RECOGNITION TECHNIQUES. Pakistan Journal of Science. 2016; 68(3).

58. Butt AH, Rasool N, Khan YD. A Treatise to Computational Approaches Towards Prediction of Membrane Protein and Its Subtypes. The Journal of membrane biology. 2017; 250(1):55–76. https://doi.org/10.1007/s00232-016-9937-7 PMID: 27866233

59. Butt AH, Rasool N, Khan YD. Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC. Molecular biology reports. 2018:1–12.

60. Ehsan A, Mahmood K, Khan YD, Khan SA, Chou K-C. A novel modeling in mathematical biology for classification of signal peptides. Scientific reports. 2018; 8(1):1039. https://doi.org/10.1038/s41598-018-19491-y PMID: 29348418

61. Ghauri A, Khan Y, Rasool N, Khan S, Chou K. pNitro-Tyr-PseAAC: Predict nitrotyrosine sites in proteins by incorporating five features into Chou's general PseAAC. Current pharmaceutical design. 2018.

62. Khan YD, Ahmad F, Anwar MW. A neuro-cognitive approach for iris recognition using back propagation. World Applied Sciences Journal. 2012; 16(5):678–85.

63. Khan YD, Khan NS, Farooq S, Abid A, Khan SA, Ahmad F, et al. An Efficient Algorithm for Recognition of Human Actions. The Scientific World Journal. 2014; 2014.

64. Khan YD, Jamil M, Hussain W, Rasool N, Khan SA, Chou K-C. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. Journal of theoretical biology. 2018.

65. Khan YD, Khan SA, Ahmad F, Islam S. Iris recognition using image moments and k-means algorithm. The Scientific World Journal. 2014; 2014.

66. Gluhovsky A, Agee E. Estimating higher-order moments of nonlinear time series. Journal of Applied Meteorology and Climatology. 2009; 48(9):1948–54.

67. Zhu H, Shu H, Zhou J, Luo L, Coatrieux J-L. Image analysis by discrete orthogonal dual Hahn moments. Pattern Recognition Letters. 2007; 28(13):1688–704.

68. Bishop CM. Neural networks for pattern recognition: Oxford university press; 1995.

69. Haykin S. Neural networks: a comprehensive foundation: Prentice Hall PTR; 1994.

70. Petersen B, Lundegaard C, Petersen TN. NetTurnP–neural network prediction of beta-turns by use of evolutionary information and predicted protein sequence features. PLoS One. 2010; 5(11):e15079. https://doi.org/10.1371/journal.pone.0015079 PMID: 21152409

71. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. JNar 1998; 26(9):2230–6.

72. Chou K-C. Prediction of signal peptides using scaled window. Jp 2001; 22(12):1973–9.

73. Awais M, Hussain W, Khan YD, Rasool N, Khan SA, Chou K-C, et al. 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. JIAtocb 2019.

74. Qiu WR, Sun BQ, Xiao X, Xu D, Chou KC. iPhos-PseEvo: Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into General PseAAC via Grey System Theory. Molecular Informatics. 2017; 36(5–6).

75. 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. JO 2016; 7(23):34180.

76. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 2017; 8(3):4208–17. https://doi.org/10.18632/oncotarget.13758 PMID: 27926534.

77. Liu B, Yang F, Huang DS, Chou KC. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics. 2018; 34(1):33–40. https://doi.org/10.1093/bioinformatics/btx579 PMID: 28968797.

78. Ehsan A, Mahmood K, Khan YD, Khan SA, Chou KC. A Novel Modeling in Mathematical Biology for Classification of Signal Peptides. Scientific Reports. 2018; 8:1039. https://doi.org/10.1038/s41598-018-19491-y PMID: 29348418.

79. Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics. 2018. https://doi.org/10.1016/j.ygeno.2018.01.005 PMID: 29360500.

80. 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. JAb 2013; 436(2):168–77.

81. Khan A, Majid A, Hayat M. CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. JCb, chemistry 2011; 35(4):218–29.

82. Chou K-C. Some remarks on predicting multi-label attributes in molecular biosystems. JMB 2013; 9(6):1092–100.

83. 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. Journal of theoretical biology. 2019; 460:195–203. https://doi.org/10.1016/j.jtbi.2018.10.021 PMID: 30312687

84. Cui X, Yu Z, Yu B, Wang M, Tian B, Ma Q, et al. UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. JC 2019; 184:28–43.