*Send Orders for Reprints to reprints@benthamscience.net*



**RESEARCH ARTICLE**

# iHyd-LysSite (EPSV): Identifying Hydroxylysine Sites in Protein Using Statistical Formulation by Extracting Enhanced Position and Sequence Variant Feature Technique

Muhammad Khalid Mahmood[1], Asma Ehsan[1,*], Yaser Daanial Khan[2] and Kuo-Chen Chou[3]

[1]*Department of Mathematics, University of the Punjab, Lahore, Pakistan;* [2]*Faculty of Information Technology, University of Management and Tecnology, Lahore, Pakistan;* [3]*Gordon Life Science Institute, Boston, MA 02478, USA*

**ARTICLE HISTORY**

Received: December 04, 2019
Revised: May 14, 2020
Accepted: May 15, 2020

*DOI:*
*10.2174/1389202921999200831142629*

**Abstract:** ***Introduction*:** Hydroxylation is one of the most important post-translational modifications (PTM) in cellular functions and is linked to various diseases. The addition of one of the hydroxyl groups (OH) to the lysine sites produces hydroxylysine when undergoes chemical modification.

***Methods*:** The method which is used in this study for identifying hydroxylysine sites based on powerful mathematical and statistical methodology incorporating the sequence-order effect and composition of each object within protein sequences. This predictor is called "iHyd-LysSite (EPSV)" (identifying hydroxylysine sites by extracting enhanced position and sequence variant technique). The prediction of hydroxylysine sites by experimental methods is difficult, laborious and highly expensive. *In silico* technique is an alternative approach to identify hydroxylysine sites in proteins.

***Results*:** The experimental results require that the predictive model should have high sensitivity and specificity values and must be more accurate. The self-consistency, independent, 10-fold cross-validation and jackknife tests are performed for validation purposes. These tests are resulted by using three renowned classifiers, Neural Networks (NN), Random Forest (RF) and Support Vector Machine (SVM) with the demanding prediction rate. The overall predictive outcomes are extraordinarily superior to the results obtained by previous predictors. The proposed model contributed an excellent prediction rate in the system for NN, RF, and SVM classifiers. The sensitivity and specificity results using all these classifiers for jackknife test are 96.08%, 94.99%, 98.16% and 97.52%, 98.52%, 80.95%.

***Conclusion*:** The results obtained by the proposed tool show that this method may meet the future demand of hydroxylysine sites with a better prediction rate over the existing methods.

**Keywords:** Hydroxylysine, PTMs, ANN, cross-validation, predictive model, post-translational modifications.

## 1. INTRODUCTION

Numerous proteins experience a broad collection of post-translational modifications. There are two types of modifications; one is called reversible, while another is named as non-reversible. Reversible modifications are related to physiological procedures and significant in the functioning of organisms, whereas later one is related to pathological causes and diseases [1]. Hydroxylation is one of the essential reversible post-translational modifications in protein. In this modification, at least one hydroxyl group is attached to an amino acid by modifying it [2]. The hydroxylation of proline and lysine is the main type of hydroxylated residue in the protein chain, contained in collagen to a large extent [3]. The hydroxylation of proline happens in a gamma-carbon atom, which forms a vital constituent of collagen called hydroxyproline. It is used to maintain the triple helix structure of collagen and in hypoxia through hypoxia-inducible factors hydroxyproline is also expedient [4]. Lack of ascorbate produces deficiencies in hydroxyproline, influences less stability of collagen, which causes metabolic disorder or disease [5]. Another kind of protein hydroxylation is imparted as hydroxylation of a lysine residue, also exclusively produced in collagen [6]. This type of hydroxylation happens in the delta-carbon atom to form hydroxylysine (Fig. **1**) and associated with secretion as well as function in the extracellular matrix [7]. Thus, in the field of biomedical research and drug development, the identification of hydroxyproline and hydroxylysine gives significant information [8].

Mass spectrometry is an experimental method to predict the hydroxylysine site in the protein. The experimental prediction of the hydroxylysine site is pretty difficult, tedious and overpriced [7, 9]. In contrast, the *in-silico* method is much more handy and

*Address correspondence to this author at the Department of Mathematics, University of the Punjab, Lahore, Pakistan;
E-mail: asmak.pu@gmail.com

1389-2029/20 $65.00+.00 

useful in order to predict hydroxylysine sites. This methodology gives the desire results in no time and cost. This is a fundamental approach in bioinformatics in the prediction of the protein modified residue in the process of a post-translational modification. Furthermore, most of the computational algorithms have been developed in order to understand the complex molecular structure and to predict hydroxylation sites [10]. Many similar methods related to post-translational modifications involve, prediction of threonine phosphorylation sites, tyrosine nitration, tyrosine phosphorylation and so forth are described in the series of very recent published articles [11-16]. The predictor "iHyd-PseAAC" was developed for identifying hydroxylysine and hydroxyproline sites in proteome by incorporating the dipeptide position-specific propensity into the general form of pseudo amino acid composition [8]. Another scheme, "iHyd-PseCp", for identifying hydroxyproline and hydroxylysine sites in protein, was developed by Qiu, Wang-Ren *et al.* [17] based on the sequence-coupled information into the general pseudo amino acid composition. The number of encoding schemes based on the composition of k-spaced amino acid pairs (CKSAAP), Amino Acid Composition (AAC), Binary Encoding (BE) and so forth is used for the prediction of phosphorylation, S-sulfenylation and lysine succinylation sites [18-20]. The composition of k-spaced amino acid pairs is an interesting and an effective features extraction technique proposed for identifying lysine formylation sites. By incorporating general pseudo components and Chou's 5-steps rule, this scheme into the CKSAAP method is used to encode formylation sites [21]. The k-spaced amino acid pairs (CKSAAP) encoding scheme is also used in predicting antifreeze proteins [22] and protein phosphorylation sites [23]. Nanni, L. *et al.* [24] developed a technique based on wavelet images and Chou's pseudo amino acid composition for the classification of protein. The technique which is used in this study is taken from the recent work [25, 26], and the predictor is called "iHyd-LysSite (EPSV)" to identify hydroxylysine sites in proteins.

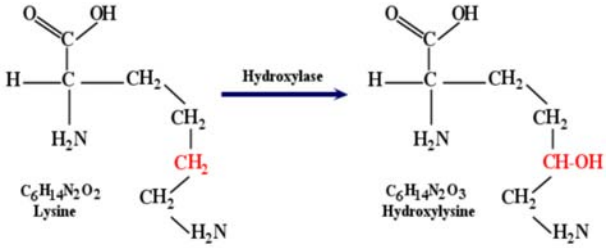


**Fig. (1).** Figure shows the formation of hydroxylysine in the process of protein hydroxylation in lysine residue. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

## 2. METHODS

### 2.1. Benchmark Dataset

The dataset for hydroxylated proteins is established from the UniProt database. The dataset for the term "hydroxylysine" was searched in the field of "modified residue" with PTM/Processing annotation. In order to construct a stringent benchmark dataset, the entries glossed with terms probable, potential, or by similarity were excluded. Against this query, 281 protein sequences were obtained that include hydroxylysine sites. For the sake of convenience, the records found with hydroxylysine sites are denoted as a positive sample and symbolically represented as $\mathbb{L}^+$. Later on, the converse query was run for the sequences not containinghydroxylysine sites. In the result of this query, 500 sequences were obtained. This dataset is considered as a negative dataset and symbolically represented as $\mathbb{L}^-$. The overall dataset comprised of 781 sequences, the sum of the positive and negative dataset, mathematically expressed in Eq. (1). After cutting down the duplicate sequences along with those having homology greater than 60 %, the positive dataset is reduced to 185 samples and negative dataset to 497, accumulatively forming 682 samples. The supplementary information of datasets can be found in Supplementary Tables **S1** and **S2**, respectively.

$$\mathbb{L} = \mathbb{L}^+ + \mathbb{L}^- \tag{1}$$

### 2.2. Construction of Algorithm

To formulate the protein sequences and to classify them according to their attributes, we adopted the scheme as employed by Ehsan *et al.* [25, 26]. The algorithm for peptides classification was encoded in 220 features incorporated three attributes, namely, hydrophobicity, hydrophilicity and side-chain mass of amino acid. This method particularly focuses on the composition and order of each monomer and gives fixed length vector while featuring the polypeptide sample [25, 26]. In this work, by using this methodology, we will identify the hydroxylysine site in an uncharacterized protein sequence. According to this scheme, a peptide **P** sample formulation is described below:

$$P = M_1 M_2 M_3 M_4 M_5 M_6 M_7 \cdots M_n \tag{2}$$

Where $M_1$, $M_2$, $M_3$, ..., $M_n$ represent the amino acid monomers linked by a peptide bond and n represents the number of amino acid residues within the polypeptide sequence (2). Fig. (**2**) shows the proposed scheme for the hydroxylysine site and the feature vector corresponding to the lysine modification site in an uncharacterized protein sequence can be obtained by (3).

$$\sqsubseteq_i = \omega_i + (\omega_i - 1)!\ \tau(M_i, M_i) + \frac{1}{38}[(r-0)_0\{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_0 + \{(s-r)_1\{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_1 + (s-r)_2\{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_2 + (s-r)_3\{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_3 + \cdots + (r-s)_{m-1}\{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_{m-1} + (n-t)_m\{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_m] \tag{3}$$

Where $M_i$ represents a lysine residue and $\omega_i$ denotes the number of repetitions of $M_i$ in sequence, while j varies for all other amino acid residues except $M_i$. $\tau$ represents the pairing of $M_i$ and $M_j$ residues in every possible way by consolidating the difference of each position for $M_i$ residue. In addition, r, s, and t show the corresponding positions of "$M_i$" in (2). Now let,

$$X_0 = \{\sum_{j=1}^{20} q_j\tau(M_j, M_i) + \sum_{j=1}^{20} q_j\tau(M_i, M_j)\}_0 \quad (4)$$

$$X_k = \{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_k \quad (5)$$

$$X_m = \{\sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_j, M_i) + \sum_{\substack{j=1 \\ j\neq i}}^{20} q_j\tau(M_i, M_j)\}_m \quad (6)$$

Equation (3) in term of (4) to (6) can be written in compact form as:

$$\sqsubseteq_i = \omega_i + (\omega_i - 1)!\ \tau(M_i, M_i) + \frac{1}{38}[(r-0)_0X_0 + \sum_{s>r}(s-r)_kX_k + (n-t)_mX_m];$$

$$1 \le r < s < t \le n, k = 1,2,3,\ldots,m-1 \quad (7)$$

The term $(r-0)_0X_0 + \sum_{s>r}(s-r)_kX_k + (n-t)_mX_m$ in (7) can be evaluated by the following constraints

$$\begin{cases} \sum_{s>r}(s-r)_kX_k + (n-t)_mX_m, if\ 1 \le r < s, t < n \\ (r-0)_0X_0 + \sum_{s>r}(s-r)_kX_k + (n-t)_mX_m, if\ 1 < r < s, t < n \\ (r-0)_0X_0 + \sum_{s>r}(s-r)_kX_k, if\ 1 < r < s, t = n \end{cases} \quad (8)$$

The ordinal numbers $j = 1,2,3,\ldots,20$ in (3) represent the amino acids in alphabetical order named as A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Each $M_j$ can cyclically repeat itself more than once and form a peptide sample of length n. As $\sqsubseteq_i$ represents the feature vector corresponding to the ith residue for "K". Similarly, we can define feature vectors for all residues. Thus, we can relate feature vector $\sqsubseteq_1, \sqsubseteq_2, \sqsubseteq_3, \ldots, \sqsubseteq_{20}$ for all twenty amino acids then the set of twenty feature vectors is given by:

$$\sqsubseteq_1 = \omega_1 + (\omega_1 - 1)!\ \tau(M_1, M_1) + \frac{1}{38}[(r-0)_0\{\sum_{\substack{j=2 \\ j\neq 1}}^{20} q_j\tau(M_j, M_1) + \sum_{\substack{j=2 \\ j\neq 1}}^{20} q_j\tau(M_1, M_j)\}_0 + \sum_{s>r}(s-r)_k\{\sum_{\substack{j=2 \\ j\neq 1}}^{20} q_j\tau(M_j, M_1) + \sum_{\substack{j=2 \\ j\neq 1}}^{20} q_j\tau(M_1, M_j)\}_k + (n-t)_m\{\sum_{\substack{j=2 \\ j\neq 1}}^{20} q_j\tau(M_j, M_1) + \sum_{\substack{j=2 \\ j\neq 1}}^{20} q_j\tau(M_1, M_j)\}_m], k = 1,2,3,\ldots,m-1$$

$$\sqsubseteq_2 = \omega_2 + (\omega_2 - 1)!\ \tau(M_2, M_2) + \frac{1}{38}[(r-0)_0\{\sum_{\substack{j=1 \\ j\neq 2}}^{20} q_j\tau(M_j, M_2) + \sum_{\substack{j=1 \\ j\neq 2}}^{20} q_j\tau(M_2, M_j)\}_0 + \sum_{s>r}(s-r)_k\{\sum_{\substack{j=1 \\ j\neq 2}}^{20} q_j\tau(M_j, M_2) + \sum_{\substack{j=1 \\ j\neq 2}}^{20} q_j\tau(M_2, M_j)\}_k + (n-t)_m\{\sum_{\substack{j=1 \\ j\neq 2}}^{20} q_j\tau(M_j, M_2) + \sum_{\substack{j=1 \\ j\neq 2}}^{20} q_j\tau(M_2, M_j)\}_m], k = 1,2,3,\ldots,m-1$$

$$\sqsubseteq_3 = \omega_3 + (\omega_3 - 1)!\ \tau(M_3, M_3) + \frac{1}{38}[(r-0)_0\{\sum_{\substack{j=1 \\ j\neq 3}}^{20} q_j\tau(M_j, M_3) + \sum_{\substack{j=1 \\ j\neq 3}}^{20} q_j\tau(M_3, M_j)\}_0 + \sum_{s>r}(s-r)_k\{\sum_{\substack{j=1 \\ j\neq 3}}^{20} q_j\tau(M_j, M_3) + \sum_{\substack{j=1 \\ j\neq 3}}^{20} q_j\tau(M_3, M_j)\}_k + (n-t)_m\{\sum_{\substack{j=1 \\ j\neq 3}}^{20} q_j\tau(M_j, M_3) + \sum_{\substack{j=1 \\ j\neq 3}}^{20} q_j\tau(M_3, M_j)\}_m], k = 1,2,3,\ldots,m-1$$

$$\vdots$$

$$\sqsubseteq_{20} = \omega_{20} + (\omega_{20} - 1)!\,\tau(M_{20}, M_{20}) + \frac{1}{38}[(r-0)_0\{\sum_{\substack{j=1 \\ j\neq 20}}^{20} q_j\tau(M_j, M_{20}) + \sum_{\substack{j=1 \\ j\neq 20}}^{20} q_j\tau(M_{20}, M_j)\}_0 + \sum_{s>r}(s-r)_k\{\sum_{\substack{j=1 \\ j\neq 20}}^{20} q_j\tau(M_j, M_{20}) + \sum_{\substack{j=1 \\ j\neq 20}}^{20} q_j\tau(M_{20}, M_j)\}_k + (n-t)_m\{\sum_{\substack{j=1 \\ j\neq 20}}^{20} q_j\tau(M_j, M_{20}) + \sum_{\substack{j=1 \\ j\neq 20}}^{20} q_j\tau(M_{20}, M_j)\}_m], k = 1,2,3,\ldots,m-1 \quad (9)$$

The above set of vectors consists of sixty components with respect to three choices of each pair $\tau(M_i, M_j) i, j = 1,2,\ldots,20$, evaluated by using Eqs. (10) to (12) for the hydrophobic property. Similarly, hydrophilicity and side-chain mass attributes give sixty components individually. In this regard, we get 180 components while remaining forty components incorporating a number of occurrence of twenty amino acid residues and the sum of the positions of the corresponding occurrence of each object.

$$\tau(M_i, M_j) = \sqrt{\zeta_b^*(M_i)^2|\zeta_b^*(M_i) - \zeta_b^*(M_j)|^2} + \frac{|(\zeta_1^*(M_i) - \overline{\zeta}_1^*(aa))(\zeta_2^*(M_j) - \overline{\zeta}_2^*(aa)|}{\sqrt{\sum_{i=1}^{20}(\zeta_1^*(M_i) - \overline{\zeta}_1^*(aa))^2 \sum_{j=1}^{20}(\zeta_2^*(M_j) - \overline{\zeta}_2^*(aa))^2}} \quad (10)$$

$$\tau(M_i, M_j) = \sqrt{\zeta_b^*(M_i)^2|\zeta_b^*(M_i) - \zeta_b^*(M_j)|^2} + \frac{|(\zeta_1^*(M_i) - \overline{\zeta}_1^*(aa))(\zeta_3^*(M_j) - \overline{\zeta}_3^*(aa)|}{\sqrt{\sum_{i=1}^{20}(\zeta_1^*(M_i) - \overline{\zeta}_1^*(aa))^2 \sum_{j=1}^{20}(\zeta_3^*(M_j) - \overline{\zeta}_3^*(aa))^2}} \quad (11)$$

$$\tau(M_i, M_j) = \sqrt{\zeta_b^*(M_i)^2|\zeta_b^*(M_i) - \zeta_b^*(M_j)|^2}$$

$$+\frac{|(\zeta_2^*(M_i) - \overline{\zeta}_2^*(aa))(\zeta_3^*(M_j) - \overline{\zeta}_3^*(aa)|}{\sqrt{\sum_{i=1}^{20}(\zeta_2^*(M_i) - \overline{\zeta}_2^*(aa))^2 \sum_{j=1}^{20}(\zeta_3^*(M_j) - \overline{\zeta}_3^*(aa))^2}} \quad (12)$$

Where $\zeta_b^*, b = 1,2,3$ are normalized values of naturally occurring values of hydrophobicity, hydrophilicity and side-chain mass, respectively. These are normalized to the values obtained from the same source, which was taken by Ehsan *et al.* [16]. The mean of normalized values of all twenty amino acids (aa) residues with respect to properties listed above is denoted by $\overline{\zeta_b}^*, b = 1,2,3$. The normalized values are obtained by using equation (13) within normalization range N.

$$\begin{cases} \zeta_1^*(aa) = [\frac{2N}{(\zeta_{1_{(max)}} - \zeta_{1_{(min)}})}(\zeta_1(aa) - \zeta_{1_{(max)}})] + N \\ \zeta_2^*(aa) = [\frac{2N}{(\zeta_{2_{(max)}} - \zeta_{2_{(min)}})}(\zeta_2(aa) - \zeta_{2_{(max)}})] + N \\ \zeta_3^*(aa) = [\frac{2N}{(\zeta_{3_{(max)}} - \zeta_{3_{(min)}})}(\zeta_3(aa) - \zeta_{3_{(max)}})] + N \end{cases} \quad (13)$$

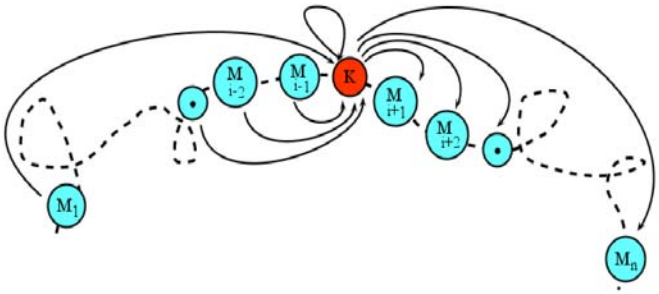


**Fig. (2).** Adopted formulation scheme of the proposed method [16]. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The classifiers, which are used to train the extracted feature vector data, are neural networks (NN), random forest (RF), and support vector machine (SVM). NN algorithm works as a neuron system and every last output of the neuron used as input of the next neuron. In the decision-making problems, neural networks play a key role in solving problems. To identify and incorporate all obscure structure and vague information in the wide collection of datasets, Multi-Layer Perceptron (MLP) is an excellent tool to overcome this difficulty. In any classification problem, MLP is fitted better, as it is adjusted finely by changing the number of hidden layer neurons, training parameters and training algorithms to generate excellent results. To train the extracted feature set, a Multi-Layer Perceptron (MLP) is used (Fig. **3**). The basic strength of neural networks is its flexibility. It has numerous parameters that can be fine tuned to provide the best results. After extensive probing and testing, a neural network was set up having 50 neurons in the hidden layer. Adaptive gradient descent algorithm was incorporated for training, which uses a variable learning rate for optimal convergence. The feature vectors for each sample set are assembled into a large array. In the array, each row represents the feature vector corresponding to a single sequence, whereas each column made up of extracted feature components. Since there are 220 features for each sample, so each row consists of 220 columns. The total columns in feature vectors were 682 out of which 185 were positive samples and weights for each layer randomly adjusted with 75 neurons wereutilized. Moreover, to adjust the weight for each epoch, the back propagation algorithm was employed, while outcomes were obtained after 2693 iterations by the use of the gradient descent method for the learning rate.

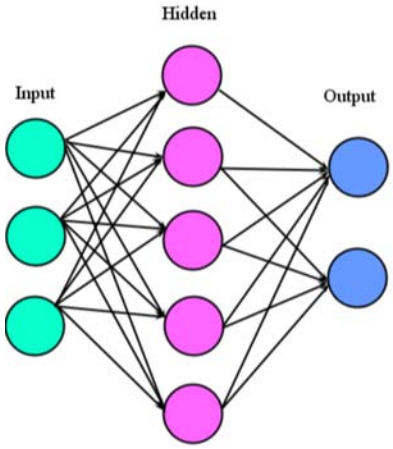


**Fig. (3).** Neural Network. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The results were simulated on MATLAB R2017 version and were duplicated on the python ver 3.6 platform along with Scikit Learn 0.20 for neural network training and simulation bearing identical results. Random forest is another ensemble learning technique for classification. It creates various decision trees on entire data samples by using various learning algorithms to collect prediction results from all of them and decide the final solution upon voting. The support vector machine is mostly used for classification problems. The feature data are plotted over n-space, then draw a line between the two classes by finding hyper-plane for the sake of classification.

## 3. RESULTS AND DISCUSSION

### 3.1. Metrics Evaluation

To evaluate the predictive quality of the proposed predictor, one of the most important and easiest methods was adopted, which is also utilized by Chou [27]. The following set of four metrics based on this formulation was employed in the list of publications [25-27]. In the current study, the four convoluted measures, sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficients (MCC), respectively, were employed to assess the performance of "iHyd-LysSite (EPSV)" predictor as expressed in Ref [8, 17]:

$$\begin{cases} Sn = 1 - \dfrac{L_-^+}{L^+} \; 0 \le Sn \le 1 \\ Sp = 1 - \dfrac{L_+^-}{L^-} \; 0 \le Sp \le 1 \\ Acc = 1 - \dfrac{L_-^+ + L_+^-}{L^+ + L^-} \; 0 \le Acc \le 1 \\ MCC = \dfrac{1 - (\frac{L_-^+}{L^+} + \frac{L_+^-}{L^-})}{\sqrt{(1 + \frac{L_+^- - L_-^+}{L^+})(1 + \frac{L_-^+ - L_+^-}{L^-})}} - 1 \le MCC \le 1 \end{cases} \quad (14)$$

Where $L^+$ describes the total number of accurate predictions of hydroxylysine sites, $L_-^+$ indicates the number of true predictions of hydroxylysine sites wrongly predicted as non-hydroxylysine site; the total number of non-hydroxylysine predicted sites is denoted by $L^-$, while $L_+^-$ is the number of non-hydroxylysine sites investigated as hydroxylysine site. It can be clearly seen from the above equation when $L_-^+ = 0$ describing no true hydroxylysine sites are wrongly predicted to be of non-hydroxylysine sites, which gives $Sn = 1$. When $L_-^+ = L^+$ describing that all the true hydroxylysine sites are wrongly predicted to be of non-hydroxylysine site, we have the sensitivity $Sn = 0$. Similarly, in the case of $L_+^- = 0$, describing none of the non-hydroxylysine sites are wrongly investigated to be as hydroxylysine site, gives $Sp = 1$, while $L_+^- = L^-$ meaning that all non-hydroxylysine sites investigated to be of true hydroxylysine predicted sites, we have specificity $Sp = 0$. On the other hand, when $L_-^+ = L_+^- = 0$ indicating that there are none true hydroxylysine sites and none of the non-hydroxylysine sites are wrongly predicted in positive as well as a negative dataset, we have accuracy $Acc = 1$ and $MCC = 1$. When $L_-^+ = L^+$ and $L_+^- = L^-$ describing that all true hydroxylysine investigated sites are wrongly predicted as non-hydroxylysine sites in a positive dataset and all non-hydroxylysine sites are wrongly investigated as hydroxylysine sites in a negative dataset that gives the overall accuracy $Acc = 0$ and $MCC = -1$; while for $L_-^+ = L^+/2$ and $L_+^- = L^-/2$ we obtained the accuracy $Acc = 0.5$ and $MCC = 0$ describing not good than a random estimate.

Moreover, the set of equations defined in Eq. (14) is only applicable for single-labeled systems and multi-label system, which is useful in systems biology, system medicine and biomedicine [28] defined by more perplexed metrics as given in Ref. [29].

### 3.2. Test Method

In order to score the metrics given in Eq. (14) and to evaluate the performance of the predictor, the following validation methods, self-consistency test, independent dataset test, 10-fold cross-validation test and jackknife test are frequently used. In the jackknife validation process, the test was performed by removing each sample from the given dataset for test purposes, while the remaining dataset was used to train the predictive model. The test is then conducted on the rest of the data trained by a predictive classifier. Moreover, the jackknife test was used to evaluate several predictors as expended in a series of literature [30-32]. While 10-fold cross-validation is partitioned into 10 dissimilar datasets by splitting the dataset for both positive and negative class and outcomes that are generated by taking the mean of all partition outcomes. Each partition gives the independent dataset test individually. These tests are scored by using the following three classifiers for validation purposes, namely, Neural Network (NN), Random Forest (RF), and Support Vector Machine (SVM). The results obtained by using Eq. (14) for all four metrics are given in Table **1**. It can be observed from Table **1**, the accuracy and recall value obtained by employing the proposed predictor "iHyd-LysSite (EPSV)" throughout all classifiers is higher, which gives the results for correctly identified hydroxylated sites. The accuracy graph for the 10-fold cross-validation test is shown in Fig. (**4**). Precision is a positive predictive value (PPV), used to describe the relationship between all true positive predictions and all positive predicted conditions. This test is like a screening test when it returns a positive result or correctly identified hydroxylated sites. It is a probability that protein sequences with a positive screening test result indeed have the hydroxylated sites. The precision values table for all varying classifiers is given in Table **1**. The Receiver Operating Characteristic (ROC) is a two-dimensional graphical representation used to explain the performance of the predictive model by the area under the curve (AUC) or area under the ROC curve. The value of AUC ranges from 0 to 1. The AUC with value 0.0 represents the 100% wrong prediction, while AUC = 1.0 is obtained when the prediction is 100% correct. The more area under the curve, the more accurate the model. The ROC for a 10-fold cross-validation test for all three classifiers is shown in Fig. (**5**), and the comparison of all classifiers for self-consistency test is presented in Fig. (**6**).

### 3.3. Comparison with Previous Methods

In this study, a comparison is established by the former prediction methods by using a rigorous jackknife test to check the quality of the proposed model "iHyd-LysSite (EPSV)". The comparison is made among all classifiers, neural network (NN), random forest (RF), and support vector machine (SVM). The jackknife results for all classifiers obtained by using the proposed model "iHyd-LysSite (EPSV)" for the above metrics in Eq. (14) are given in (Table **2**). The examination is prepared with two existing predictors, the "iHyd-PseAAC" [8], and " iHyd-PseCp" [17]. These methods have also achieved the metric scores using the jackknife test method and it is easy to see from Table **2** that, the accuracy (Acc), stability (MCC), sensitivity (Sn), and specificity (Sp) scores evaluated by the proposed predictor are superior than calculated by existing predictors. It can also be noticed that all classifiers contribute to excellent scores in the result of Eq. (14). It is also demonstrated by the cross-validation test, the overall prediction accuracies of the system for all three classifiers are 96.77%, 97.31 and 84.38. It can be observed that the overall accuracy of the predictor "iHydLysSite (EPSV)" is higher than the existing predictors.

Due to the following reasons, the proposed predictor is more reliable and robust, in prediction. First is the formulation of sequence, which is convenient in handling the diverse length sequences in a generous way without skipping

**Table 1. The values of all four metrics for three classifiers obtained by using the proposed predictor "iHyd-LysSite (EPSV)".**

| Classifiers NN | | | | | RF | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Tests** | Sn (%) | Sp (%) | Acc (%) | MCC | Sn (%) | Sp (%) | Acc (%) | MCC | Sn (%) | Sp (%) | Acc (%) | MCC |
| **Self-consistency** | 95.15 | 99.39 | 98.00 | 0.95 | 100.00 | 99.60 | 99.74 | 0.99 | 86.01 | 92.93 | 90.40 | 0.88 |
| **Independent** | 93.00 | 100.00 | 97.40 | 0.95 | 95.10 | 97.12 | 96.32 | 0.93 | 81.32 | 92.36 | 88.09 | 0.87 |
| **Cross-validation** | 96.14 | 97.57 | 96.77 | 0.92 | 95.04 | 98.60 | 97.31 | 0.94 | 98.22 | 81.00 | 84.38 | 0.89 |
| **Jackknife** | 96.08 | 97.52 | 97.14 | 0.93 | 94.99 | 98.52 | 97.24 | 0.90 | 98.16 | 80.95 | 84.33 | 0.84 |

**Precision Table**

| Classifiers NN | | RF | SVM |
|---|---|---|---|
| **Tests** | PPV | PPV | PPV |
| **Self-consistency** | 0.89 | 0.99 | 0.88 |
| **Independent** | 0.88 | 0.93 | 0.87 |
| **Cross-validation** | 0.92 | 0.97 | 0.78 |
| **Jackknife** | 0.88 | 0.92 | 0.74 |

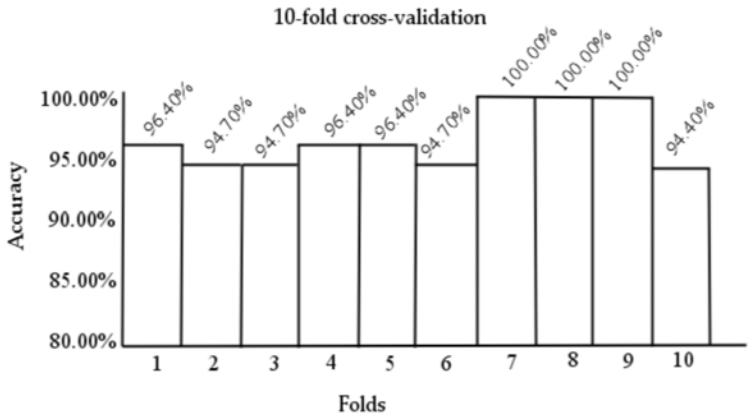


**Fig. (4).** The graph shows the 10-fold cross-validation performed on the overall dataset and the corresponding accuracy for each fold test, the results are generated by employing the Neural network (NN) classifier.

any information of the sequence, and makes pairwise couplings in every possible combination with amino acids. Second is the fixed length vector, which always imparts with a non-variable feature vector that equally separates the proteins according to their attributes. Due to this reason, each sample could rigorously classify and conveniently recognize. The third is about correlation expression, this correlation mainly takes part in scoring the feature vector, that is manipulated by incorporating each attribute group. Each and every expression deals with some specific metric and statistical expressions. For the sake of convenience, every amino acid was standardized with a suitable range, that the value of each property of amino acid lies between this range. Moreover, it is observed that, in comparison with previous methods, the proposed predictor outcomes are more superior and better than the former prediction rate.

### 3.4. Friendly User Web-Server

"User-friendly and publicly accessible web-servers represent the current trend for developing various computational methods [33], as reflected by a serious of recent publications see *e.g.* [15, 34-37]. Actually, they have significantly enhanced the impacts of computational biology on medical science and driving medicinal chemistry into an unprecedented revolution [38], here we shall do our best to provide a web-server for the predictor presented in this paper as soon as possible."

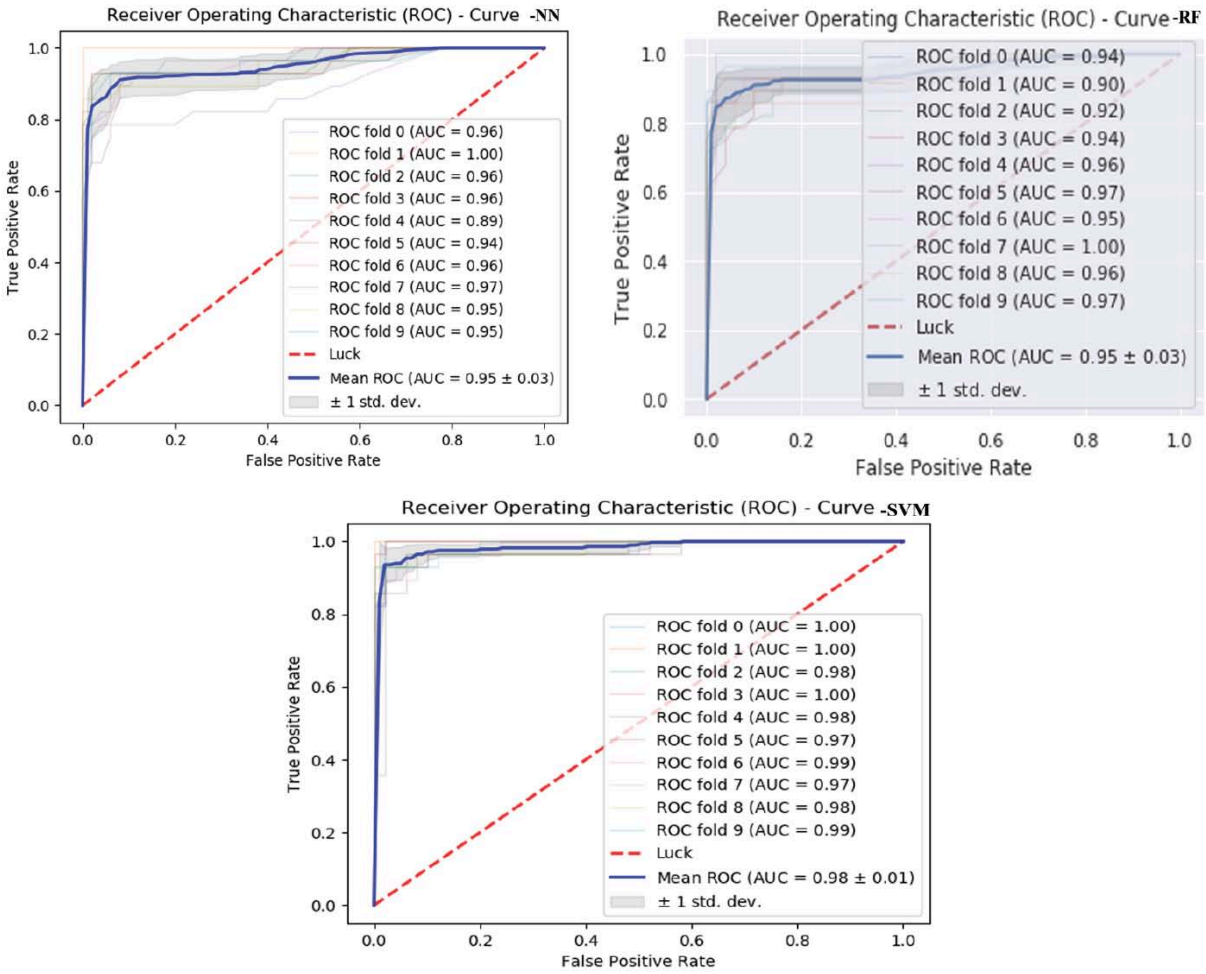


**Fig. (5).** The ROC curves obtained from the classifiers, NN, RF, SVM for the 10-fold cross-validation test. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

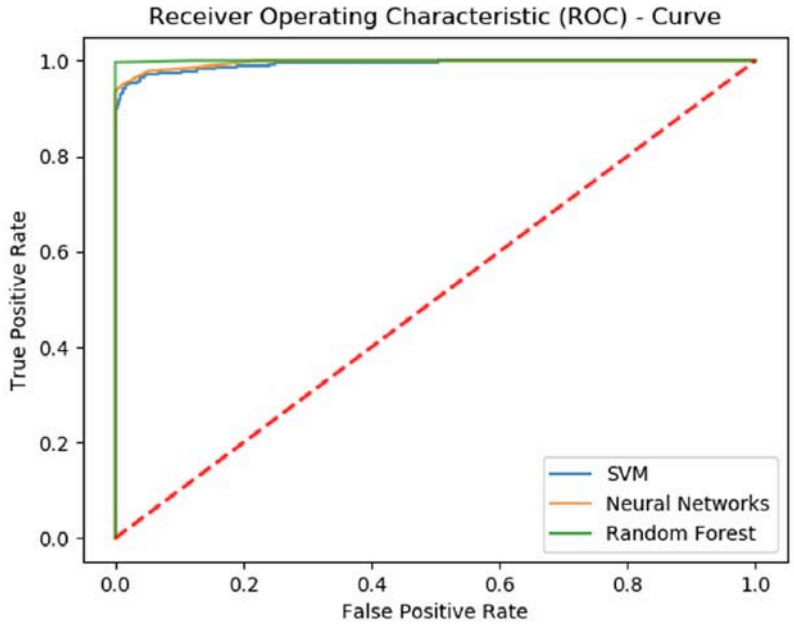


**Fig. (6).** The comparison of NN, RF, SVM ROC curves for self-consistency test. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

**Table 2. A comparison of the proposed model "iHyd-LysSite (EPSV)" with the previous methods using the jackknife test validated by NN, RF, and SVM classifiers.**

| Methods | Sn[a] | Sp[a] | Acc[a] | MCC[a] |
|---|---|---|---|---|
| **iHyd-PseAAC** | 87.85 | 83.01 | 83.56 | 0.50 |
| **iHyd-PseCp** | 78.77 | 99.08 | 97.08 | 0.86 |
| **iHyd-LysSite (EPSV)-NN[b]** | 96.08 | 97.52 | 97.14 | 0.93 |
| **iHyd-LysSite (EPSV)-RF[b]** | 94.99 | 98.52 | 97.24 | 0.90 |
| **iHyd-LysSite (EPSV)-SVM [b]** | 98.16 | 80.95 | 84.33 | 0.84 |

(***a***) definition of metrics in Eq. (14), (***b***) proposed predictor "iHyd-LysSite (EPSV)".

## CONCLUSION

In cellular functions, the Post-Translational Modification (PTM) of protein is of vital importance. Covalent addition of any functional group to the proteins produces PTM. Hydroxylation is one of the PTM reactions, which is mostly occurring on three residues, proline, lysine and asparagine. On the maturation of collagen fibers, hydroxyproline and hydroxylysine are significant, while, hydroxyasparagine is important for antifungal and anti-toxin drugs. Hydroxylysine is the hydroxylated class of lysine residue and plays a central role in both biomedical research and drug development against cancer and many other diseases. A powerful computational approach was adopted for identifying the potential hydroxylysine sites in proteins. In the current work, we prove that "iHyd-LysSite (EPSV)" is a predictor that has an excellent prediction proficiency for identifying hydroxylysine sites on a comparison with the former techniques. For this purpose, the methodology is used taken from a recent published article given in Ref. [25]. To validate the potency of the proposed model, the exhaustive jackknife test was performed. The model is verified with three main classifiers, Neural Network (NN), Random Forest (RF) and Support Vector Machine (SVM). Then 96.08%, 94.99%, 98.16% sensitivity and 97.52%, 98.52%, and 80.95% specificity results have been obtained for the jackknife test using the above three classifiers. It is concluded that the proposed predictor has the potential of more improvement in the computed result as in a continuous sequence, there are so rapidly increasing combinations of lysine residues.

## 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

None.

## CONFLICT OF INTEREST

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

## ACKNOWLEDGEMENTS

Declared none.

## SUPPLEMENTARY MATERIAL

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

## REFERENCES

[1] Xie, H.; Vucetic, S.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J. Proteome Res.*, **2007**, *6*(5), 1917-1932. http://dx.doi.org/10.1021/pr060394e PMID: 17391016

[2] Kaelin, W.G.; William, G. Proline hydroxylation and gene expression. *Annu. Rev. Biochem.,* **2005**, *74*, 115-128. http://dx.doi.org/10.1146/annurev.biochem.74.082803.133142 PMID: 15952883

[3] Chopra, R.K.; Ananthanarayanan, V.S. Conformational implications of enzymatic proline hydroxylation in collagen. *Proc. Natl. Acad. Sci. USA,* **1982**, *79*(23), 7180-7184. http://dx.doi.org/10.1073/pnas.79.23.7180 PMID: 6296823

[4] Berra, E.; Ginouvès, A.; Pouysségur, J. The hypoxia-inducible-factor hydroxylases bring fresh air into hypoxia signalling. *EMBO Rep.,* **2006**, *7*(1), 41-45. http://dx.doi.org/10.1038/sj.embor.7400598 PMID: 16391536

[5] Salnikow, K.; Kasprzak, K.S. Ascorbate depletion: a critical step in nickel carcinogenesis? *Environ. Health Perspect.,* **2005**, *113*(5), 577-584. http://dx.doi.org/10.1289/ehp.7605 PMID: 15866766

[6] Yamauchi, M.; Shiiba, M. Lysine hydroxylation and cross-linking of collagen. *Methods Mol. Biol.,* **2008**, *446*, 95-108. http://dx.doi.org/10.1007/978-1-60327-084-7_7 PMID: 18373252

[7] Richards, A.A.; Stephens, T.; Charlton, H.K.; Jones, A.; Macdonald, G.A.; Prins, J.B.; Whitehead, J.P. Adiponectin multimerizati-

on is dependent on conserved lysines in the collagenous domain: evidence for regulation of multimerization by alterations in post-translational modifications. *Mol. Endocrinol.,* **2006**, *20*(7), 1673-1687.
http://dx.doi.org/10.1210/me.2005-0390 PMID: 16497731

[8] 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

[9] Cockman, M.E.; Webb, J.D.; Kramer, H.B.; Kessler, B.M.; Ratcliffe, P.J. Proteomics-based identification of novel factor inhibiting hypoxia-inducible factor (FIH) substrates indicates widespread asparaginyl hydroxylation of ankyrin repeat domain-containing proteins. *Mol. Cell. Proteomics,* **2009**, *8*(3), 535-546.
http://dx.doi.org/10.1074/mcp.M800340-MCP200 PMID: 18936059

[10] Hu, L.L.; Niu, S.; Huang, T.; Wang, K.; Shi, X.H.; Cai, Y.D. Lysine hydroxylation and cross-linking of collagen. *Methods Mol. Biol.,* **2010**, *446*, 95-108.

[11] 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.

[12] Butt, A.H.; Khan, Y.D. Prediction of S-Sulfenylation sites using statistical moments based features *via* Chou'S 5-Step rule. *Int. J. Pept. Res. Ther.,* **2019**, *2019*, 1-11.
http://dx.doi.org/10.1007/s10989-019-09931-2

[13] Malebary, S.J.; Rehman, M.S.U.; Khan, Y.D. iCrotoK-PseAAC: Identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. *PLoS One,* **2019**, *14*(11), e0223993.
http://dx.doi.org/10.1371/journal.pone.0223993 PMID: 31751380

[14] Khan, S.A.; Khan, Y.D.; Ahmad, S.; Allehaibi, K.H. N-MyristoylG-PseAAC: sequence-based prediction of N-myristoyl glycine sites in proteins by integration of PseAAC and statistical moments. *Lett. Org. Chem.,* **2019**, *16*(3), 226-234.
http://dx.doi.org/10.2174/1570178616666181217153958

[15] Liu, Y.; Wang, M.; Xi, J.; Luo, F.; Li, A. PTM-ssMP: a web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *Int. J. Biol. Sci.,* **2018**, *14*(8), 946-956.
http://dx.doi.org/10.7150/ijbs.24121 PMID: 29989096

[16] Basu, S.; Plewczynski, D. AMS 3.0: prediction of post-translational modifications. *BMC Bioinformatics,* **2010**, *11*(1), 210.
http://dx.doi.org/10.1186/1471-2105-11-210 PMID: 20423529

[17] 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.

[18] Hasan, M.M.; Rashid, M.M.; Khatun, M.S.; Kurata, H. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. *Sci. Rep.,* **2019**, *9*(1), 8258.
http://dx.doi.org/10.1038/s41598-019-44548-x PMID: 31164681

[19] Hasan, M.M.; Guo, D.; Kurata, H. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. *Mol. Biosyst.,* **2017**, *13*(12), 2545-2550.
http://dx.doi.org/10.1039/C7MB00491E PMID: 28990628

[20] Hasan, M.M.; Khatun, M.S.; Kurata, H. Large-scale assessment of bioinformatics tools for lysine succinylation sites. *Cells,* **2019**, *8*(2), 95.
http://dx.doi.org/10.3390/cells8020095 PMID: 30696115

[21] Ju, Z.; Wang, S.Y. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs *via* Chou's 5-steps rule and general pseudo components. *Genomics,* **2020**, *112*(1), 859-866.
http://dx.doi.org/10.1016/j.ygeno.2019.05.027 PMID: 31175975

[22] Usman, M.; Lee, J.A. Afp-cksaap: prediction of antifreeze proteins using composition of k-spaced amino acid pairs with deep neural network. *arXiv preprint,* **1910**.

[23] Zhang, S.; Li, X.; Fan, C.; Wu, Z.; Liu, Q. Application of machine learning techniques to predict protein phosphorylation sites. *Lett. Org. Chem.,* **2019**, *16*(4), 247-257.
http://dx.doi.org/10.2174/1570178615666180907150928

[24] Nanni, L.; Brahnam, S.; Lumini, A. Wavelet images and Chou's pseudo amino acid composition for protein classification. *Amino Acids,* **2012**, *43*(2), 657-665.
http://dx.doi.org/10.1007/s00726-011-1114-9 PMID: 21993538

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

[26] Ehsan, A.; Mahmood, M.K.; Khan, Y.D.; Barukab, O.M.; Khan, S.A.; 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. *Curr. Genomics,* **2019**, *20*(2), 124-133.
http://dx.doi.org/10.2174/1389202920666190325162307 PMID: 31555063

[27] Chou, K.C. Prediction of protein signal sequences and their cleavage sites. *Proteins,* **2001**, *42*(1), 136-139.
http://dx.doi.org/10.1002/1097-0134(20010101)42:1<136::AID-PROT130>3.0.CO;2-F PMID: 11093267

[28] 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

[29] 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

[30] Li, S.; Li, H.; Li, M.; Shyr, Y.; Xie, L.; Li, Y. Improved prediction of lysine acetylation by support vector machines. *Protein Pept. Lett.,* **2009**, *16*(8), 977-983.
http://dx.doi.org/10.2174/092986609788923338

[31] Shi, M.G.; Huang, D.S.; Li, X.L. A protein interaction network analysis for yeast integral membrane protein. *Protein Pept. Lett.,* **2008**, *15*(7), 692-699.
http://dx.doi.org/10.2174/092986608785133627 PMID: 18782064

[32] Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.,* **2007**, *248*(3), 546-551.
http://dx.doi.org/10.1016/j.jtbi.2007.06.001 PMID: 17628605

[33] Salvatore, M.; Shu, N.; Elofsson, A. The SubCons webserver: a user friendly web interface for state□of□the□art subcellular localization prediction. *Prot. Sci.,* **2018**, *27*, 195-201.
http://dx.doi.org/10.1002/pro.3297

[34] van Zundert, G.C.P.; Rodrigues, J.P.G.L.M.; Trellet, M.; Schmitz, C.; Kastritis, P.L.; Karaca, E.; Melquiond, A.S.J.; van Dijk, M.; de Vries, S.J.; Bonvin, A.M.J.J. The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.,* **2016**, *428*(4), 720-725.
http://dx.doi.org/10.1016/j.jmb.2015.09.014 PMID: 26410586

[35] Ghouzam, Y.; Postic, G.; Guerin, P.E.; de Brevern, A.G.; Gelly, J.C. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Sci. Rep.,* **2016**, *6*(1), 28268.
http://dx.doi.org/10.1038/srep28268 PMID: 27319297

[36] Wang, D.; Liu, D.; Yuchi, J.; He, F.; Jiang, Y.; Cai, S.; Li, J.; Xu, D. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.,* **2020**, *48*(W1), W140-W146.
http://dx.doi.org/10.1093/nar/gkaa275 PMID: 32324217

[37] Gnanavel, M.; Mehrotra, P.; Rakshambikai, R.; Martin, J.; Srinivasan, N.; Bhaskara, R.M. CLAP: a web-server for automatic clas-

sification of proteins with special reference to multi-domain proteins. *BMC bioinformatics*, **2014**, *15*(1), 343. http://dx.doi.org/10.1186/1471-2105-15-343

[38] Weng, G.; Wang, E.; Wang, Z.; Liu, H.; Zhu, F.; Li, D.; Hou, T. HawkDock: a web server to predict and analyze the protein-protein complex based on computational docking and MM/GBSA. *Nucleic Acids Res.,* **2019**, *47*(W1), W322-W330. http://dx.doi.org/10.1093/nar/gkz397 PMID: 31106357