Send Orders for Reprints to reprints@benthamscience.net

124 *Current Genomics,* **2019**, *20*, 124-133

RESEARCH ARTICLE

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

Asma Ehsan[1,*], Muhammad K. Mahmood[1], Yaser D. Khan[2], Omar M. Barukab[2], Sher A. Khan[3] and Kuo-Chen Chou[4]

[1]Department of Mathematics, University of the Punjab, Lahore, Pakistan; [2]Faculty of Information Technology, University of Management and Technology, Lahore, Pakistan; [3]King Abdul Aziz University, Faculty of Computing and Information Technology in Rabigh, Jeddah, KSA; [4]Gordon Life Science Institute, Boston, MA 02478, USA

ARTICLE HISTORY

Received: January 18, 2019
Revised: March 15, 2019
Accepted: March 18, 2019

DOI: 10.2174/1389202920666190325162307

Abstract: ***Background*:** In various biological processes and cell functions, Post Translational Modifications (PTMs) bear critical significance. Hydroxylation of proline residue is one kind of PTM, which occurs following protein synthesis. The experimental determination of hydroxyproline sites in an uncharacterized protein sequence requires extensive, time-consuming and expensive tests.

***Methods*:** With the torrential slide of protein sequences produced in the post-genomic age, certain remarkable computational strategies are desired to overwhelm the issue. Keeping in view the composition and sequence order effect within polypeptide chains, an innovative *in-silico* predictor *via* a mathematical model is proposed.

***Results*:** Later, it was stringently verified using self-consistency, cross-validation and jackknife tests on benchmark datasets. It was established after a rigorous jackknife test that the new predictor values are superior to the values predicted by previous methodologies.

***Conclusion*:** This new mathematical technique is the most appropriate and encouraging as compared with the existing models.

Keywords: PseAAC, Hydroxylation of proline, Post Translational Modifications (PTMs), Sequence-coupling model, Mammalian proteins, Hydroxyproline.

1. INTRODUCTION

Collagens are profoundly plenteous mammalian proteins which possess abundant hydroxyproline [1] that plays a key role in its stability. The structure of collagen is stringy and long; nearly a quarter or even more of total protein content in mammals is comprised of collagen [2]. In medical applications, collagens work as a major constituent while contributing to wound healing [3], burns surgery [4] and cosmetic surgery [5]. Their asymmetrical behavior and irregular movements may contribute to stomach disease [6] and lung cancer [7]. The ability to predict hydroxyproline (HyP) sites as a result of post-translational modifications in proteins provides precious information useful for both biomedical research and medication evolution [8]. Hydroxyproline is a non-essential amino acid which means that it is mostly synthesized with other amino acids in the liver and need not to be obtained directly through systemic ingestion. Proline undergoes hydroxylation by the conversion of CH_2 group in proline residue into $CH-OH$ group or a hydroxyl group [8] as shown in Fig. (**1**).

Owing to its significance for an in-depth understanding of the cellular biological process and discovering drug against cancers and other major diseases, many efforts have been made by other scientists in this regard [9-17]. Although, experimental techniques based on mass spectrometry exist that are used to determine hydroxylation sites of a given protein [18], however, this is laborious, tedious and high-priced. As a multitude of proteomic sequences are gathered into databanks each day, it is extremely desirable to devise an integrated and robust computational technique incorporating the composition and sequence order effect to determine potential hydroxylation sites with greater accuracy. Researchers have proposed a few methodologies for this purpose. However, the existing predictors lack the most pertinent details of features obscured within the primary sequences that prove crucial for reaching an accurate decision. Hydroxylation process had been of great interest to many researchers. Quantification of hydroxyproline was estimated by Colgrave, *et al.* [1] by using multiple-reaction-monitoring mass spectrometry. A mathematical modeling has been developed to understand

*Address correspondence to this author at the Department of Mathematics, University of the Punjab, Lahore, Pakistan; Fax: +92-99230329;
E-mail: asmak.pu@gmail.com

1389-2029/19 $58.00+.00

the microbial behavior and their communities [19]. It was shown that the hydroxyproline and hydroxylysine in collagen were integrated by a clear extraordinary pathway, in which proline and lysine were hydroxylated after they were consolidated into a comprehensive polypeptide antecedent of collagen. Berg, *et al.* [20] defined a system that was set up to examine the inadequacy of collagen in connective tissues occurring due to lack of ascorbates to some extent.

The isolation and partial characterization of highly purified protocollagen proline hydroxylase and hydroxylation of proline in synthetic polypeptides with purified procollagen hydroxylase were elaborated by Halme *et al.* [21] and Kivirikko *et al.* [22]. Morgan, *et al.* [23] investigated, in terms of the distribution, the frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. Hydroxylation of lysine and crosslinking of collagens have been discussed in "Posttranslational Modifications of Proteins" [24]. Shi, Shao-Ping, *et al.* [25] presented a new method named as PredHydroxy to mechanize the forecast of the proline and lysine hydroxylation locales in term of position weight of 8 high-quality amino acid indices and support vector machines. The metabolism for the proline, hydroxyproline and a survey of activity of proline with the changing environment were also studied [26, 27]. Employing support vector machine and developing a tool for prediction of hydroxyproline sites were proposed by ZR Yang [28]. Hu, Le-Le, *et al.* [29] developed a sequence-based methodology for predicting hydroxylation of hydroxyproline and hydroxylysine. Xu, Yan, *et al.* [8] predicted hydroxyproline and hydroxylysine in proteins using dipeptide position and specific propensity into pseudo amino acid composition. An improved approach over this proficiency was proposed by Qiu, Wang-Ren, *et al.* [30] by integrating a sequence-coupled effect into general PseAAC.

2. RESULTS

To develop a worthwhile predictor for a biological phenomenon, one should observe the Chou's 5-step rule [31]. It is indeed good to present the new prediction method by observing the Chou's 5-step rule as many researchers followed this fundamental rule in their papers, published very recently [9, 32-38]. In the first step, benchmark dataset is accumulated for training and testing the predictor; in the next step, a mathematical model is formulated which sieves out the most momentous features of the polypeptide sequence. Later the feature vector is integrated into a prediction algorithm for training. Once the training is completed, the trained model is thoroughly tested and validated. Lastly, a web-server is developed for open use of the prediction model. In this study, the first four steps have been meticulously performed, however, the last step has been kept open for future work.

3. ACCURACY METRICS

In order to measure the predictive quality of the predictor, the following metrics are commonly used: Acc is used to quantify the comprehensive accuracy of the predictor, MCC is a stable measure of overall accuracy of the model, Sn is used to estimate sensitivity, and Sp is used for specificity [39]. To evaluate the prediction rate of the proposed model, this set of metrics is followed which are also employed by Ehsan *et al.* [40]. The formulation for the actual prediction of hydroxylated $\mathbb{H}^+$ and non-hydroxylated $\mathbb{H}^-$ site of proline is given below.

$$\mathbb{H}^+ = \frac{\mathbb{P}^+ - \mathbb{P}^+_-}{\mathbb{P}^+} \tag{1}$$

$$\mathbb{H}^- = \frac{\mathbb{P}^- - \mathbb{P}^-_+}{\mathbb{P}^-} \tag{2}$$

Where $\mathbb{P}^+$ and $\mathbb{P}^+_-$ represent the total number of peptides which was correctly predicted with proline hydroxylated site and the number of hydroxylated peptides which was incorrectly predicted as a non-hydroxylated proline site, respectively. Likewise $\mathbb{P}^-$ and $\mathbb{P}^-_+$ represent the total actual count of non-hydroxylated peptides and the number of wrongly predicted hydroxylated peptides, respectively.

$$\mathbb{H} = \frac{\mathbb{H}^+\mathbb{P}^+ + \mathbb{H}^-\mathbb{P}^-}{\mathbb{P}^+ + \mathbb{P}^-} = 1 - \frac{\mathbb{P}^+_- + \mathbb{P}^-_+}{\mathbb{P}^+ + \mathbb{P}^-} \tag{3}$$

It has been observed that when there are zero incorrectly predicted hydroxylated and non-hydroxylated proline peptides such that $\mathbb{P}^+_- = \mathbb{P}^-_+ = 0$ then equation (1) to (3) gives $\mathbb{H}^+ = \mathbb{H}^- = 1$ and $\mathbb{H} = 1$ signifying the highest possible accuracy rate. Subsequently, when $\mathbb{P}^+_- = \mathbb{P}^-_+ \neq 0$, then the prediction would be less than 1. There are a number of statistical equations which are used to measure the performance of the predictor given in eq (4).

$$\begin{cases} Sn = \frac{TP}{TP+FN} \\ Sp = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+TN+FP+FN} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(FN+TN)(FP+TN)(TP+FN)}} \end{cases} \tag{4}$$

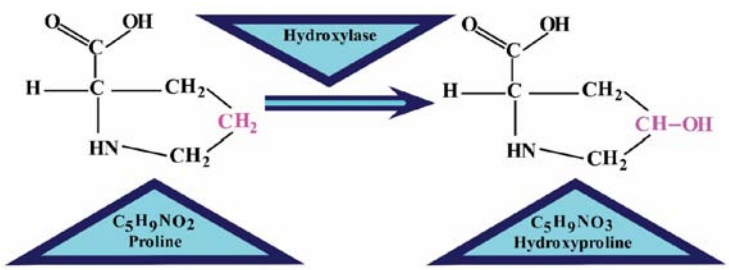

Fig. (1). Figure shows how CH_2 group is converted into $CH-OH(-OH)$ group in the process of proline hydroxylation.

Where TP, TN, FP and FN represent the true positive, true negative, false positive and false negative values, respectively. Expressions in equations (5) and (6) represent the symbols in terms of equation (1) to (3). It is also advantageous to use the intuitive metrics of Equations (5)-(6) to replace the traditional Equation (4). Either the set of traditional metrics copied from maths books or the intuitive metrics derived from the Chou's symbols [41-43] are valid only for the single-label systems (where each sample only belongs to one class). For the multi-label systems (where a sample may simultaneously belong to several classes), whose existence has become more frequent in system biology [32, 33, 36, 44], system medicine [45] and biomedicine [46], a completely different set of metrics as defined in the study represnted as reference [47] is absolutely needed.

$$\begin{cases} TP = \mathbb{P}^{+} - \mathbb{P}^{+}_{-} \\ TN = \mathbb{P}^{-} - \mathbb{P}^{-}_{+} \\ FP = \mathbb{P}^{-}_{+} \\ FN = \mathbb{P}^{+}_{-} \end{cases} \tag{5}$$

$$\begin{cases} Sn = 1 - \frac{\mathbb{P}^{+}_{-}}{\mathbb{P}^{+}} \\ Sp = 1 - \frac{\mathbb{P}^{-}_{+}}{\mathbb{P}^{-}} \\ Acc = \mathbb{H} = 1 - \frac{\mathbb{P}^{+}_{-} + \mathbb{P}^{-}_{+}}{\mathbb{P}^{+} + \mathbb{P}^{-}} \\ MCC = \frac{1 - (\frac{\mathbb{P}^{+}_{-}}{\mathbb{P}^{+}} + \frac{\mathbb{P}^{-}_{+}}{\mathbb{P}^{-}})}{\sqrt{(1 + \frac{\mathbb{P}^{-}_{+} - \mathbb{P}^{+}_{-}}{\mathbb{P}^{+}})(1 + \frac{\mathbb{P}^{+}_{-} - \mathbb{P}^{-}_{+}}{\mathbb{P}^{-}})}} \end{cases} \tag{6}$$

It is relevant to discuss the following cases of the above equation (6), if $\mathbb{P}^{+}{}_{-} = 0$ then there is no incorrectly predicted hydroxylated proline peptides as non-hydroxylated proline peptides such that $Sn = 1$. Similarly, when $\mathbb{P}^{+}{}_{-} = \mathbb{P}^{+}$, it indicates that all hydroxylated proline peptides were incorrectly predicted as non-hydroxylated proline peptides, hence the sensitivity was computed as $Sn = 0$. Furthermore, $\mathbb{P}^{-}{}_{+} = 0$ yields specificity, $and\ Sp = 1$ represents that not even one non-hydroxylated proline peptide was incorrectly predicted as a hydroxylated proline peptide. Likewise $\mathbb{P}^{-}{}_{+} = \mathbb{P}^{-}$ yields specificity, and $Sp = 0$ represents that all non-hydroxylated proline peptides were incorrectly predicted as hydroxylated proline peptides. Also, $Acc = \mathbb{H} = 1$ implies that all sequences of hydroxylated $\mathbb{H}^{+}$ and non-hydroxylated $\mathbb{H}^{-}$ proline peptides were predicted correctly such that $\mathbb{P}^{+}{}_{-} = \mathbb{P}^{-}{}_{+} = 0$. Further, the performance of binary classifications is often measured by Matthew correlative coefficient (MCC). There were three cases herein, $\mathbb{P}^{+}{}_{-} = \mathbb{P}^{-}{}_{+} = 0$ indicates that no incorrectly predicted sequences were found both for hydroxylated $\mathbb{H}^{+}$ and non-hydroxylated $\mathbb{H}^{-}$ peptides yielding $MCC = 1$. In the second case, $\mathbb{P}^{+}{}_{-} = \frac{\mathbb{P}^{+}}{2}$ and $\mathbb{P}^{-}{}_{+} = \frac{\mathbb{P}^{-}}{2}$ generated $MCC = 0$ indicating that this prediction was not more accurate than the random prediction. Lastly, with values of $\mathbb{P}^{+}{}_{-} = \mathbb{P}^{+}$ and $\mathbb{P}^{-}{}_{+} = \mathbb{P}^{-}$, $MCC = -1$ was obtained signifying a totally wrong binary classification and complete disagreement between the observed and predicted values.

4. VALIDATION METHOD

The metrics given in equation (6) are used to describe three frequently used test methods namely, independent dataset test, K-fold cross-validation test, and jackknife test. These tests are considered beneficial in validating the quality of the predictor. The jackknife test is considered the least arbitrary because it can agree to specific results for particularly obtained benchmark dataset as explained earlier in a study [31]. To study the statistical analysis of the new predictor, a comparison was made using the jackknife test with previous methodologies [8, 30]. In this study, all of these validation tests were employed to evaluate the quality of the proposed methodology. In addition, K-fold cross-validation test is based on sub-sampling to validate the classifier since several partitioning permutations exist therefore it cannot avoid ambiguity [8].

5. COMPARISON WITH PREVIOUS METHODS

Values given in Table **1** are the scores of the four metrics attained by the proposed predictor using the independent dataset test, 10-fold cross-validation test, and jackknife test on the dbptm benchmark dataset, while, Table **2** represents the scores of similar metrics using the most updated dataset obtained from UniProt. Furthermore, Table **3** shows a comparison with the existing techniques. Two existing predictors have been depicted, namely "iHyd-PseAAC" [8], and " iHyd-PseCp" [30], for identifying the hydroxyproline sites. These methods also achieved the metrics scores using the jackknife test method. It can be observed from Table **3** that the accuracy (Acc), stability (MCC), sensitivity (Sn), and specificity (Sp) scores evaluated by the newly proposed predictor are superior than those reported by the existing predictors. A comparison with previous methods was made using two benchmark datasets extracted from (a) dbptm and (b) uniprot database. To understand the complex biological systems, the graphical representation gives a valuable vision as represented by the list of earlier articles [48-50]. The same is depicted as a comparison in graphical representation showing the Receiver Operating Characteristic (ROC) [51] of the proposed predictor and previously existing predictors. In Fig. (**2**), the red curve represents the ROC curve for iHyd-PseAAC and green curve for iHyd-PseCp, while blue solid and dotted curves represent the ROC plotted by using the proposed predictor on dbptm and uniprot benchmark datasets. It is evident from the figure below that the area under the blue dotted and solid curves is extraordinarily larger than that under the red and green curves. Undoubtedly, the novel proposed predictor is certainly an improved approach over the existing predictors.

The superior performance of the proposed system can be rationalized by a number of scientific and theoretical reasons. Some of these are discussed here. Firstly, the proposed model is a formulation based on the composition and sequence of primary structure which can conveniently handle diverse length sequences in a generous way without

Table 1. Three tests result on set of metrics using proposed model on dbptm benchmark.

Tests	Sn (%)	Sp (%)	Acc (%)	MCC
Independent dataset test	98.30	98.02	98.77	0.96
Cross-Validation	98.73	94.87	96.85	0.93
Jackknife test	98.68	94.82	96.80	0.90

Table 2. Three tests result on four metrics using proposed model on recent uniprot benchmark.

Tests	Sn (%)	Sp (%)	Acc (%)	MCC
Independent dataset test	98.38	99.54	98.80	0.95
Cross-Validation	97.07	94.62	96.06	0.91
Jackknife test	97.02	94.57	96.01	0.88

Table 3. A comparison of the proposed model with the previous methods to identify hydroxylation of proline using jackknife test in the validation of benchmark datasets extracted from (a) dbptm and (b) uniprot.

Predictors	Sn (%)	Sp (%)	Acc (%)	MCC
iHyd-PseAAC	80.66	80.54	80.57	0.51
iHyd-PseCp	86.35	99.12	96.58	0.89
iHyd-PseAAC (EPSV)[a]	98.68	94.82	96.80	0.90
iHyd-PseAAC (EPSV)[b]	97.02	94.57	96.01	0.88

skipping any obscure information and form pairwise couplings in every possible permutation of amino acid residues. Secondly, it generates a fixed length vector, which imparts a non-variable size feature vector that equally separates proteins according to their attributes. This aspect enables the predictor to rigorously classify and conveniently recognize each sample. Thirdly, the correlation expression is the main mechanism that contributes towards the computation of a feature vector. It has been configured by incorporating each attribute group. Each expression deals with some specific metric and statistical expressions. For the sake of convenience, every property of amino acids was standardized numerically within a suitable range. Also, it has been observed that in comparison with previous methods proposed, the predictor outcomes are more superior and better than the former prediction rate.

6. WEB-SERVER

User-friendly and publicly accessible web-servers represent the current trend for developing various computational methods [52], as reflected by a series of recent publications [32, 33, 35, 36, 44]. Actually, they have significantly enhanced the impacts of computational biology in medical science [53], driving medicinal chemistry into an unprecedented revolution [54], here we shall do our best to provide a web-server for the predictor presented in this paper as soon as possible.

7. DISCUSSION

The proposed model is a new predictor to identify hydroxylation of proline. It can be analysed from Table **3** that the accuracy calculated for the proposed model is **96.80** and **96.01** which is higher than the accuracy calculated using previous predictors, that is 80.57 and 96.58. Also, MCC values were 0.90 and 0.88 which were superior to both the predictors *i.e.* iHyd-PseAAC and iHyd-PseCp. The proposed model was validated using benchmark datasets extracted from dbptm as well as from UniProt database.

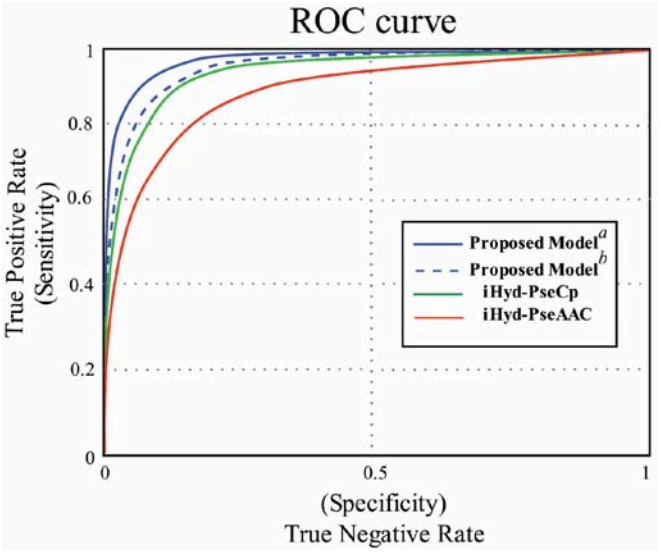

Fig. (2). Comparison of the proposed model with the curves plotted with iHyd-PseAAC and iHyd-PseCp predictors.

8. METHODS

8.1. Benchmark Dataset

According to Chou's 5-step rule [31], the extraction of benchmark dataset is a crucial step that leads to the acquisition of a robust, diverse and updated dataset. In this study, a stringent benchmark dataset has been borrowed from two roots. One of the datasets is received from the resource http://www.uniprot.org/, and the other is leased from a post-translational modification database dbPTM 3.0 [55] that has also been utilized by Xu *et al.* [8]. The following two steps are used to select a stringent benchmark dataset.

Step-1: The data extracted from UniProt database, consists of positive and negative samples that represent the hydroxylated and non-hydroxylated polypeptide sequences at proline site. A query is generated to select protein sequences in the PTM/processing field as hydroxyproline. Entries annotated with any experimental assertion in Feature Table (FT) were exclusively selected.

Step-2: After a rigorous adoption of the above step, a first-rate benchmark dataset of hydroxyproline was collected. Total samples of 816 and 24,980 for positive and negative were extracted, respectively. After obtaining the duplicates, both were cut down to 782 and 24971 unique values. For the sake of convenience, $\mathbb{R}_{\eta^+}$ and $\mathbb{R}_{\eta^-}$ represent the positive and negative set of the hydroxylated polypeptides, respectively. Further, let $R = \mathbb{R}_{\eta^+} + \mathbb{R}_{\eta^-} = 782 + 24971$ be the total sum of these two. Also, it can be easily seen that there exist more negative peptides than positive peptides in nature. Thus, $\mathbb{R}_{\eta^-} \gg \mathbb{R}_{\eta^+}$. Similarly, to extract another stringent benchmark dataset, the dbdtm 3.0 [55] was employed. The dataset was easily available in FASTA format and conveniently were downloaded for hydroxylation (positive and negative). There were found 226 positive sets and 3,865 negative sets. A demonstration in term of a Flowchart is given in Fig. (**3**), to understand the above steps. The primary structure of hydroxylated and non-hydroxylated proline sites can be found in Supplementary Tables **S1**, **S2**, **S3** and **S4** respectively.

9. SAMPLE FORMULATION AND ALGORITHM DEVELOPMENT

According to the Chou's second and third step [31], a powerful mathematical formulation is proposed that can accurately reflect their indispensable correlation to arrange the sample in an effective way, also used by Ehsan *et al.* [40]. Considering a protein sample **P,** consisting of L amino acid residues.

$$P = \eta_1\eta_2\eta_3\eta_4\eta_5\eta_6\eta_7 \cdots \eta_L \tag{7}$$

Where η_1 is the first amino acid residue, η_2 is the second amino acid residue, and so on up to η_L, the last residue of protein sequence **P**, where L indicates the length of the sequence (7). To identify the post translational modification in proline site, a computational methodology has been persuaded. This method upholds the sequence order effect and is adopted using the whole sequence data together with the occurrence of each amino acid residue $\lambda_{\hat{b}}$ of type $\hat{b}: 1 \leq \hat{b} \leq 20$ (any one of the residues among twenty amino acid residues). Expression (8) to (11) describes the whole formulation strategy. The number of occurrences $\lambda_{\hat{b}}$ of residue $\eta_{\hat{b}}$ and the possible number of correlated factors ν of $\hat{b}$ with itself, such that $(\lambda_{\hat{b}} - 1)!\, \nu(\eta_{\hat{b}}, \eta_{\hat{b}})$ is linked to expression (8). While, mean factors M_0, M_i and M_j are connected with the deviation factors of $\eta_{\hat{b}}$ at their respective positions and are represented by the expression (9), followed by condition (10). Whereas, M_i runs over deviation factors and these factors are linked by a local mean. This deviation is denoted by $(q-p)_i$, provided the positions of $\hat{b}$ are labeled by p and q in $\eta_p = \eta_q = \eta_{\hat{b}}$, the polypeptide chain. While the subscript i denotes the frequency of occurrence of deviation factors for similar amino acid residues discarding the occurrence at the first and last position residue $\hat{b}$, based on n total occurrences of $\hat{b}$; similarly, M_j is labeled for the difference, $L - r$, and r represent the exact position of the residue $\hat{b}$ appearing at n_{th} of its occurrence in (7) while $\eta_r = \eta_{\hat{b}}$ and $1 \leq p < q < r \leq L$ denote the $\hat{b}$ amino acid residues in the corresponding positions.

$$\lambda_{\hat{b}} + (\lambda_{\hat{b}} - 1)!\ \ \nu(\eta_{\hat{b}}, \eta_{\hat{b}}) \tag{8}$$

$$[(p-0)_0 M_0 + \textstyle\sum_{q>p} (q-p)_i M_i + (L-r)_n M_{j=n}];\ \ 1 \leq p < q < r \leq L,\ \ i = 1,2,3,\ldots,n-1 \tag{9}$$

$$(9) \Rightarrow \begin{cases} \sum_{q>p} (q-p)_i M_i + (L-r)_n M_{j=n},\ \ if\ \ 1 \leq p < q, r < L \\ (p-0)_0 M_0 + \sum_{q>p} (q-p)_i M_i + (L-r)_n M_{j=n},\ \ if\ \ 1 < p < q, r < L \\ (p-0)_0 M_0 + \sum_{q>p} (q-p)_i M_i,\ \ if\ \ 1 < p < q, r = L \end{cases} \tag{10}$$

Combining expressions (8) and (9) and using constraint (10) yield the template for manipulating feature component related to $\hat{b}$, given in (11).

$$\lambda_{\hat{b}} + (\lambda_{\hat{b}} - 1)!\ \nu(\eta_{\hat{b}}, \eta_{\hat{b}}) + [(p-0)_0 M_0 + \textstyle\sum_{q>p} (q-p)_i M_i + (L-r)_n M_{j=n}],\ \ i\ \ = 1,2,3,\ldots,n-1 \tag{11}$$

While M_i and M_j denote the number occurrences of $\hat{b}$ before and after, with the remaining residues, respectively.

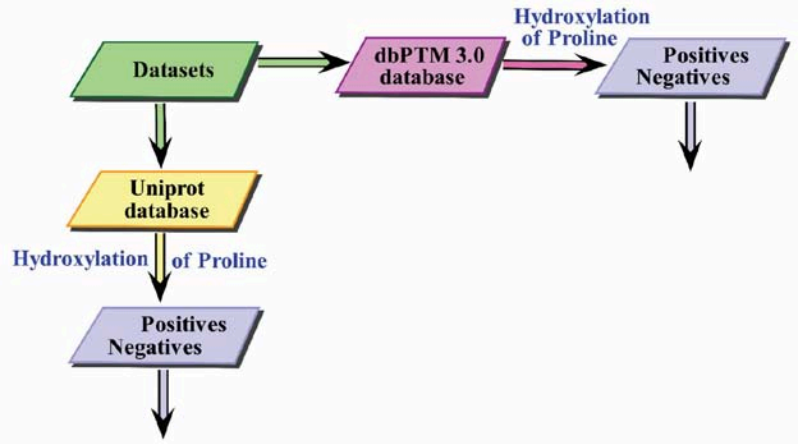

Fig. (3). Flowchart is representing the database sources used to retrieve the datasets.

These are given in eq (12) and (13).

$$M_i = \{X_i + Y_i\}, \qquad i = 1,2,3,\ldots,n-1$$
$$M_{j=n} = \{X + Y\} \tag{12}$$

Where

$$X = X_i = \frac{1}{38}[\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_0 \nu(\eta_{\hat{b}},\eta_\iota)]$$
$$Y = Y_i = \frac{1}{38}[\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_0 \nu(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_{\hat{b}},\eta_\iota)] \tag{13}$$

Where $f_\iota, 1 < \iota < 20$ shows the occurrence of binary function ν related to residue $\hat{b}$ with any of the remaining nineteen residues and f_0 stands for none of its occurrence with others. $\varsigma_{s,t}$, is defined as the pair function Â£h for all combinations of all residues, whereas the pair function $\nu(\eta_s,\eta_t)$ in terms of $\varsigma_{s,t}$ is defined as $\nu(\eta_s,\eta_t) = \varsigma_{s,t}; s = t = 1,2,3\ldots,20$, elaborated in a matrix (14). Equation (15) assigns all the possible pair factors concerning X and Y together with (16). If a pair $\nu(\eta_s,\eta_t)$ is found, then $\varsigma_{s,t}$ is labeled as 1 otherwise it will be assigned 0 value. Additionally, (15) admits to (14) with entries $\varsigma_{s,t}$, $\varsigma_{s,s}$ and $\varsigma_{t,s}$ specifying lower triangular matrix for X. Accordingly, the diagonal entries signify the combination among analogous residues and upper triangular matrix for Y.

$$\begin{pmatrix} \varsigma_{1,1} & \varsigma_{1,2} & \varsigma_{1,3} & \cdots & \varsigma_{1,20} \\ \varsigma_{2,1} & \varsigma_{2,2} & \varsigma_{2,3} & \cdots & \varsigma_{2,20} \\ \varsigma_{3,1} & \varsigma_{3,2} & \varsigma_{3,3} & \cdots & \varsigma_{3,20} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varsigma_{20,1} & \varsigma_{20,2} & \varsigma_{20,3} & \cdots & \varsigma_{20,20} \end{pmatrix} \tag{14}$$
$$\begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 1 \\ 1 & 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 0 \end{pmatrix} = M_i = M_j = \{X_i + Y_i\} = \{X + Y\} =$$

$$\begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 0 & 0 & 1 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \tag{15}$$

Where

$$\varsigma_{s,t} = \begin{cases} 1 & ,when\ \nu(\eta_s,\eta_t)\ exists\ for\ both\ s = t\ or\ s \neq t \\ 0 & ,otherwise \end{cases} \tag{16}$$

The manipulation of feature components in a matrix form, incorporating all the amino acid residues given in (17) can be viewed as an extension of (11).

$$\lambda_{\hat{b}} + (\lambda_{\hat{b}} - 1)! \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} + \frac{1}{38}[(p-0)_0 M_0 + \sum_{q>p}$$
$$(q-p)_i \begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 1 \\ 1 & 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 0 \end{pmatrix} +$$
$$(L-r)_n \begin{pmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 1 \\ 1 & 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 0 \end{pmatrix}] \tag{17}$$

Expression (11) together with equation (13) yields the component of feature vector that is $\hat{b}$, elaborated in eq (18) and (19).

$$\Gamma_{\hat{b}} =$$
$$\lambda_{\hat{b}} + (\lambda_{\hat{b}} - 1)!\ \ \nu(\eta_{\hat{b}},\eta_{\hat{b}}) + \frac{1}{38}[(p-0)_0\{\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_\iota,\eta_{\hat{b}}) +$$
$$\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_{\hat{b}},\eta_\iota)\}_0 + \{(q-p)_1\{\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_\iota,\eta_{\hat{b}}) +$$
$$\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_{\hat{b}},\eta_\iota)\}_1 + (q-p)_2\{\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_\iota,\eta_{\hat{b}}) +$$
$$\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_{\hat{b}},\eta_\iota)\}_2 + (q-p)_3\{\sum_{\substack{\iota=1 \\ \iota\neq\hat{b}}}^{20} f_\iota \nu(\eta_\iota,\eta_{\hat{b}}) +$$

$$\sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_{\hat{b}},\eta_\iota)\}_3+\ldots+(q-p)_{n-1}\{\sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_{\hat{b}},\eta_\iota)\}_{n-1}\} + (L-r)_n\{\sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_{\hat{b}},\eta_\iota)\}_n] \quad (18)$$

Or

$$\Gamma_{\hat{b}} = \lambda_{\hat{b}} + (\lambda_{\hat{b}}-1)!\ \ v(\eta_{\hat{b}},\eta_{\hat{b}}) + \frac{1}{38}[(p-0)_0\{\sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_{\hat{b}},\eta_\iota)\}_0 + \sum_{q>p}(q-p)_i\{\sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_{\hat{b}},\eta_\iota)\}_i + (L-r)_n\{\sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_\iota,\eta_{\hat{b}}) + \sum_{\substack{\iota=1\\ \iota\neq\hat{b}}}^{20} f_\iota v(\eta_{\hat{b}},\eta_\iota)\}_n],\ i=1,2,3,\ldots,n-1. \quad (19)$$

The structural scheme of the proposed formulation can be understood by considering l_{th} term of a sequence (7), say, η_l, which mirrors the amino acid residues say "A". It must be noted that η_l makes a pair with its adjacent residues before and after the l_{th} residue in terms of $v(\eta_\iota,\eta_l)$ and $v(\eta_l,\eta_\iota)$ exemplified by blue and pink curvy lines and pairs η_l with itself which is denoted by muddy green loops as shown in Fig. **4**. The procedure must be followed till η_m appears in m_{th} place such that $\eta_l=\eta_m=A$. Correspondingly, a similar procedure will be adopted for η_m. The feature component agreeing to residue "A" is substituted in equation (20).

$$\Gamma_A = \lambda_A + (\lambda_A-1)!\,v(A,A) + \frac{1}{38}[(p_l-0)_0\{\sum_{\substack{\iota=1\\ \iota\neq A}}^{20} f_\iota v(\eta_\iota,A) + \sum_{\substack{\iota=1\\ \iota\neq A}}^{20} f_\iota v(A,\eta_\iota)\}_0 + (q_m-p_l)\{\sum_{\substack{\iota=1\\ \iota\neq A}}^{20} f_\iota v(\eta_\iota,A) + \sum_{\substack{\iota=1\\ \iota\neq A}}^{20} f_\iota v(A,\eta_\iota)\} + (L-r_m)\{\sum_{\substack{\iota=1\\ \iota\neq A}}^{20} f_\iota v(\eta_\iota,A) + \sum_{\substack{\iota=1\\ \iota\neq A}}^{20} f_\iota v(A,\eta_\iota)\}] \quad (20)$$

Where $\iota=1,2,3\ldots,20$ are the amino acid residues in ascending order. For simplicity, taking $\eta_1,\eta_2,\eta_3,\ldots,\eta_{20}$ as the 20 amino acids in an alphabetical order for further generalization and η_{21} onwards the 20 residues that periodically replicate themselves. Supposing $\Gamma_1,\Gamma_2,\Gamma_3,\ldots,\Gamma_{20}$ are their associate feature components. These are given in equation (21).

$$\Gamma_1 = \lambda_1 + (\lambda_1-1)!\ \ v(\eta_1,\eta_1) + \frac{1}{38}[(p-0)_0\{\sum_{\substack{\iota=1\\ \iota\neq 1}}^{20} f_\iota v(\eta_\iota,\eta_1) + \sum_{\substack{\iota=1\\ \iota\neq 1}}^{20} f_\iota v(\eta_1,\eta_\iota)\}_0 + \sum_{q>p}(q-p)_i\{\sum_{\substack{\iota=1\\ \iota\neq 1}}^{20} f_\iota v(\eta_\iota,\eta_1) + \sum_{\substack{\iota=1\\ \iota\neq 1}}^{20} f_\iota v(\eta_1,\eta_\iota)\}_i + (L-r)_n\{\sum_{\substack{\iota=1\\ \iota\neq 1}}^{20} f_\iota v(\eta_\iota,\eta_1) + \sum_{\substack{\iota=1\\ \iota\neq 1}}^{20} f_\iota v(\eta_1,\eta_\iota)\}_n],\quad i=1,2,3,\ldots,n-1$$

$$\Gamma_2 = \lambda_2 + (\lambda_2-1)!\ \ v(\eta_2,\eta_2) + \frac{1}{38}[(p-0)_0\{\sum_{\substack{\iota=1\\ \iota\neq 2}}^{20} f_\iota v(\eta_\iota,\eta_2) + \sum_{\substack{\iota=1\\ \iota\neq 2}}^{20} f_\iota v(\eta_2,\eta_\iota)\}_0 + \sum_{q>p}(q-p)_i\{\sum_{\substack{\iota=1\\ \iota\neq 2}}^{20} f_\iota v(\eta_\iota,\eta_2) + \sum_{\substack{\iota=1\\ \iota\neq 2}}^{20} f_\iota v(\eta_2,\eta_\iota)\}_i + (L-r)_n\{\sum_{\substack{\iota=1\\ \iota\neq 2}}^{20} f_\iota v(\eta_\iota,\eta_2) + \sum_{\substack{\iota=1\\ \iota\neq 2}}^{20} f_\iota v(\eta_2,\eta_\iota)\}_n],\quad i=1,2,3,\ldots,n-1$$

$$\Gamma_3 = \lambda_3 + (\lambda_3-1)!\ \ v(\eta_3,\eta_3) + \frac{1}{38}[(p-0)_0\{\sum_{\substack{\iota=1\\ \iota\neq 3}}^{20} f_\iota v(\eta_\iota,\eta_3) + \sum_{\substack{\iota=1\\ \iota\neq 3}}^{20} f_\iota v(\eta_3,\eta_\iota)\}_0 + \sum_{q>p}(q-p)_i\{\sum_{\substack{\iota=1\\ \iota\neq 3}}^{20} f_\iota v(\eta_\iota,\eta_3) + \sum_{\substack{\iota=1\\ \iota\neq 3}}^{20} f_\iota v(\eta_3,\eta_\iota)\}_i + (L-r)_n\{\sum_{\substack{\iota=1\\ \iota\neq 3}}^{20} f_\iota v(\eta_\iota,\eta_3) + \sum_{\substack{\iota=1\\ \iota\neq 3}}^{20} f_\iota v(\eta_3,\eta_\iota)\}_n],\quad i=1,2,3,\ldots,n-1$$

$$\vdots$$

$$\Gamma_{20} = \lambda_{20} + (\lambda_{20}-1)!\ v(\eta_{20},\eta_{20}) + \frac{1}{38}[(p-0)_0\{\sum_{\substack{\iota=1\\ \iota\neq 20}}^{20} f_\iota v(\eta_\iota,\eta_{20}) + \sum_{\substack{\iota=1\\ \iota\neq 20}}^{20} f_\iota v(\eta_{20},\eta_\iota)\}_0 + \sum_{q>p}(q-p)_i\{\sum_{\substack{\iota=1\\ \iota\neq 20}}^{20} f_\iota v(\eta_\iota,\eta_{20}) + \sum_{\substack{\iota=1\\ \iota\neq 20}}^{20} f_\iota v(\eta_{20},\eta_\iota)\}_i + (L-r)_n\{\sum_{\substack{\iota=1\\ \iota\neq 20}}^{20} f_\iota v(\eta_\iota,\eta_{20}) + \sum_{\substack{\iota=1\\ \iota\neq 20}}^{20} f_\iota v(\eta_{20},\eta_\iota)\}_n],\quad i=1,2,3,\ldots,n-1 \quad (21)$$

The three main characteristics of amino acids, that is hydrophobicity, hydrophilicity and side chain mass of amino acids mainly take part in the above set of twenty feature components. Every characteristic relates 60 entries as coordinates,which contribute to 180 coordinates in total influenced by equation (22) to (24), $w=1,2,3$ identifies the characteristics.

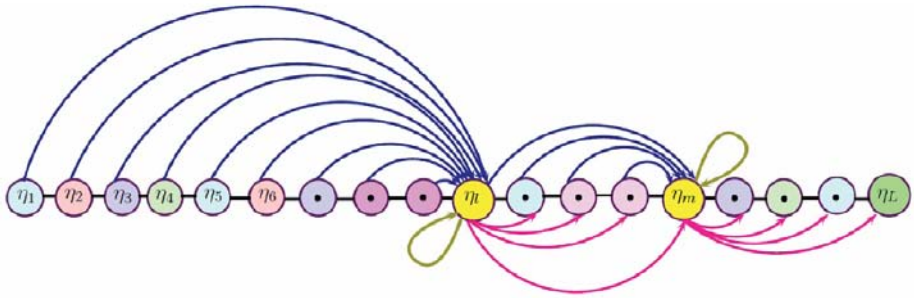

Fig. (4). Graphical representation shows how to formulate the sequence for classification.

$$\nu(\eta_s,\eta_t)=\sqrt{\Omega_w^*(\eta_s)^2|\Omega_w^*(\eta_s)-\Omega_w^*(\eta_t)|^2}+\frac{|(\Omega_1^*(\eta_s)-\overline{\Omega}_1^*(\hat{b}))(\Omega_2^*(\eta_t)-\overline{\Omega}_2^*(\hat{b})|}{\sqrt{\sum_{s=1}^{20}(\Omega_1^*(\eta_s)-\overline{\Omega}_1^*(\hat{b}))^2\sum_{t=1}^{20}(\Omega_2^*(\eta_t)-\overline{\Omega}_2^*(\hat{b}))^2}} \quad (22)$$

$$\nu(\eta_s,\eta_t)=\sqrt{\Omega_w^*(\eta_s)^2|\Omega_w^*(\eta_s)-\Omega_w^*(\eta_t)|^2}+\frac{|(\Omega_1^*(\eta_s)-\overline{\Omega}_1^*(\hat{b}))(\Omega_3^*(\eta_t)-\overline{\Omega}_3^*(\hat{b})|}{\sqrt{\sum_{s=1}^{20}(\Omega_1^*(\eta_s)-\overline{\Omega}_1^*(\hat{b}))^2\sum_{t=1}^{20}(\Omega_3^*(\eta_t)-\overline{\Omega}_3^*(\hat{b}))^2}} \quad (23)$$

$$\nu(\eta_s,\eta_t)=\sqrt{\Omega_w^*(\eta_s)^2|\Omega_w^*(\eta_s)-\Omega_w^*(\eta_t)|^2}+\frac{|(\Omega_2^*(\eta_s)-\overline{\Omega}_2^*(\hat{b}))(\Omega_3^*(\eta_t)-\overline{\Omega}_3^*(\hat{b})|}{\sqrt{\sum_{s=1}^{20}(\Omega_2^*(\eta_s)-\overline{\Omega}_2^*(\hat{b}))^2\sum_{t=1}^{20}(\Omega_3^*(\eta_t)-\overline{\Omega}_3^*(\hat{b}))^2}} \quad (24)$$

Where Ω_1^*,Ω_2^*, andΩ_3^* represent the normalized hydrophobicity, hydrophilicity and side-chain mass, respectively, and $\overline{\Omega_1}^*$, $\overline{\Omega_2}^*$, and $\overline{\Omega_3}^*$ indicate the mean of the normalized values corresponding to the 20 amino acids $\hat{b}$ related to w attributes. The values used in (22) to (24) are normalized by using (25), and standardized in a range (-T, T), where T is the count for $\hat{b}$ amino acids to be standardized. Entries for hydrophobicity are picked from Tanford C. [56], and for hydrophilicity, entries are taken from Hopp T.P., Woods K.R. [57], while the values of side-chain mass can be found in most of the books given in the bibliography.

$$\Omega_1^*(\hat{b})=[\frac{2T}{(\Omega_{1_{(max)}}-\Omega_{1_{(min)}})}(\Omega_1(\hat{b})-\Omega_{1_{(max)}})]+T$$

$$\Omega_2^*(\hat{b})=[\frac{2T}{(\Omega_{2_{(max)}}-\Omega_{2_{(min)}})}(\Omega_2(\hat{b})-\Omega_{2_{(max)}})]+T$$

$$\Omega_3^*(\hat{b})=[\frac{2T}{(\Omega_{3_{(max)}}-\Omega_{3_{(min)}})}(\Omega_3(\hat{b})-\Omega_{3_{(max)}})]+T \quad (25)$$

The feature set is categorized into a vector with 220 components, of which, the first sixty are constructed by virtue of the hydrophobic nature of amino acids, the next sixty components depict their hydrophilic nature, the subsequent sixty components are related to side chain mass, whereas the last forty reflect the position and composition of each amino acid residue. The feature vectors hence obtained for the training data are clamped to a neural network for training. Once the training is completed , the trained network apparently gains the experience to categorize arbitrary input with an appreciable precision. While the process is carried on, the network normalizes its weights with a minimum slip.

Multilayer Perceptron (MLP) is an excellent model that can uncover and identify obscure patterns in diversified data sets. MLP is best suited for any classification problem as it can be fine tuned by changing the number of hidden layer neurons, training parameters and training algorithm to provide the best outcome. A Multi-Layer Perceptron (MLP) was trained using the extracted feature set for this purpose (Fig. **5**). The feature vectors for the samples were assembled into a large array. Each row of the array represents the feature vector for a single sequence while each column represents a feature item extracted. Since 220 features were extracted for each sample; therefore each row had 220 columns while the total columns were 25796; out of which, 816 were positive samples.The weights of each layer were initialized randomly while a hidden layer with 75 neurons was used. Further, back propagation algorithm was used to adjust the weights after each epoch. Convergence was achieved after 2693 iterations while using gradient descent method for learning rate.

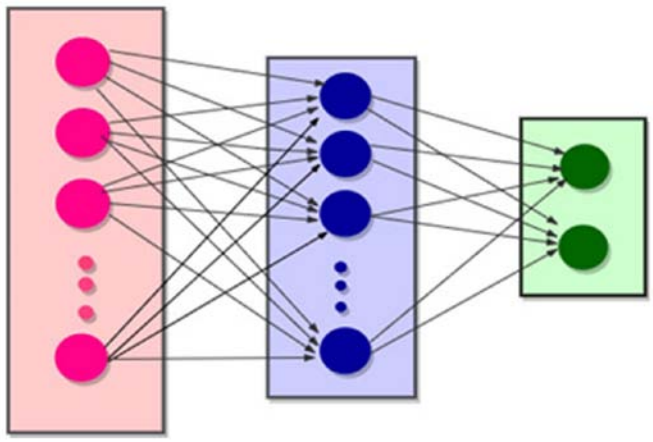

Fig. (5). neural network.

The results were simulated on MATLAB R2017 version and were duplicated on python ver 3.6 platform along with Scikit Learn 0.20 for neural network training and simulation bearing identical results.

The algorithm which is developed by the following above method is called iHyd-PseAAC (EPSV), where "i" represents the first word of "identify", Hyd is used for "hydroxylation" and Pse-AAC is the general term used for pseudo amino acid composition. Also the term "EPSV" stands for "enhanced position and sequence variant" technique which is used to construct an algorithm for polypeptide sequence.

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.

AUTHORS' CONTRIBUTIONS

A.E. proposed an algorithm and conducted experiments, K.M. supervised the results. Y.D. supervised the validation process. S.K. worked on the appropriateness of the suggested model. O.M. worked on the data collection and feature extraction. K.C. originally formulated the problem contributed equally to analyze and improve the results. All authors reviewed the manuscript.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available in the Kaggle at www.kaggle.com, reference number https://www.kaggle.com/ydkhan/starter-proline-hydroxylation-f4c3c873-a".

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 web site along with the published article.

REFERENCES

[1] Colgrave, M.L.; Peter G.A; and Jones, A. Hydroxyproline quantification for the estimation of collagen in tissue using multiple reaction monitoring mass spectrometry. *J. Chromatogr. A.*, **2008**, *1212*(1-2), 150-153.

[2] Gelse, K.; Pöschl, E.; and Aigner, T. Collagens—structure, function, and biosynthesis. *Adv. Drug Deliv. Rev.*, **2003**, *55*(12), 1531-1546.

[3] Ruszczak, Zbigniew. Effect of collagen matrices on dermal wound healing. *Adv. Drug Deliv. Rev.*, **2003**, *55*(12), 1595-1611.

[4] Lee, C.H.; Singla, A.; and Lee, Y. Biomedical applications of collagen. *Int. J. Pharm.*, **2001**, *221*(1-2), 1-22.

[5] Becker, G.D.; Lawrence A.A; and Hackett, J. Collagen-assisted healing of facial wounds after mohs surgery. *Laryngoscope*, **1994**, *104*(10), 1267-1270.

[6] Guszczyn, T.; Soboleweki, K. Deregulation of collagen metabolism in human stomach cancer. *Pathobiology*, **2004**, *71*(6), 308-313.

[7] Sunila, E.S., and Kuttan, G. A preliminary study on antimetastatic activity of Thuja occidentalis L. in mice model. *Immunopharmacol. Immunotoxicol.*, **2006**, *28*(2), 269-280.

[8] Xu, Y.; Wen, X.; Shao, X. J.; Deng, N.Y.; and 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.

[9] Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; and Chou, K.C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*, **2019**, *111*(1), 96-102.

[10] Xu, Y.; Ding, J.; Wu, L.Y.; and Chou, K.C. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition *PLoS one* , **2013**, *8*(2), e55844.

[11] Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; and Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *Peer J.*, **2013**, 1, e171.

[12] Jia, C.; Lin, X.; and Wang, Z. Prediction of protein s-nitrosylation sites based on adapted normal distribution bi-profile bayes and chou's pseudo amino acid composition. *Int. J. Mol. Sci.*, **2014**, *15*(1), 10410-10423.

[13] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; and Chou, K.C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.*, **2016**, *394*(1), 223-230.

[14] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; and Chou, K.C. iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget*, **2016**, *7*(23), 34558-34570.

[15] Jia, J.; Zhang, L.; Liu, Z.; Xiao, X.; and Chou, K.C. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics*, **2016**, *32*(1), 3133-3141.

[16] Khan, Y.D.; Rasool, N.; Hussain, W.; Khan, S.A.; and Chou, K.C. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Anal. Biochem.*, **2018**, *550*(1), 109-116.

[17] Khan, Y.D.; Rasool, N.; Hussain, W.; Khan, S.A.; Chou, K.C. iPhosY-PseAAC: Identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. *Mol. Biol. Rep.*, **2018**, *550*(109-116), (doi: 10.1016/j.ab.2018.04.021).

[18] 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.

[19] Ang, K.S.; Lakshmanan, M.; Lee, N.R.; Lee, D.Y. Metabolic modeling of microbial community interactions for health, environmental and biotechnological applications. *Curr. Genom.*, **2018**, *19*(8), 712-722.

[20] Berg, R.A.; Steinmann, B.; Rennard, S.I.; and Crystal, R.G. Ascorbate deficiency results in decreased collagen production: under-hydroxylation of proline leads to increased intracellular degradation. *Arch. Biochem. Biophys.*, **1983**, *226*(2), 681-686.

[21] Halme, J.; Kivirikko, K.I.; and Simons, K. Isolation and partial characterization of highly purified protocollagen proline hydroxylase. *Biochim. Biophys. Acta.* **1970**, *198*(3), 460-470.

[22] Kivirikko, K.I.; and Prockop, D.J. Hydroxylation of proline in synthetic polypeptides with purified protocollagen hydroxylase. *J. Biol. Chem.*, **1967**, *242*(18), 4007-4012.

[23] Morgan, A.A.; and Rubenstein, E. Proline: The distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. *PloS One*, **2013**, *8*(1), e53785.

[24] Yamauchi, M.; and Shiiba, M. Lysine hydroxylation and crosslinking of collagen, In: *Posttranslational modifications of proteins*; *Humana Press*: New York, **2002**, 277-290.

[25] Shi, S.P.; Chen, X.; Xu, H.D.; and Qiu, J.D. PredHydroxy: Computational prediction of protein hydroxylation site locations based on the primary structure. *Mol. Biosyst.*, **2015**, *11*(3), 819-825.

[26] Wu, G.; Bazer, F.W.; Burghardt, R.C.; Johnson, G.A.; Kim, S.W.; Knabe, D.A.; Li, P.; Li, X.; McKnight, J.R.; Satterfield, M.C.; Spencer, T.E. Proline and hydroxyproline metabolism: Implications for animal and human nutrition. *Amino acids*, **2011**, *40*(4), 1053-1063.

[27] Hayat, S.; Hayat, Q.; Alyemeni, M.N.; Wani, A.S.; Pichtel, J.; Ahmad, A. Role of proline under changing environments: A review. *Plant Sig. Behav.*, **2012**, *7*(11), 1456-1466.

[28] Yang, Z.R. Predict collagen hydroxyproline sites using support vector machines. *J. Comput. Biol.*, **2009**, *16*(5), 691-702.
[29] Hu, L.L.; Niu, S.; Huang, T.; Wang, K.; Shi, X.H.; and Cai, Y.D. Prediction and analysis of protein hydroxyproline and hydroxylysine. *PLoS One*, **2010**, *5*(12), e15917.
[30] 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.
[31] Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.*, **2011**, *273*(1), 236-247.
[32] Cheng, X.; Xiao, X.; Chou, K.C. pLoc-mPlant: Predict subcellular localization of multi-location plant proteins *via* incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.*, **2017**, *13*(1), 1722-1727.
[33] Xiao, X.; Cheng, X.; Su, S.; Mao, Q.; Chou, K.C. pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins. *Nat. Sci.*, **2017**, *9*(1), 331-349.
[34] Wang, J.; Li, J.; Yang, B.; Xie, R.; Marquez-Lago, T.T.; Leier, A.; Hayashida, M.; Akutsu, T.; Zhang, Y.; Chou, K.C.; Selkrig, J.; Zhou, T.; Song, J.; Lithgow, T. Bastion3: A two-layer approach for identifying type III secreted effectors using ensemble learning. *Bioinformatics*, **2018**, doi: 10.1093/bioinformatics/xxxxx.
[35] Chou, K.C.; Cheng, X.; and Xiao, X. pLoc-bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. *Genomics*, **2018**, (doi: 10.1016/j.ygeno.2018.08.007).
[36] Xiao, X.; Cheng, X.; Chen, G.; Mao, Q. pLoc-bal-mGpos: Predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. *Genomics*, **2018**, doi:10.1016/j.ygeno.2018.05.017.
[37] Khan, Y.D.; Jamil, M.; Hussain, W.; Rasool, N.; Khan, S.A.; Chou, K.C. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *J. Theor. Biol.*, **2019**, *463*(1), 47-55.
[38] Jia, J.; Li, X.; Qiu, W.; Xiao, X.; and Chou, K.C. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. *J. Theor. Biol.*, **2019**, *460*(1), 195-203.
[39] Chen, J.; Liu, H.; Yang, J.; and Chou, K.C. Prediction of linear b-cell epitopes using amino acid pair antigenicity scale. *Amino acids*, **2007**, *33*(1), 423-428.
[40] Ehsan, A.; Mahmood, K.; Khan, Y. D.; Khan, S. A.; and Chou, K.C. A novel modeling in mathematical biology forclassification of signal peptides. *Sci. Reports*, **2018**, *8*(1), 1039.
[41] Chou, K.C. Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct., Funct., Genet.*, **2001**, 42, 136-139.
[42] Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.*, **2001**, *14*(1), 75-79.
[43] Chou, K.C. Prediction of signal peptides using scaled window. *Peptides*, **2001**, *22*(1), 1973-1979.
[44] Cheng, X.; Xiao, X.; and Chou, K.C. pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning *via* general PseAAC. *Genomics*, **2018**, *110*(1), 231-239.
[45] Cheng, X.; Zhao, S.G.; Xiao, X.; and Chou, K.C. iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics*, **2017**, *33*(3), 341-346.
[46] Qiu, W.R.; Sun, B.Q.; Xiao, X.; Xu, Z.C.; and Chou, K.C. iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, **2016**, *32*(1), 3116-3123.
[47] Chou, K.C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. BioSyst.*, **2013**, 9, 1092-1100.
[48] Chou, K.C. Graphic rule for drug metabolism systems. *Curr. Drug Metab.*, **2010**, *11*(1), 369-378.
[49] Chou, K.C.; Lin, W.Z.; and Xiao, X. Wenxiang: A web-server for drawing wenxiang diagrams. *Nat. Sci.*, **2011**, *3*(1), 862.
[50] Wu, Z.C.; Xiao, X.; and Chou, K.C. 2d-mh: A web-server for generating graphic representation of protein sequences basedon the physicochemical properties of their constituent amino acids. *J. Theor. Biol.*, **2010**, *267*(1), 29-34.
[51] Davis J.; Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* ACM, **2006**, pp. 233-240.
[52] Chou, K.C.; Shen, H.B. Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.*, **2009**, *1*(1), 63-92.
[53] Chou, K.C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.*, **2015**, *11*(1), 218-234.
[54] Chou, K.C. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem.*, **2017**, *17*(1), 2337-2358.
[55] Lu, C.T.; Huang, K.Y.; Su, M.G.; Lee, T.Y.; Bretana, N.A.; Chang, W.C.; Chen, Y.J.; Chen, Y.J. and Huang, H.D. Dbptm 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.*, **2012**, *41*(1), 295-305.
[56] Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am.Chem. Soc.*, **1962**, *84*(1), 4240-4247.
[57] Hopp, T.P. and Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.*, **1981**, *78*(1), 3824-3828.