iHyd-ProSite: A novel Computational Approach for Identifying Hydroxylation Sites in Proline Via Mathematical Modeling

Muhammad Khalid Mahmood[1*☯], Asma Ehsan[1☯], Yaser Daanial Khan[2☯]

1 Department of Mathematics, University of the Punjab, Lahore, Punjab, Pakistan
2 Faculty of Information Technology, University of Management and Tecnology, Lahore, Punjab, Pakistan

☯These authors contributed equally to this work.
* khalid.math@pu.edu.pk

Abstract

In various cellular functions, post translational modifications (PTM) of protein play a vital role. The addition of certain functional group through a covalent bond to the protein induces PTM. The number of PTMs are identified which are closely linked with diseases for example cancer and neurological disorder. Hydroxylation is one of the PTM, modified proline residue within a polypeptide sequence. The defective hydroxylation of proline causes absences of ascorbic acid in human which produce scurvy, and many other dominant health issues. Undoubtedly, the prediction of hydroxylation sites in proline residues is of challenging frontier. The experimental identification of hydroxyproline site is quite difficult, high-priced and time-consuming. The diversity in protein sequences instigates to develop a computational tool to identify hydroxylated site within short time with excellent prediction accuracy to handle such proteomics problems. In this work a novel in silico predictor is developed through rigorous mathematical modeling to identify which site of proline is hydroxylated and which site is not? Then performance of the predictor was verified using three validations tests, namely self-consistency test, cross-validation test and jackknife test over the benchmark dataset. A comparison was established for jackknife test with the previous methods. In comparison with previous predictors the proposed tool is more accurate than the existing techniques. Hence this scheme is highly useful and inspiring in contrast to all previous predictors.

Introduction

In mammals collagens are extremely abundant protein comprised of proline modified residue during the chemical process such as hydroxylation and produce hydroxyproline [1]. Collagens are stringy and long in nature, most of the protein in mammals consists of almost a quarter part of collagen [2]. In the treatment of wound healing [3], burn and cosmetic surgeries [4, 5] collagen mainly works as a medicinal drug. Most of the dominant human diseases like stomach and lung cancer [6, 7] are closely related with the defects and irregularities in hydroxylation process. Thus the identification of hydroxyproline (HyP) sites in proteins gives valuable data helpful to both biomedical research and drug development [8]. Hydroxyproline obtained by the addition of a hydroxyl group (-OH) to the proline (P) residue modifies the CH group into the COH group [8] (see Fig. 1).

Fig 1. Conversion of proline (Pro) residue into hydroxyproline. The figure is to show that, hydroxylation action attaches the -OH group to proline (Pro) to convert CH group to COH and modify proline residue into hydroxyproline.

The number of scientists has been making their contributions [9–12] in order to understand the cellular biological process and finding out medicines for cancer and for various other diseases. The prediction of hydroxylation sites in the lab by using method mass spectrometry [13] is difficult to conduct, expensive and very lengthy process. Every day, the large number of protein sequences is collected in the data bank and to classify them according to their functional properties is a crucial. It is highly worthy to build an efficient computational predictor for the classification of targeted hydroxylation sites within polypeptide sequences with improved prediction accuracy. Many researchers have been developed a couple of methods in this regard. Still, all these previous methods are insufficient to incorporate all components of features vague in the polypeptide sequence that become difficult to get exact prediction. Many scientists had been shown their great interest in hydroxylation process. Colgrave, et al. [1] was computed quantification of hydroxyproline by using multiple reaction monitoring mass spectrometry. In order to understand the microbial activity and their communities, a mathematical model has been developed [14]. A system was defined to study the insufficiency of collagen in connective tissues that encountered by lack of ascorbic acid [15].

Halme et al. [16] and kiviriko et al. [17] were explained the separation and classification of extremely purified protocollagen proline hydroxylase as well as proline hydroxylation in synthetical proteins with pure procollagen hydroxylase. In human proteome, the functional character of proline and polyproline based on distribution, frequency and positioning was investigated by Morgan, et al. [18]. Yamauchi et al. [19] were elaborated the Hydroxylation of lysine and cross-linking of collagens. By using a position weight of 8 high-quality amino acid indices and via support vector machines, Shi, Shao-Ping, et al. [20] were proposed a novel technique named as PredHydroxy for the forecast of the proline and lysine hydroxylation locales. Moreover, the functional study of proline with mutable surroundings and the metabolism of proline, hydroxyproline were examined in [21, 22]. ZR Yang [23] developed a tool for the prediction of hydroxyproline sites by utilizing support vector machine. A sequence-based formulation for identifying hydroxyproline and hydroxylysine were developed by Hu, Le-Le, et al. [24]. Using dipeptide position and specific propensity into pseudo amino acid composition Xu, Yan, et al. [8] predicted hydroxyproline and hydroxylysine in proteins. Qiu, Wang-Ren, et al. [25] was suggested an enhanced method over this technique by assimilating a sequence coupled effect into general PseAAC.

Materials and methods

Benchmark Dataset

The acquiring of benchmark dataset is critical, as indicated by Chou's 5-step rule [26] that prompts the attaining of a powerful, assorted and improved dataset. In order to obtain a stringent benchmark dataset, two resources have been used in the current study. One of the supported datasets is obtained from the universal protein database http://www.uniprot.org/, while the other dataset is borrowed from a posttranslational modification database dbPTM 3.0 [42]. Thus, a stringent benchmark datasets are obtained by employing the following two steps.

Step-1: The extracted dataset from UniProt database, contains two sets of protein sequences. One of the set represents hydroxylated protein sequences at proline site and labeled as positive sample. Likewise, other set consists of non-hydroxylated protein sequences at proline site, tagged as negative sample. An inquiry is produced to choose polypeptide sequences in the PTM/processing field as hydroxyproline. Records construed with any experimental assertion in Feature Table (FT) were only chosen. After a thorough selection of the described query, a stringent benchmark dataset of hydroxyproline was obtained. There were found records of 816 and 24980 for hydroxylated and non-hydroxylated sequences. The records were reduced to 782 and 24971 respectively, after removing duplicates.

Step-2: Likewise, to obtain another stringent benchmark dataset the dbdtm 3.047 were utilized. The dataset was effectively accessible in FASTA format and advantageously were downloaded for hydroxylation (hydroxylated and non-hydroxylated). There were discovered 226 hydroxylated records and 3,865 non-hydroxylated samples. The primary dataset of hydroxylated and non-hydroxylated proline sites can be found in Supplementary Tables S1, S2, S3 and S4 separately.

Method

In order to identify target proline sites with hydroxylation, an excellent methodology is proposed as indicated in the Chou's second and third step [26]. This technique is developed by incorporating all indispensible components of polypeptide sequences that can perfectly indicate their correlation to assemble the sequence in an effective way. The alternate formulation was also employed by Ehsan et al. [31, 32], impart as prominent prediction rate in proteomics problems. Consider a protein sample $\mathbb{C}$ consists of Z amino acid residues.

$$\mathbb{C} = U_1U_2U_3U_4U_5U_6U_7 \cdots U_Z \tag{1}$$

Where U_1 indicates the first amino acid residue with in string $\mathbb{C}$, U_2 is the second amino acid component, U_3 is third component and so on up till U_Z last amino acid residue of the protein sequence $\mathbb{C}$ as given in (1). In this methodology the formulation is handled by upholding sequence information. By considering the sequence order effect and composition of each term of sequence an advantageous factor introduced, known as weight factor. This is used to maintain position and composition of all components of sequence and denoted by T_i. While the significant terms L, M, and N indicate the count factor of each residue with its contiguous residues in both forward and reverse direction. The weight factor T_i is characterized as: the product of the position with the occurrence of the each term of the sequence among the similar residues. This whole scheme is based on expression (2) to handle the diverse length of the polypeptide sequences. The term $\{L + M + N\}$ describes the weighted mean of all possible coupling between similar residues that is after the first residue and before it occurred again.

$$\frac{(T_1)\{L+M+N\} + (T_2)\{L+M+N\} + (T_3)\{L+M+N\} + ... + (T_n)\{L+M+N\}}{T_1 + T_2 + T_3 + ... + T_n} \tag{2}$$

Where the weight factors $T_1, T_2, T_3, ..., T_n$ depends upon the repeated terms of the sequence of type $\tilde{r} : 1 \leq \tilde{r} \leq 20$. Each T_i is estimated between the two consecutive terms T_i and T_{i+1} before $\tilde{r}$ and when it occurred again. All weight factors characterization of amino acid of type $\tilde{r}$ in terms of mathematical form is represented as: $\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i} : i = 1, 2, 3, ..., n$ (α represent the occurrence of residue of type $\tilde{r}$ at their

corresponding positions β in the sequence varies with i, represents n time appearance of $\tilde{r}$) and L, M, N is the estimated count of the three correlated factors in both forward and backward direction from $\tilde{r}$ with its contiguous residues except $\tilde{r}$ until $\tilde{r}$ appeared again. The demonstration of the above process is mathematically expanded in Eqs. (3) and (4) which collectively set up a mean, $\frac{\sum_i^n T_i\{L+M+N\}}{\sum_i T_i}$ as given in expression (2). Whereas i depends upon the number of compositions of residue of type $\tilde{r}$ in concatenation. Moreover, non-occurrence will assign zero value corresponding to the weight factor, so this weight is neglected and only considered the weight factors for has occurred objects.

$$\begin{aligned}
T_i &= \alpha_{\tilde{r}_i}\beta_{\tilde{r}_i}, \qquad i = 1,2,3...,n; \quad \alpha_{\tilde{r}}, \beta_{\tilde{r}} \in \mathbb{N} \\
&Or \\
T_1 &= \alpha_{\tilde{r}_1}\beta_{\tilde{r}_1}, \quad \alpha_{\tilde{r}}, \beta_{\tilde{r}} \in \mathbb{N} \\
T_2 &= \alpha_{\tilde{r}_2}\beta_{\tilde{r}_2}, \quad \alpha_{\tilde{r}}, \beta_{\tilde{r}} \in \mathbb{N} \\
T_3 &= \alpha_{\tilde{r}_3}\beta_{\tilde{r}_3}, \quad \alpha_{\tilde{r}}, \beta_{\tilde{r}} \in \mathbb{N} \\
&\vdots \\
T_n &= \alpha_{\tilde{r}_n}\beta_{\tilde{r}_n}, \quad \alpha_{\tilde{r}}, \beta_{\tilde{r}} \in \mathbb{N}
\end{aligned} \tag{3}$$

$$\begin{aligned}
L &= \frac{1}{39}[\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_m, U_{\tilde{r}}) + h_0\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_0\chi(U_{\tilde{r}}, U_m)] \\
M &= \frac{1}{39}[\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_0\chi(U_m, U_{\tilde{r}}) + h_{\tilde{r}}\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_0\chi(U_{\tilde{r}}, U_m)] \\
N &= \frac{1}{39}[\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_0\chi(U_m, U_{\tilde{r}}) + h_0\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_{\tilde{r}}, U_m)]
\end{aligned} \tag{4}$$

Where $h_m,\ 1 \leq m \leq 20$ symbolizes the repeated coupling function χ of residue $\tilde{r}$ with all amino acid residues also $U_m \neq U_{\tilde{r}}$. For non-occurrence, it is denoted by h_0. The coupling function of $U_{\tilde{r}}$ with itself is denoted by $\chi(U_{\tilde{r}}, U_{\tilde{r}})$ and the frequency of this pair is represented by $h_{\tilde{r}}$. For the sake of convenience, consider $\chi(U_i, U_j) = \omega_{i,j};\ i = j = 1,2,3,...,20$. Whereas $\omega_{i,j}$ represents all possible coupling factors for all amino acid residue with each other. A complete interpretation for all possible correlation is given in matrix representation (5) and in term of L, M and N separately assigned in (6). The matrix (6) is adopted by constraint (7), when the pair $\chi(U_i, U_j)$ appeared, then $\omega_{i,j}$ gives 1 otherwise it is attributed as number zero.

$$\begin{pmatrix}
\omega_{1,1} & \omega_{1,2} & \omega_{1,3} & \dots & \omega_{1,20} \\
\omega_{2,1} & \omega_{2,2} & \omega_{2,3} & \dots & \omega_{2,20} \\
\omega_{3,1} & \omega_{3,2} & \omega_{3,3} & \dots & \omega_{3,20} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\omega_{20,1} & \omega_{20,2} & \omega_{20,3} & \dots & \omega_{20,20}
\end{pmatrix} \tag{5}$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 & \dots & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 & \dots & 1 \\ 0 & 0 & 1 & 1 & \dots & 1 \\ 0 & 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \tag{6}$$

Where

$$\omega_{i,j} = \begin{cases} 1 & ,when\ \chi(U_i, U_j)\ exists\ for\ both\ i = j\ or\ i \neq j \\ 0 & ,otherwise \end{cases} \tag{7}$$

Use Eqs. (3) and (4) in expression (2) to get feature vector reflecting the residue $U_{\tilde{r}}$ of type $\tilde{r}$. The desired feature vector is given in Eq. (8) and compactly defined in Eq. (9).

$$\begin{aligned} \lambda_{U_{\tilde{r}}} &= \frac{1}{39\{(\alpha_{\tilde{r}_1}\beta_{\tilde{r}_1}) + (\alpha_{\tilde{r}_2}\beta_{\tilde{r}_2}) + (\alpha_{\tilde{r}_3}\beta_{\tilde{r}_3}) + ... + (\alpha_{\tilde{r}_n}\beta_{\tilde{r}_n})\}}[(\alpha_{\tilde{r}_1}\beta_{\tilde{r}_1})\{\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_m, U_{\tilde{r}}) \\ &+ h_{\tilde{r}}\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_{\tilde{r}}, U_m)\} \\ &+ (\alpha_{\tilde{r}_2}\beta_{\tilde{r}_2})\{\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_m, U_{\tilde{r}}) + h_{\tilde{r}}\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_{\tilde{r}}, U_n)\} \\ &+ (\alpha_{\tilde{r}_3}\beta_{\tilde{r}_3})\{\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_m, U_{\tilde{r}}) + h_{\tilde{r}}\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_{\tilde{r}}, U_m)\} \\ &\vdots \\ &+ (\alpha_{\tilde{r}_n}\beta_{\tilde{r}_n})\{\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_m, U_{\tilde{r}}) + h_{\tilde{r}}\chi(U_{\tilde{r}}, U_{\tilde{r}}) + \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_{\tilde{r}}, U_m)\}] \end{aligned} \tag{8}$$

Or

$$\begin{aligned} \lambda_{U_{\tilde{r}}} &= \frac{1}{39\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})}[\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})\{\sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_m, U_{\tilde{r}}) + h_{\tilde{r}}\chi(U_{\tilde{r}}, U_{\tilde{r}}) \\ &+ \sum_{\substack{m=1 \\ m\neq\tilde{r}}}^{20} h_m\chi(U_{\tilde{r}}, U_m)\}] \end{aligned} \tag{9}$$

In order to understand the mechanism of proposed model consider ith term U_i of sequence (1), reflects the first alphabetical letter of amino acid residues, say, "A". Notice its occurrences as well as its corresponding positions in the sequence. U_i makes pairing with its contiguous residues in reverse and forward direction. The *ith* residue in term of $\chi(U_m, U_i)$ and $\chi(U_i, U_m)$ represented by green and blue curved lines (see

Fig 2. A mechanism for sequence formulation. The figure is to show the graphical demonstration of scheme feature vector for the residue "A", representing how "A" make pairs with its contiguous residues in both directions up to next residue.

Fig. 2). This process will be continued until next U_j occurs at jth position such that $U_i = U_j = A$. Similarly, the same steps will be conducted for U_j. The feature component corresponding to residue "A" is interpreted in Eq. (10).

$$\begin{aligned}\lambda_A &= \frac{1}{39\{(\alpha_{\tilde{r}_1}\beta_{\tilde{r}_i}) + (\alpha_{\tilde{r}_2}\beta_{\tilde{r}_j})\}}[(\alpha_{\tilde{r}_1}\beta_{\tilde{r}_i})\{\sum_{\substack{m=1\\U_m\neq A}}^{20} h_m\chi(U_m, A) + h_A\chi(A, A) \\ &+ \sum_{\substack{m=1\\U_m\neq A}}^{20} h_m\chi(A, U_m)\} + (\alpha_{\tilde{r}_2}\beta_{\tilde{r}_j})\{\sum_{\substack{m=1\\U_m\neq A}}^{20} h_m\chi(U_m, A) + h_A\chi(A, A) \\ &+ \sum_{\substack{m=1\\U_m\neq A}}^{20} h_m\chi(A, U_m)\}] \qquad (10)\end{aligned}$$

Where the numeral values $m = 1, 2, 3..., 20$ indicates the twenty amino acid residues of alphabetical order. For more convenience, assume that $U_1, U_2, U_3, ..., U_{20}$ represents 20 amino acids of alphabetical order labeled as: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y and U_{21} onwards the 20 residues cyclically repeats themselves then λ_1,λ_2,λ_3,...,λ_{20} be their corresponding feature components. The set of twenty feature components is given in Eq. (11) as follows.

$$
\begin{aligned}
\lambda_1 &= \frac{1}{39\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})}[\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})\{\sum_{\substack{m=2\\m\neq 1}}^{20} h_m\chi(U_m,U_1)+h_1\chi(U_1,U_1)\\
&+ \sum_{\substack{m=2\\m\neq 1}}^{20} h_m\chi(U_1,U_m)\}]\\
\lambda_2 &= \frac{1}{39\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})}[\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})\{\sum_{\substack{m=1\\m\neq 2}}^{20} h_m\chi(U_m,U_2)+h_2\chi(U_2,U_2)\\
&+ \sum_{\substack{m=1\\m\neq 2}}^{20} h_m\chi(U_2,U_m)\}]\\
\lambda_3 &= \frac{1}{39\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})}[\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})\{\sum_{\substack{m=1\\m\neq 3}}^{20} h_m\chi(U_m,U_3)+h_3\chi(U_3,U_3)\\
&+ \sum_{\substack{m=1\\m\neq 3}}^{20} h_m\chi(U_3,U_n)\}] \qquad (11)\\
&\vdots\\
\lambda_{20} &= \frac{1}{39\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})}[\sum_{i=1}^{n}(\alpha_{\tilde{r}_i}\beta_{\tilde{r}_i})\{\sum_{\substack{m=1\\m\neq 20}}^{20} h_m\chi(U_m,U_{20})+h_{20}\chi(U_{20},U_{20})\\
&+ \sum_{\substack{m=1\\m\neq 20}}^{20} h_m\chi(U_{20},U_m)\}]
\end{aligned}
$$

The above set of twenty feature vectors depends upon three properties of amino acids such that, hydrophobicity, hydrophilicity and side chain mass of amino acids, can be calculated by employing Eqs. (12) to (14). These equations can expand as per choice of attributes of amino acids other than these three properties of amino acid. For extended properties l of amino acids a compact representation is elaborated in Eq. (15).

$$
\begin{aligned}
\chi(U_i,U_j) &= \frac{|\aleph^*_{i_1}(U_j)|}{1+|\aleph^*_{i_1}(U_i)+\aleph^*_{i_1}(U_j)|}+\frac{1}{2^3}[\frac{|\aleph^*_{i_1}(U_j)-\aleph^*_{i_1}(U_i)|}{1+|\aleph^*_{i_1}(U_j)-\aleph^*_{i_1}(U_i)|}\\
&+ \frac{|\aleph^*_{i_2}(U_j)-\aleph^*_{i_2}(U_i)|}{1+|\aleph^*_{i_2}(U_j)-\aleph^*_{i_2}(U_i)|}+\frac{|\aleph^*_{i_3}(U_j)-\aleph^*_{i_3}(U_i)|}{1+|\aleph^*_{i_3}(U_j)-\aleph^*_{i_3}(U_i)|}] \qquad (12)
\end{aligned}
$$

$$
\begin{aligned}
\chi(U_i,U_j) &= \frac{|\aleph^*_{i_2}(U_j)|}{1+|\aleph^*_{i_2}(U_i)+\aleph^*_{i_2}(U_j)|}+\frac{1}{2^3}[\frac{|\aleph^*_{i_1}(U_j)-\aleph^*_{i_1}(U_i)|}{1+|\aleph^*_{i_1}(U_j)-\aleph^*_{i_1}(U_i)|}\\
&+ \frac{|\aleph^*_{i_2}(U_j)-\aleph^*_{i_2}(U_i)|}{1+|\aleph^*_{i_2}(U_j)-\aleph^*_{i_2}(U_i)|}+\frac{|\aleph^*_{i_3}(U_j)-\aleph^*_{i_3}(U_i)|}{1+|\aleph^*_{i_3}(U_j)-\aleph^*_{i_3}(U_i)|}] \qquad (13)
\end{aligned}
$$

$$\chi(U_i,U_j) = \frac{|\aleph^*_{i_3}(U_j)|}{1+|\aleph^*_{i_3}(U_i)+\aleph^*_{i_3}(U_j)|} + \frac{1}{2^3}[\frac{|\aleph^*_{i_1}(U_j)-\aleph^*_{i_1}(U_i)|}{1+|\aleph^*_{i_1}(U_j)-\aleph^*_{i_1}(U_i)|} + \frac{|\aleph^*_{i_2}(U_j)-\aleph^*_{i_2}(A_i)|}{1+|\aleph^*_{i_2}(U_j)-\aleph^*_{i_2}(U_i)|} + \frac{|\aleph^*_{i_3}(U_j)-\aleph^*_{i_3}(U_i)|}{1+|\aleph^*_{i_3}(U_j)-\aleph^*_{i_3}(U_i)|}] \quad (14)$$

Or

$$\chi(U_i,U_j) = \frac{|\aleph^*_{l}(U_j)|}{1+|\aleph^*_{l}(U_i)+\aleph^*_{l}(U_j)|} + \frac{1}{2^l}\sum_{l}[\frac{|\aleph^*_{l}(U_j)-\aleph^*_{l}(U_i)|}{1+|\aleph^*_{l}(U_j)-\aleph^*_{l}(U_i)|} \quad (15)$$

Whereas $\aleph^*_{i_1}$, $\aleph^*_{i_2}$, $\aleph^*_{i_3}$ are the values of hydrophobicity, hydrophilicity and side-chain mass of amino acid residues that are normalized by using Eq. (16) against the pair of U_i and U_j. The normalization index that is used to normalize the values given in Eqs. (12) to (14) lies between (-S, S), where S is the normalizing count for $\tilde{r}$ amino acids. Here the number 5 is used for normalization. The original values for hydrophobicity and hydrophilicity were taken from the main source, employed by Ehsan at el. [31, 32], while the values for side-chain mass of amino acid residues was taken from any text book of biochemistry.

$$\begin{aligned}
\aleph^*_{i_1}(\tilde{r}) &= \{\frac{2S}{(\aleph_{i_{1_{(max)}}}-\aleph_{i_{1_{(min)}}})}(\aleph_{i_1}(\tilde{r})-\aleph_{i_{1_{(max)}}})\}+S \\
\aleph^*_{i_2}(\tilde{r}) &= \{\frac{2S}{(\aleph_{i_{2_{(max)}}}-\aleph_{i_{2_{(min)}}})}(\aleph_{i_2}(\tilde{r})-\aleph_{i_{2_{(max)}}})\}+S \\
\aleph^*_{i_3}(\tilde{r}) &= \{\frac{2S}{(\aleph_{i_{3_{(max)}}}-\aleph_{i_{3_{(min)}}})}(\aleph_{i_3}(\tilde{r})-\aleph_{i_{3_{(max)}}})\}+S
\end{aligned} \quad (16)$$

The provided modelling consist of 100 dimensions by comprising the protein features and to classify them according to their functional properties and attributes. These features are divided as: the very first twenty feature vectors corresponds to hydrophobic property, while next twenty matches for hydrophilic attribute. Similarly the succeeding twenty vectors indicate side chain mass property of amino acid residues and last forty vectors related to the position and composition of each amino acids. For the identification of diverse protein sequence this novel technique establishes a wonderful result. For the sake of classification these extracted feature vectors are further passed through a training-testing process by using the rigorous classifier, neural networks (NN).

The neural network is an extraordinary tool for decision making problems and to classify patterns in available diversified data sets. It is typically arranged in layers and learn from its experience using input data and able to modify their weights according to provided data. Subsequent to the training process is finished the system apparently acts such that makes it fit to arrange each given input inside a worthy level of precision. Its connectionist structural design comprises of 100 input layer neurons, 50 hidden layer neurons and two output neurons that classify hydroxylated and non-hydroxylated protein samples. The back propagation method was used for training of the multilayered neural network. In order to get the higher prediction rate and to decrease the error rate a gradient descent method was employed with adaptive learning rate.

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. This procedure is done in the flowchart as given underneath (see Fig. 3).

Fig 3. Flowchart describes the training and validation process. The prediction process done in the following steps (a) extraction of two class dataset (b) extract feature vectors using the proposed tool (c) train/test dataset using neural network classifier.

Results

In order to build up a beneficial predictor for an organic development, the Chou's 5-step rule [26] are noticeable. Undoubtedly, it is useful to develop a new predictor by employing Chou's 5-step rule. A number of researchers [27–30] had used this method in their work, published very recently. The prediction analysis is done in some steps: firstly the stringent benchmark data set is collected for training and testing purpose of proposed predictor, in a second step, a powerful mathematical tool developed that select the major and most significant features of the amino acid polymers. Then the developed feature vector incorporated into an identifying formulation for the sake of training. When the process of training is completed, the trained model is completely tested and validated. Finally, a web-server is created for open use of the proposed predictor. In the current study, the initial four steps have been carefully performed, while, the last step has been kept open for future work.

Statistical Measures

To evaluate the performance of the proposed model "iHyd-ProSite", a set of four metrics are followed, which were employed by Ehsan et al. [31,32]. The following these four metrics are: sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficients (MCC) respectively, used for proposed algorithm evaluation. Either the set of traditional metrics copied from maths books or the intuitive metrics derived from the Chou's symbols [33,34] is valid only for the single-label systems (where each sample only belongs to one class). For the multilabel systems (where a sample may simultaneously belong to several classes), whose existence has become more frequent in system biology [28,35], system medicine [36] and biomedicine [37], a completely different set of metrics as defined in the study represented as a reference [38] is absolutely needed.

$$\begin{cases} Sn = 1 - \frac{\P_-^+}{\P^+} & 0 \le Sn \le 1 \\ Sp = 1 - \frac{\P_+^-}{\P^-} & 0 \le Sp \le 1 \\ Acc = 1 - \frac{\P_-^+ + \P_+^-}{\P^+ + \P^-} & 0 \le Acc \le 1 \\ MCC = \frac{1 - \left(\frac{\P_-^+}{\P^+} + \frac{\P_+^-}{\P^-}\right)}{\sqrt{\left(1 + \frac{\P_+^- - \P_-^+}{\P^+}\right)\left(1 + \frac{\P_-^+ - \P_+^-}{\P^-}\right)}} & -1 \le MCC \le 1 \end{cases} \tag{17}$$

Consider $\P^+$ and $\P^-$ represents the all correctly predicted positive and negative records for hydroxylated and non-hydroxylated site in proline (Pro) within polypeptide sequences. Similarly, if the positive records are wrongly predicted as negative records, is denoted by $\P_-^+$ and when negative records are wrongly predicted in terms of positive records is represented by $\P_+^-$. It is relevant to discuss the following

cases of the above Eq. (17). If $¶^+{}_- = 0$, then there is no incorrectly predicted positive records as negative records, then produce $Sn = 1$. In another case when $¶^+{}_- = ¶^+$ indicating that all positive records were incorrectly predicted in terms of negative records so, the sensitivity is $Sn = 0$. Similarly for $¶^-{}_+ = 0$ gives specificity $Sp = 1$ representing no any negative record peptide was incorrectly predicted in terms of positive records, while for $¶^-{}_+ = ¶^-$ gives specificity $Sp = 0$ denoting all negative instances were wrongly predicted as positive instances. On the other hand the prediction accuracy $Acc = 1$ when, there is no any incorrectly predicted sequences were found both positive and negative cases such that, $¶^+{}_- = ¶^-{}_+ = 0$. When $¶^+{}_- = ¶^+$ and $¶^-{}_+ = ¶^-$, brings out misclassification so the overall accuracy is $Acc = 0$. Furthermore, the performance of binary classifications is often measured by Matthew correlative coefficient (MCC). There are three cases here, for $¶^+{}_- = ¶^-{}_+ = 0$ indicating that there is no incorrectly predicted record were found for both positive and negative instances so we obtain $MCC = 1$. In the second case when $¶^+{}_- = \frac{¶^+}{2}$ and $¶^-{}_+ = \frac{¶^-}{2}$ we obtain $MCC = 0$ indicating the inaccurate prediction. Lastly, when $¶^+{}_- = ¶^+$ and $¶^-{}_+ = ¶^-$ we obtain $MCC = -1$ denoting the totally wrong binary classification and disagreement between observed and predicted values.

Renowned Validation Tests

In order to validate the quality of the proposed predictor the following three test methods, self-consistency test, K-fold cross validation test and jackknife test are often used. These tests are applied to score the metrics given in Eq. (17). To approve the predictor's quality these tests are viewed as valuable. By employing the statistical measures, a comparison was made using the jackknife test with the existing predictors [8, 25]. In the current study, to test the performance of proposed scheme all above test methods were employed. Additionally, for validation purpose the benchmark datasets were taken from two sources, one is from uniprot and other one is from dbptm. The results obtained by using both datasets are given in Table 1. Table 1 is divided into two main columns. The first column is explaining the values of metrics for dbptm dataset by employing all above tests. Whereas, second column is giving the values in validation of uniprot dataset. It is noticed that the values for MCC were 0.91 and 0.90 for jackknife test. The comparison of the proposed scheme with the existing techniques for jackknife test is given below.

Table 1. The values Table by using proposed predictor "iHyd-ProSite".

Metrics values for dbptm dataset					**Metrics values for uniprot dataset**			
Tests	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC
$Self - Consitency$	99.30	98.20	99.31	0.96	99.48	98.84	98.69	0.95
$Cross - Validation$	98.95	95.87	97.85	0.94	94.87	95.20	95.06	0.92
$Jackknife$	98.90	95.82	97.80	0.91	94.82	95.15	95.01	0.90

The results obtained by employing the proposed predictor by using self-consistency test, cross-validation, test and jackknife test on the set of metrics for dbptm and uniprot datasets for identifying hydroxyproline sites.

Comparison Analysis

Observe Table 2 for a comparison analysis with the existing techniques "iHyd-PseAAC" [8], and "iHyd-PseCp" [25]. The comparison was also made with the most recent publication "iHyd-PseAAC (EPSV)" [32] for identifying the hydroxyproline sites. All these techniques attained the metrics records, employing the jackknife test method. It can be noticed from Table 2 that the accuracy (Acc),

stability (MCC), sensitivity (Sn), and specificity (Sp) assess measured via operating proposed scheme are more predominant than the those values given by the former methodologies. There are two benchmark datasets borrowed from (a) dbptm and (b) uniprot database for comparison purpose with the existing schemes. Indeed, the newly proposed methodology is absolutely better suggestion throughout the past methodologies. There are a number of scientific and theoretical reasons can be explained for the improved quality of the developed scheme. Few of them are covered here. First of all, the proposed formulation is established by incorporating position and composition of primary protein structure and is beneficial to deal with the different length protein sequences in a thoughtful manner without missing any hidden data and also organize pairwise couplings in every possible permutation of amino acid residues. Second, it produces uniform dimension vectors, which contribute invariant size feature vectors that uniformly classify proteins corresponding to their properties. This concept allows the predictor to meticulously separate and appropriately distinguish each instance. Third, the correlation aspect is the principle concept that impart for computing feature vector. It has been assembled by considering each attribute group. Each expression deals with some specific metric and statistical measures. For the sake of convenience, every property of amino acids was standardized numerically within a suitable range. Also, it has been noticed that in comparison with previous methods proposed, the predicted outcomes are more superior and better than the former prediction rate.

Table 2. A comparison analysis of the proposed predictor with the existing predictors using well-known jackknife validation tests for the metrics given in Eq. (17).

Comparison Table				
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
$iHyd-PseACC(EPSV)^a$	98.68	94.82	96.80	0.90
$iHyd-PseACC(EPSV)^b$	97.02	94.57	96.01	0.88
$iHyd-ProSite^a$	98.90	95.82	97.80	0.91
$iHyd-ProSite^b$	94.82	95.15	95.01	0.90

A comparison is made for jackknife test using benchmark datasets obtained from (a)dbptm and (b)uniprot database sources. It can be seen that the results obtained by using proposed predictor "iHyd-ProSite" is much better than all previous methodologies.

Discussion

Table 2 explain that, the values of sensitivity, specificity, accuracy and methew correlation coefficient for proposed predictor are higher than all the values obtained by utilizing former schemes. Sensitivity test describes the correctly predicted hydroxylated sites which are extraordinary larger than all reported values for previous methodologies. Also the stability of the predictor is measured by MCC value, and it can be observed that MCC values obtained by using proposed scheme are greater than above reported values. Undoubtedly, the proposed scheme is much helpful for diagnosing the biological problems efficiently.

Conclusion

The novel proposed technique "iHyd-ProSite" is a new predictor to find hydroxyproline sites in protein sequences. Undoubtedly, it can be observed from the comparison Table 2 that the results obtained by using the proposed method are higher-up than all previous methods. For example, the accuracy calculated with the proposed tool (iHyd-ProSite) were **97.80** and **95.01** corresponding to two benchmark datasets obtained from databases (a) dbptm and (b) uniprot which is superior than the accuracies obtained by all previous predictors. Also MCC value were 0.91 and 0.90 with is greater than all schemes iHyd-PseAAC, iHyd-PseCp and iHyd-PseAAC (EPSV). Also the set of two data sets taken from dbptm and uniprot database were utilized for the proposed predictor validation. This technique is convenient to handle all types of biological data and can gently classify the unpredictable biological sequences. If the researchers are interested in the classification problems they should use this handy predictor, it can be helpful for future prediction problems.

Bold the title sentence. Add descriptive text after the title of the item (optional).

Conflict of Interest

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

References

1. Colgrave, Michelle L and Allingham, Peter G and Jones, Alun. Hydroxyproline quantification for the estimation of collagen in tissue using multiple reaction monitoring mass spectrometry. J. Chromatogr. A 2008 Nov;1212(1-2):150–153.

2. Gelse, Kolja and Pöschl, E and Aigner, T. Collagens structure, function, and biosynthesis. Adv. Drug Deliv. Rev. 2003 Nov;55(12):1531–1546.

3. Ruszczak, Zbigniew. Effect of collagen matrices on dermal wound healing. Adv. Drug Deliv. Rev. 2003 Nov;55(12):1595–1611.

4. Lee, Chi H and Singla, Anuj and Lee, Yugyung. Biomedical applications of collagen. Int. J. Pharm. 2001 Jun;221(1-2):1–22.

5. Becker, Gary D and Adams, Lawrence A and Hackett, James. Collagen-assisted healing of facial wounds after mohs surgery. The Laryngoscope 1994 Oct;104(10):1267–1270.

6. Guszczyn, Tomasz and Sobolewski, Krzysztof Deregulation of collagen metabolism in human stomach cancer. Pathobiology 2004 Dec;71(6):308–313.

7. Sunila, ES and Kuttan, G. A preliminary study on antimetastatic activity of Thuja occidentalis L. in mice model. IMMUNOPHARM IMMUNOT 2006 Oct;28(2):269–280.

8. Xu, Yan and Wen, Xin and Shao, Xiao-Jian and Deng, Nai-Yang and Chou, Kuo-Chen. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int. J. Mol. Sci. 2014 May;15(5):7594–7610.

9. Jia, Jianhua and Liu, Zi and Xiao, Xuan and Liu, Bingxiang and Chou, Kuo-Chen. iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget. 2016 May;7(23):34558–34570.

10. Jia, Jianhua and Zhang, Liuxia and Liu, Zi and Xiao, Xuan and Chou, Kuo-Chen. pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinform. 2016 Oct;32(20):3133–3141.

11. Khan, Yaser Daanial and Rasool, Nouman and Hussain, Waqar and Khan, Sher Afzal and Chou, Kuo-Chen. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. Anal. Biochem. 2018 Jun;550:109–116.

12. Khan, Yaser Daanial and Rasool, Nouman and Hussain, Waqar and Khan, Sher Afzal and Chou, Kuo-Chen. iPhosY-PseAAC: identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. Mol. Biol. 2018 Oct;45(6):2501–2509.

13. Cockman, Matthew E and Webb, James D and Kramer, Holger B and Kessler, Benedikt M and Ratcliffe, Peter 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. Proteom. 2009 Oct;8(3):535–546.

14. Ang, Kok S and Lakshmanan, Meiyappan and Lee, Na-Rae and Lee, Dong-Yup. Metabolic modeling of microbial community interactions for health, environmental and biotechnological applications. Curr. Genomics 2018 Dec;19(8):712–722.

15. Berg, Richard A and Steinmann, Beat and Rennard, Stephen I and Crystal, Ronald G. Ascorbate deficiency results in decreased collagen production: under-hydroxylation of proline leads to increased intracellular degradation. Arch. Biochem. Biophys. 1983 Oct;226(2):681–686.

16. Halme, Jouko and Kivirikko, Kari I and Simons, Kai. Isolation and partial characterization of highly purified protocollagen proline hydroxylase. Biochim. Biophys. Acta 1970 Mar;198(3):460–470.

17. Kivirikko, Kari I and Prockop, Darwin J. Hydroxylation of proline in synthetic polypeptides with purified protocollagen hydroxylase. J. Biol. Chem. 1967 Sep;242(18):4007–4012.

18. Morgan, Alexander A and Rubenstein, Edward. Proline: the distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. PloS one. 2013 Jan;8(1).

19. Yamauchi, Mitsuo and Shiiba, Masashi. Lysine hydroxylation and crosslinking of collagen. Posttranslational Modifications of Proteins. 2002:277–290.

20. Shi, Shao-Ping and Chen, Xiang and Xu, Hao-Dong and Qiu, Jian-Ding. PredHydroxy: computational prediction of protein hydroxylation site locations based on the primary structure. Mol. Biosyst. 2015 Dec;11(3):819–825.

21. Wu, Guoyao and Bazer, Fuller W and Burghardt, Robert C and Johnson, Gregory A and Kim, Sung Woo and Knabe, Darrell A and Li, Peng and Li, Xilong and McKnight, Jason R and Satterfield, M Carey and others. Proline and hydroxyproline metabolism: implications for animal and human nutrition. Amino acids. 2011 Aug;40(4):1053–1063.

22. Hayat, Shamsul and Hayat, Qaiser and Alyemeni, Mohammed Nasser and Wani, Arif Shafi and Pichtel, John and Ahmad, Aqil. Role of proline under changing environments: a review. Plant Signal Behav. 2012 Sep;7(11):1456–1466.

23. Yang, Zheng Rong. Predict collagen hydroxyproline sites using support vector machines. J. Comput. Biol. 2009 May;16(5):691–702.

24. Hu, Le-Le and Niu, Shen and Huang, Tao and Wang, Kai and Shi, Xiao-He and Cai, Yu-Dong. Prediction and analysis of protein hydroxyproline and hydroxylysine. PLoS One 2010 Dec;5(12).

25. Qiu, Wang-Ren and Sun, Bi-Qian and Xiao, Xuan and Xu, Zhao-Chun and Chou, Kuo-Chen. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget. 2016 Jun;7(28):44310.

26. Chou, Kuo-Chen. Some remarks on protein attribute prediction and pseudo amino acid composition. J. theor. biol. 2011 Mar;273(1):236–247.

27. Chou, Kuo-Chen and Cheng, Xiang and Xiao, Xuan. pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. Genomics. 2019 Dec;111(6):1274–1282.

28. Xiao, Xuan and Cheng, Xiang and Chen, Genqiang and Mao, Qi and Chou, Kuo-Chen. pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics. 2019 Jul;111(4):886–892.

29. Khan, Yaser Daanial and Jamil, Mehreen and Hussain, Waqar and Rasool, Nouman and Khan, Sher Afzal and Chou, Kuo-Chen. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. J. Theor. Biol. 2019 Feb;463:47–55.

30. Jia, Jianhua and Li, Xiaoyan and Qiu, Wangren and Xiao, Xuan and Chou, Kuo-Chen. iPPI-PseAAC (CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. J. Theor. Biol. 2019 Jan;460:195–203.

31. Ehsan, Asma and Mahmood, Khalid and Khan, Yaser Daanial and Khan, Sher Afzal and Chou, Kuo-Chen. A novel modeling in mathematical biology for classification of signal peptides. Sci. Rep. 2018 Jan;8(1):1–16.

32. Ehsan, Asma and Mahmood, Muhammad Khalid and Khan, Yaser Daanial and Barukab, omar Mohammed and Khan, Sher Afzal and Chou, Kuo-Chen. 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 Feb;20(2):124–133.

33. Chou, Kuo-Chen. Using subsite coupling to predict signal peptides. Prot. Eng. 2001 Feb;14(2):75–79.

34. Chou, Kuo-Chen. Prediction of signal peptides using scaled window. Peptides. 2001 Dec;22(12):1973–1979.

35. Cheng, Xiang and Xiao, Xuan and Chou, Kuo-Chen. pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics. 2018 Jul;110(4):231–239.

36. Cheng, Xiang and Zhao, Shu-Guang and Xiao, Xuan and Chou, Kuo-Chen. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinform. 2017 Feb;33(3):341–346.

37. Qiu, Wang-Ren and Sun, Bi-Qian and Xiao, Xuan and Xu, Zhao-Chun and Chou, Kuo-Chen. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinform. 2016 Oct;32(20):3116–3123.

38. Chou, Kuo-Chen. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 2013 Feb;9(6):1092–1100.

39. Chou, Kuo-Chen and Shen, Hong-Bin and others. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2009;1(2):63–92.

40. Chou, Kuo-Chen. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 2015 May;11(3):218–234.

41. Chou, Kuo-Chen. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr. Top. Med. Chem. 2017 Aug;17(21):2337–2358.

42. Lu, Cheng-Tsung and Huang, Kai-Yao and Su, Min-Gang and Lee, Tzong-Yi and Bretana, Neil Arvin and Chang, Wen-Chi and Chen, Yi-Ju and Chen, Yu-Ju and Huang, Hsien-Da. DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic acids res. 2013 Jan;41(D1):D295–D305.

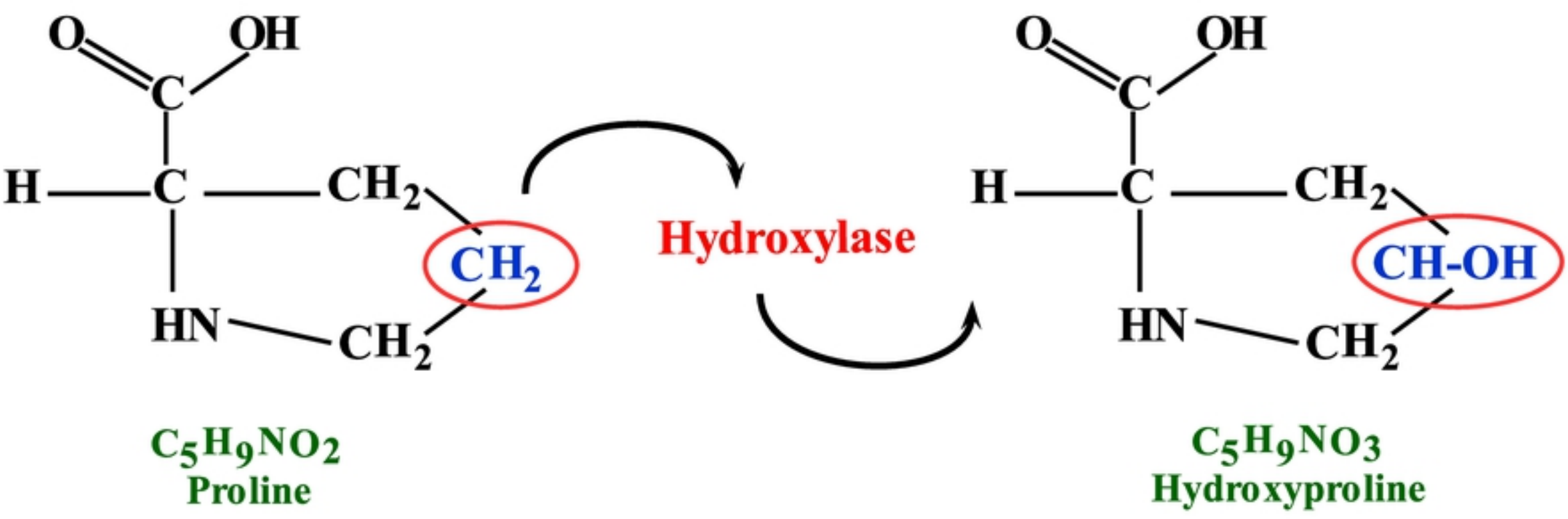
O
OH
C
H
C
CH2
CH2
HN
CH2
Hydroxylase
O
OH
C
H
C
CH2
CH-OH
HN
CH2
C5H9NO2
Proline
C5H9NO3
Hydroxyproline

Figure 1

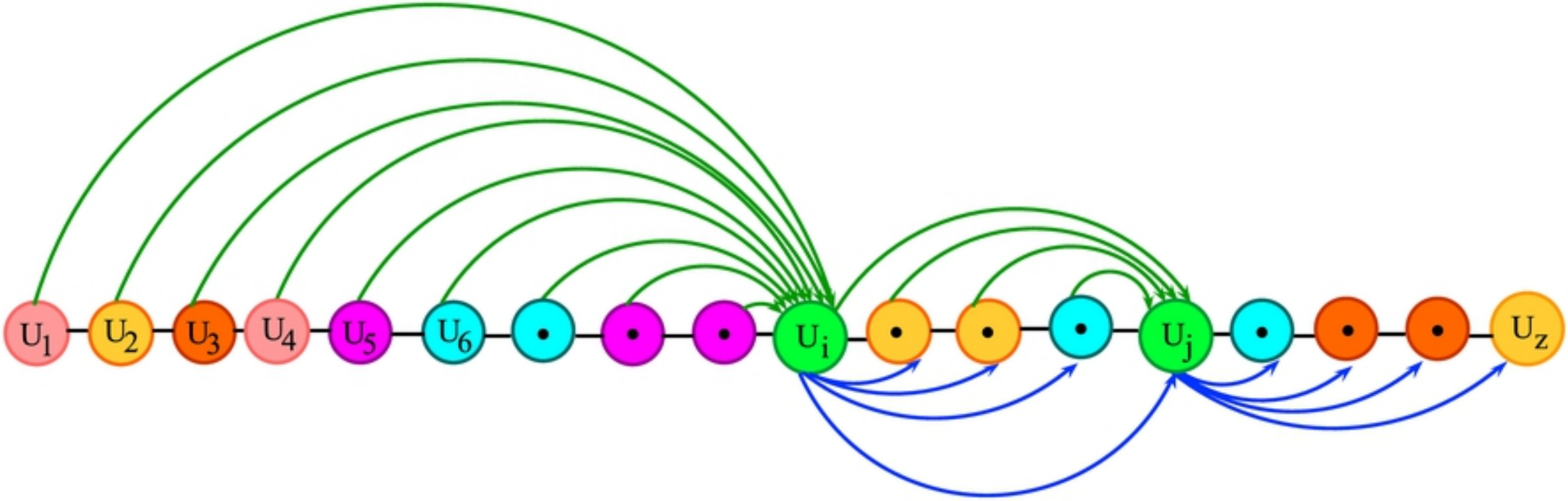

Figure 2

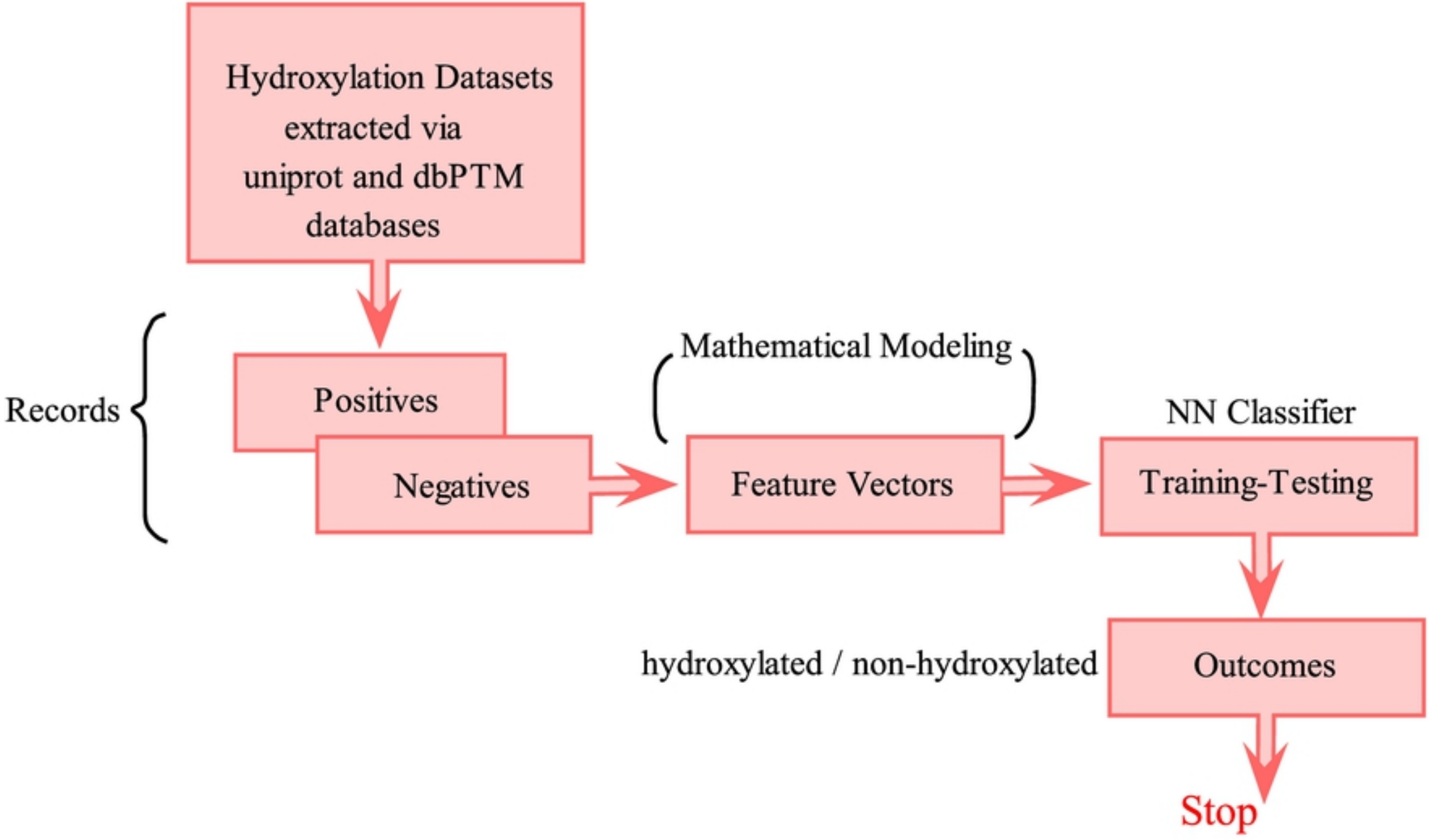
Hydroxylation Datasets extracted via uniprot and dbPTM databases
Records
Positives
Negatives
Mathematical Modeling
Feature Vectors
NN Classifier
Training-Testing
hydroxylated / non-hydroxylated
Outcomes
Stop

Figure 3