Int. J. Mol. Sci. **2014**, *15*, 7594-7610; doi:10.3390/ijms15057594

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

iHyd-PseAAC: Predicting Hydroxyproline and Hydroxylysine in Proteins by Incorporating Dipeptide Position-Specific Propensity into Pseudo Amino Acid Composition

Yan Xu [1,5,*], Xin Wen [1], Xiao-Jian Shao [2], Nai-Yang Deng [3] and Kuo-Chen Chou [4,5]

[1] Department of Information and Computer Science, University of Science and Technology Beijing, Beijing 100083, China; E-Mail: wenxinfairy@gmail.com
[2] Department of Mathematics and Information Science, Binzhou University, Binzhou 256603, China; E-Mail: shaoxiaojian@gmail.com
[3] College of Science, China Agricultural University, Beijing 100083, China; E-Mail: dengnaiyang@cau.edu.cn
[4] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia; E-Mail: kcchou@gordonlifescience.org
[5] Gordon Life Science Institute, Boston, MA 02478, USA

* Author to whom correspondence should be addressed; E-Mail: xyan@gordonlifescience.org or xuyan@ustb.edu.cn; Tel./Fax: +86-10-6233-2589.

Received: 7 February 2014; in revised form: 4 April 2014 / Accepted: 17 April 2014 / Published: 5 May 2014

Abstract: Post-translational modifications (PTMs) play crucial roles in various cell functions and biological processes. Protein hydroxylation is one type of PTM that usually occurs at the sites of proline and lysine. Given an uncharacterized protein sequence, which site of its Pro (or Lys) can be hydroxylated and which site cannot? This is a challenging problem, not only for in-depth understanding of the hydroxylation mechanism, but also for drug development, because protein hydroxylation is closely relevant to major diseases, such as stomach and lung cancers. With the avalanche of protein sequences generated in the post-genomic age, it is highly desired to develop computational methods to address this problem. In view of this, a new predictor called "iHyd-PseAAC" (identify hydroxylation by pseudo amino acid composition) was proposed by incorporating the dipeptide position-specific propensity into the general form of pseudo amino acid composition. It was demonstrated by rigorous cross-validation tests on stringent benchmark datasets that the new predictor is quite promising and may become a useful high throughput tool in this area. A user-friendly web-server for iHyd-PseAAC is accessible at

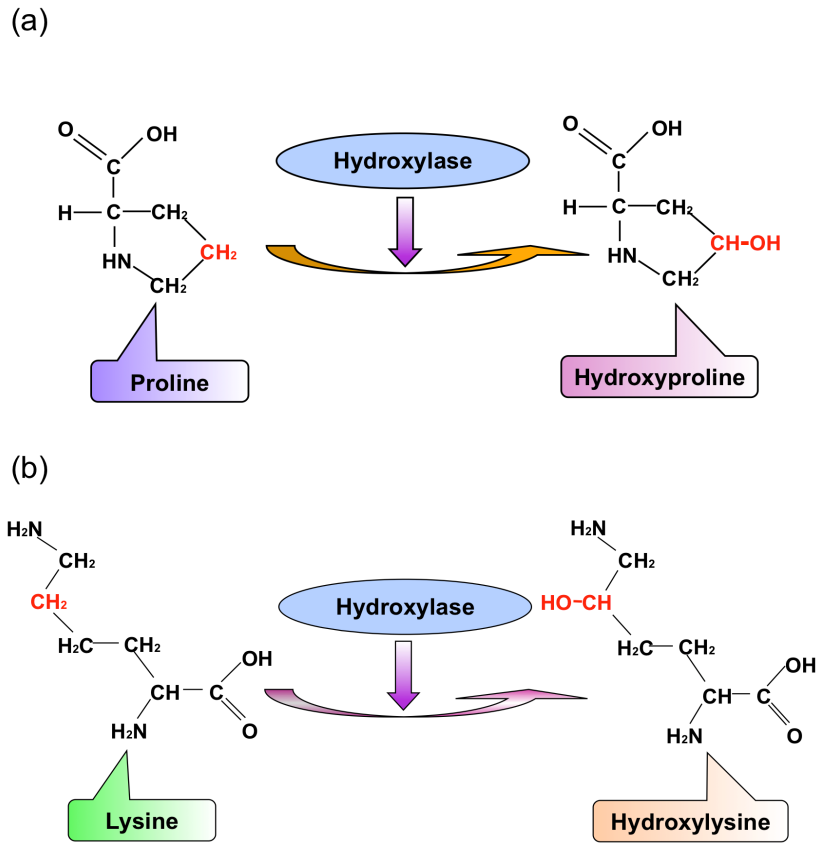
http://app.aporc.org/iHyd-PseAAC/. Furthermore, for the convenience of the majority of experimental scientists, a step-by-step guide on how to use the web-server is given. Users can easily obtain their desired results by following these steps without the need of understanding the complicated mathematical equations presented in this paper just for its integrity.

Keywords: PTMs; HyP; HyL; PseAAC; discriminant function algorithm

1. Introduction

Most proteins perform their functions after post-translational modifications (PTMs). Protein hydroxylation is one type of PTM that involves the conversion of a CH group into a COH group (Figure 1) and is closely relevant to the regulation of the transcription factor (hypoxia-inducible factor) [1]. Both the proline and lysine residues in proteins can be hydroxylated, forming hydroxyproline (Figure 1a) or HyP and hydroxylysine (Figure 1b) or HyL, respectively. However, the former is more common than the latter [2,3]. Furthermore, HyP is the key factor in stabilizing collagens [4,5], whose instability or abnormal activity may cause stomach cancer [6] and lung cancer [7,8]. Therefore, identifying the HyP and HyL sites in proteins may provide useful information for both biomedical research and drug development.

Figure 1. Schematic drawing to show protein hydroxylation occurring at (**a**) proline and (**b**) lysine to form hydroxyproline (HyP) and hydroxylysine (HyL), respectively.



Identification of hydroxylation residues with experiments was mainly done by means of mass spectrometry [1,9], which was expensive and laborious. Facing the avalanche of protein sequences generated in the post genomic age, it is highly demanded to develop a computational method for timely and effectively identifying the hydroxylation residues in proteins. However, to our best knowledge, so far, only two papers have been published in this regard [10,11]. Additionally, further development in this important area is definitely needed for the following reasons. First, with a rapidly growing database in protein hydroxylation, the benchmark datasets used in the two methods definitely need to be updated; Second, some sequence order effects were missed, which would certainly affect their prediction quality; Third, none of them provided a publicly accessible web-server, and hence, their practical usage value is substantially limited.

The present study was devoted to develop a new predictor for identifying hydroxyproline and hydroxylysine in proteins by considering the above three aspects. The principle was based on a window sliding strategy, quite similar to the popular approach developed by Garnier and Robson [12] for predicting the secondary structure of globular proteins.

As demonstrated by a series of recent publications [13–20] and summarized in a comprehensive review [21], to develop a really useful predictor for a protein or peptide system, we need to go through the following five steps: (1) select or construct a valid benchmark dataset to train and test the predictor; (2) represent the protein or peptide samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm or operation engine to conduct the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (5) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us elaborate on how to deal with these five steps.

2. Results and Discussion

2.1. Benchmark Dataset

In this study, the benchmark dataset was derived from dbPTM 3.0 [22] at http://dbptm.mbc.nctu.edu.tw/, a protein post-translational modifications database. For facilitating the description later, let us adopt Chou's peptide formulation, which was used for investigating the HIV protease cleavage sites [23,24], the specificity of GalNAc-transferase [25], as well as signal peptide cleavage sites [26–29]. According to Chou's scheme, a peptide with Pro (namely P in its single-letter code) or Lys (namely K) located at its center (Figure 2) can be expressed as:

$$\begin{cases} \mathbf{P}(\mathbb{P}) = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{P}\,\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi} \\ \mathbf{P}(\mathbb{K}) = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{K}\,\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi} \end{cases} \quad (1)$$

where the subscript, ξ, is an integer (*cf*. Figure 2), $\mathrm{R}_{-\xi}$ represents the ξ-th downstream amino acid residue from the center, R_{ξ} the ξ-th upstream amino acid residue, and so forth. Peptides $\mathbf{P}(\mathbb{P})$ and $\mathbf{P}(\mathbb{K})$ with the profile of Equation (1) can be further classified into the following categories:

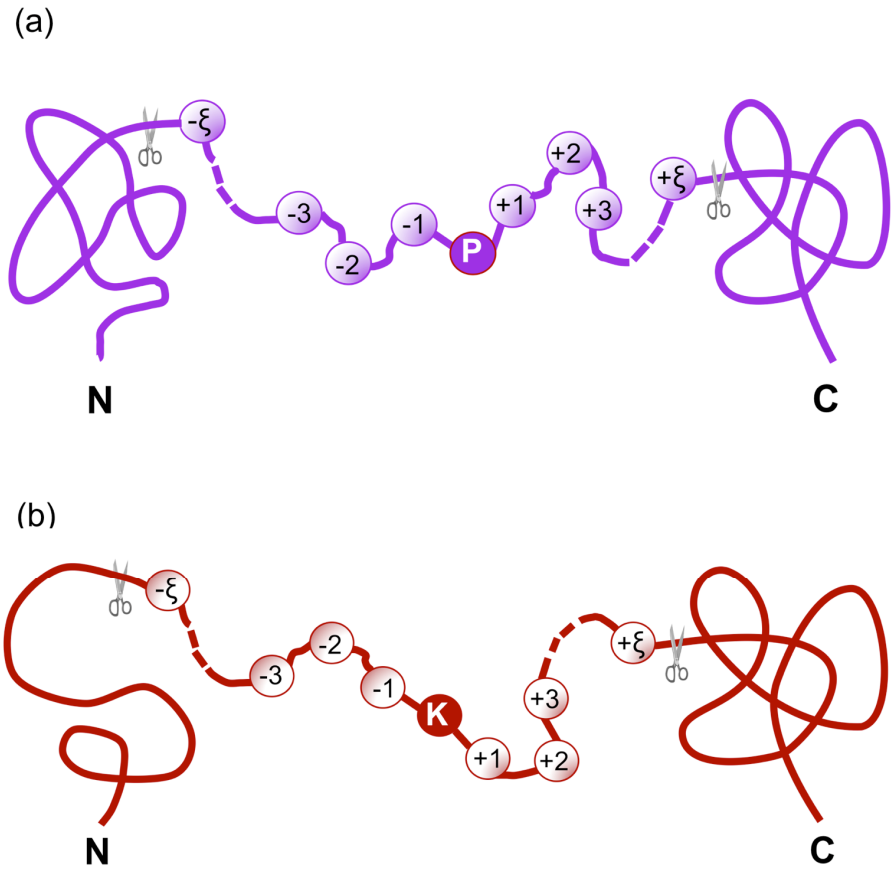
$$\mathbf{P}(\mathbb{P}) \in \begin{cases} \text{Pro-hydroxylated peptide,} & \text{if its center is a hydroxylation site} \\ \text{non-Pro-hydroxylated peptide,} & \text{otherwise} \end{cases} \quad (2)$$

and:

$$\mathbf{P}(\mathbb{K}) \in \begin{cases} \text{Lys-hydroxylated peptide,} & \text{if its center is a hydroxylation site} \\ \text{non-Lys-hydroxylated peptide,} & \text{otherwise} \end{cases} \quad (3)$$

where $\in$ represents "a member of" in the set theory.

Figure 2. An illustration to show Chou's scheme for peptides with $(2\xi+1)$ residues and their centers being **(a)** proline and (**b**) lysine. Adapted from Chou [27,29] with permission.



As pointed out by a comprehensive review [30], there is no need to separate a benchmark dataset into a training dataset and a testing dataset for examining the performance of a prediction method if it is tested by the jackknife test or subsampling cross-validation test. Thus, the benchmark dataset for the current study can be formulated as:

$$\begin{cases} \mathbb{S}_{\text{HyP}} = \mathbb{S}^{+}_{\text{HyP}} \bigcup \mathbb{S}^{-}_{\text{HyP}} \\ \mathbb{S}_{\text{HyL}} = \mathbb{S}^{+}_{\text{HyL}} \bigcup \mathbb{S}^{-}_{\text{HyL}} \end{cases} \quad (4)$$

where $\mathbb{S}_{\text{HyP}}$ is the benchmark dataset for studying hydroxyproline residues, $\mathbb{S}_{\text{HyL}}$ the benchmark dataset for studying hydroxylysine residues, $\bigcup$ the symbol for "union" in the set theory, $\mathbb{S}^{+}_{\text{HyP}}$ contains

the samples for the Pro-hydroxylated peptide only, $\mathbb{S}^{-}_{\mathrm{HyP}}$ contains the non-Pro-hydroxylated peptide only (*cf*. Equation (2)), $\mathbb{S}^{+}_{\mathrm{HyL}}$ contains the samples for the Lys-hydroxylated peptide only and $\mathbb{S}^{-}_{\mathrm{HyL}}$ contains the non-Lys-hydroxylated peptide only (*cf*. Equation (3)).

After some preliminary trials, we found that $\xi = 6$ was a good choice for the current study. Accordingly, each of the samples extracted from proteins in this study is actually a $2\xi + 1 = 13$ tuple peptide. If the upstream or downstream in a peptide sample was $3 \le \xi < 6$, the lacking residues were filled with the dummy code, @. Furthermore, to reduce the redundancy and to avoid homology bias, those peptides were excluded from the benchmark datasets that had $\ge 80\%$ pairwise sequence identity to any other in a same subset.

Finally, we obtained that the benchmark dataset, $\mathbb{S}_{\mathrm{HyP}}$, contained $636 + 2699 = 3338$ peptide samples, of which 636 were Pro-hydroxylated peptides belonging the positive subset $\mathbb{S}^{+}_{\mathrm{HyP}}$, and 2669 were non-Pro-hydroxylated peptides belonging to the negative subset, $\mathbb{S}^{-}_{\mathrm{HyP}}$; and that the benchmark dataset, $\mathbb{S}_{\mathrm{HyL}}$, contained $107 + 836 = 943$ peptide samples, of which 107 were Lys-hydroxylated peptides belonging to the positive subset, $\mathbb{S}^{+}_{\mathrm{HyL}}$, and 836 were non-Lys-hydroxylated peptides belonging to the negative subset, $\mathbb{S}^{-}_{\mathrm{HyL}}$. For the reader's convenience, the peptide sequences, as well as their hydroxylation or non-hydroxylation sites in proteins are given in the Supplementary Information, S1 and S2, for $\mathbb{S}_{\mathrm{HyP}}$ and $\mathbb{S}_{\mathrm{HyL}}$, respectively.

2.2. Feature Vector Construction

To develop a statistical method for predicting the attribute of peptides in proteins, one of the fundamental procedures was to formulate the peptide samples with an effective mathematical expression that could really reflect the intrinsic correlation with the desired target. To realize this, various feature vectors (see, e.g., [17,31–36]) were proposed to express peptides by extracting their different features into the pseudo amino acid composition [37,38] or Chou's pseudo amino acid composition [39–41] or Chou's PseAAC (pseudo amino acid composition) [42,43].

According to [21], the general form of PseAAC for a protein or peptide, $\mathbf{P}$, can be formulated by:

$$\mathbf{P} = \left[\begin{array}{ccccccc} \psi_1 & \psi_2 & \cdots & \psi_u & \cdots & \psi_\Omega \end{array} \right]^{\mathbf{T}} \tag{5}$$

where $\mathbf{T}$ is the transpose operator, while Ω is an integer to reflect the vector's dimension. The value of Ω, as well as the components ψ_u ($u = 1, 2, \cdots, \Omega$) in Equation (5) will depend on how to extract the desired information from the protein or peptide sequence. Below, let us describe how to extract the useful information from the benchmark datasets, $\mathbb{S}_{\mathrm{Hyp}}$ and $\mathbb{S}_{\mathrm{Hyl}}$, to define the peptide samples via Equation (5).

Since each of the samples concerned is a 13-tuple peptide, Equation (1) can be simplified to a more convenient form given by:

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_7\cdots\mathrm{R}_{12}\mathrm{R}_{13} \tag{6}$$

where $R_7 = P$ or K, and R_i ($i = 1, 2, \cdots, 13; i \neq 7$) can be any of the 20 native amino acids or the dummy code @, as defined above. Hereafter, let us use the numerical codes 1, 2, 3, ⋯, 20 to represent the 20 native amino acids according to the alphabetic order of their single letter codes and use 21 to represent the dummy amino acid, @. Accordingly, the number of possible different dipeptides will be $21 \times 21 = 441$, and the number of dipeptide subsite positions on the sequence of Equation (6) will be $(13 - 2 + 1) = 12$.

Now, let us introduce the following 441×12 matrix, $\mathbb{Z}$, the so-called PSDP (position-specific dipeptide propensity) matrix [36], to define the component of Equation (5):

$$\mathbb{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,12} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,12} \\ \vdots & \vdots & \ddots & \vdots \\ z_{441,1} & z_{441,2} & \cdots & z_{441,12} \end{bmatrix} \tag{7}$$

where the element:

$$z_{i,j} = F^{+}(D_i \mid j) - F^{-}(D_i \mid j) \quad (i = 1,2, \cdots, 441; j = 1,2, \cdots, 12) \tag{8}$$

and:

$$D_1 = AA, D_2 = AC, \cdots, D_{21} = A@, \cdots, D_{440} = @Y, D_{441} = @@ \tag{9}$$

In Equation (5), $F^{+}(D_i \mid j)$ is the occurrence frequency of the i-th dipeptide (i = 1,2,⋯, 441) at the j-th subsite on the sequence of Equation (6) that can be easily derived from the positive dataset in the Supplementary Information S1 or S2; while $F^{-}(D_i \mid j)$ is the corresponding occurrence frequency, but derived from the negative dataset.

Thus, the peptide, $\mathbf{P}$, of Equation (6) can be uniquely defined via the general form of PseAAC (*cf.* Equation (5)) with its dimension $\Omega = 12$ and its u-th component given by:

$$\psi_u = \begin{cases} z_{1,u} & \text{when } R_u R_{u+1} = AA \\ z_{2,u} & \text{when } R_u R_{u+1} = AC \\ \vdots & \vdots \\ z_{21,u} & \text{when } R_u R_{u+1} = A@ \\ \vdots & \vdots \\ z_{441,u} & \text{when } R_u R_{u+1} = @@ \end{cases} \quad (1 \le u \le 12) \tag{10}$$

2.3. Prediction Algorithm

Suppose $\mathbb{P}^{+}$ are the standard vectors or norms for the peptide sequences in the positive benchmark dataset, $\mathbb{S}^{+}_{HyP}$ or $\mathbb{S}^{+}_{HyL}$, and $\mathbb{P}^{-}$ are those in the negative benchmark dataset, $\mathbb{S}^{-}_{HyP}$ or $\mathbb{S}^{-}_{HyL}$. Additionally, they are defined by:

$$\begin{cases} \mathbb{P}^{+} = \left[\begin{array}{ccccccc} \overline{\psi}_1^{+} & \overline{\psi}_2^{+} & \cdots & \overline{\psi}_u^{+} & \cdots & \overline{\psi}_\Omega^{+} \end{array} \right]^{\mathbf{T}} \\ \mathbb{P}^{-} = \left[\begin{array}{ccccccc} \overline{\psi}_1^{-} & \overline{\psi}_2^{-} & \cdots & \overline{\psi}_u^{-} & \cdots & \overline{\psi}_\Omega^{-} \end{array} \right]^{\mathbf{T}} \end{cases} \tag{11}$$

where:

$$\begin{cases} \overline{\psi}_u^{+} = \dfrac{1}{N^{+}} \sum_{k=1}^{N^{+}} \psi_{u,k}^{+} \\ \overline{\psi}_u^{-} = \dfrac{1}{N^{-}} \sum_{k=1}^{N^{-}} \psi_{u,k}^{-} \end{cases} \quad (u = 1, 2, \cdots, \Omega) \tag{12}$$

where N^{+} is the total number of Pro-hydroxylated peptides or Lys- hydroxylated peptides in the benchmark dataset, $\mathbb{S}_{\mathrm{HyP}}^{+}$ or $\mathbb{S}_{\mathrm{HyL}}^{+}$, and $\psi_{u,k}^{+}$ the u-th component for the k-th Pro-hydroxylated peptide or Lys-hydroxylated peptide in the PseAAC space (see Equations (5) and (10)); whereas N^{-} and $\psi_{u,k}^{-}$ have the same meanings, but are for the k-th non-Pro-hydroxylated peptide or non-Lys-hydroxylated peptide.

For a query peptide, $\mathbf{P}$, as formulated by Equation (5), suppose $\mathbb{D}\left(\mathbf{P}, \mathbb{P}^{+}\right)$ is its similarity to the norm of hydroxylated peptides and $\mathbb{D}\left(\mathbf{P}, \mathbb{P}^{-}\right)$ its similarity to the norm of non-hydroxylated peptides, as formulated by:

$$\begin{cases} \mathbb{D}\left(\mathbf{P}, \mathbb{P}^{+}\right) = \sqrt{\sum_{u=1}^{\Omega} \left(\psi_u - \overline{\psi}_u^{+}\right)^2} \\ \mathbb{D}\left(\mathbf{P}, \mathbb{P}^{-}\right) = \sqrt{\sum_{u=1}^{\Omega} \left(\psi_u - \overline{\psi}_u^{-}\right)^2} \end{cases} \tag{13}$$

Thus, according to the discriminant function algorithm [24,44], the attribute of the query peptide, $\mathbf{P}$, can be formulated as:

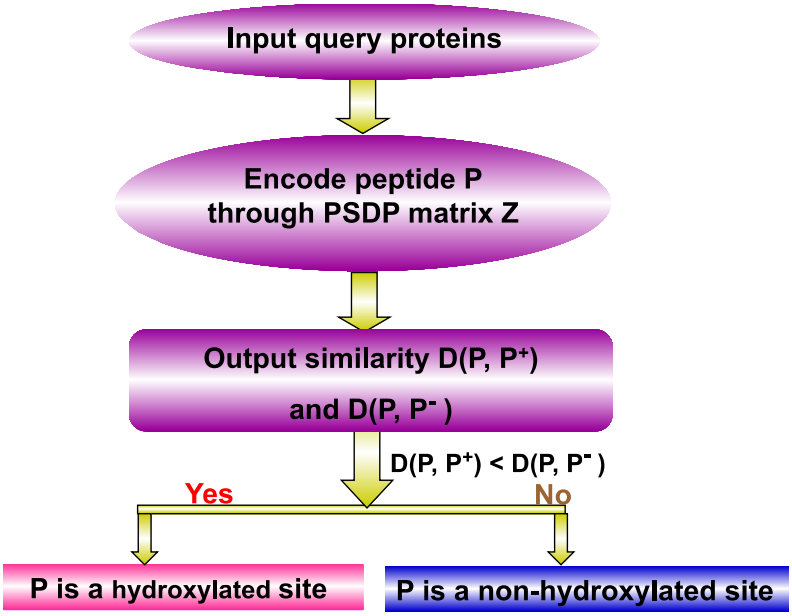
$$\mathbf{P} \in \begin{cases} \text{Hydroxylated peptide}, & \text{if } \mathbb{D}\left(\mathbf{P}, \mathbb{P}^{+}\right) < \mathbb{D}\left(\mathbf{P}, \mathbb{P}^{-}\right) \\ \text{non-hydroxylated peptide}, & \text{otherwise} \end{cases} \tag{14}$$

If there was a tie between $\mathbb{D}\left(\mathbf{P}, \mathbb{P}^{+}\right)$ and $\mathbb{D}\left(\mathbf{P}, \mathbb{P}^{-}\right)$, the query peptide would be randomly assigned between the hydroxylated peptide and non-hydroxylated peptide categories. However, this kind of tie case rarely happened and actually never happened in our study.

The predictor established via the above procedures is called iHyd-PseAAC, where "i" stands for the first character of "identify", "Hyd" for "hydroxylation" and "PseAAC" for the general form of the pseudo amino acid composition [21] used to formulate the peptide sequences.

A flowchart of the predictor is given in Figure 3 to illustrate how iHyd-PseAAC was working during the process of prediction.

Figure 3. Flowchart to show the process of how the iHyd-PseAAC (identify hydroxylation pseudo amino acid composition) predictor works in identifying the hydroxylated sites in proteins. PSDP, position-specific dipeptide propensity.



3. Experimental Section

3.1. A Set of Metrics for Measuring Prediction Quality

To provide a more intuitive and easier-to-understand method to measure the prediction quality, the following set of four metrics based on the formulation used by Chou [26–28] in predicting signal peptides was adopted. According to Chou's formulation, the sensitivity, specificity, overall accuracy, and Matthews correlation coefficient can be respectively expressed as [18,33,36,45]:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ \mathrm{Mcc} = \dfrac{1 - \left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-} \right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+} \right) \left(1 + \dfrac{N_-^+ - N_+^-}{N^-} \right)}} \end{cases} \quad (15)$$

where N^+ is the total number of the hydroxylated Pro-peptides (or Lys-peptides) investigated, while N_-^+ is the number of hydroxylated Pro-peptides (or Lys-peptides) incorrectly predicted as the non-hydroxylated Pro-peptides (or Lys-peptides); N^- is the total number of the non-hydroxylated Pro-peptides (or Lys-peptides) investigated, while N_+^- is the number of the non-hydroxylated Pro-peptides (or Lys-peptides) incorrectly predicted as the hydroxylated Pro-peptides (or Lys-peptides).

According to Equation (15), we can easily see the following. When $N_-^+ = 0$, meaning none of the hydroxylated Pro-peptides (or Lys-peptides) was mispredicted to be a non-hydroxylated Pro-peptide (or Lys-peptides), we have the sensitivity $\mathrm{Sn} = 1$; while $N_-^+ = N^+$, meaning that all the hydroxylated Pro-peptides (or Lys-peptides) were mispredicted to be the non-hydroxylated Pro-peptides (or Lys-peptides), we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$, meaning none of the non-hydroxylated Pro-peptides (or Lys-peptides) was mispredicted, we have the specificity $\mathrm{Sp} = 1$; while $N_+^- = N^-$, meaning all the non-hydroxylated Pro-peptides (or Lys-peptides) were incorrectly predicted as hydroxylated Pro-peptides (or Lys-peptides), we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$, meaning that none of the hydroxylated Pro-peptides (or Lys-peptides) in the dataset $\mathbb{S}_{\mathrm{HyP}}^+$ (or $\mathbb{S}_{\mathrm{HyL}}^+$) and none of the hydroxylated Pro-peptides (or Lys-peptides) in $\mathbb{S}_{\mathrm{HyP}}^-$ (or $\mathbb{S}_{\mathrm{HyL}}^-$) was incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$, meaning that all the hydroxylated Pro-peptides (or Lys-peptides) in the dataset $\mathbb{S}_{\mathrm{HyP}}^+$ (or $\mathbb{S}_{\mathrm{HyL}}^+$) and all the non-hydroxylated Pro-peptides (or Lys-peptides) in $\mathbb{S}_{\mathrm{HyP}}^-$ (or $\mathbb{S}_{\mathrm{HyL}}^-$) were mispredicted, we have the overall accuracy $\mathrm{Acc} = 0$. The Matthews correlation coefficient (MCC) is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$, meaning that none of the hydroxylated Pro-peptides (or Lys-peptides) in the dataset $\mathbb{S}_{\mathrm{HyP}}^+$ (or $\mathbb{S}_{\mathrm{HyL}}^+$) and none of the non-hydroxylated Pro-peptides (or Lys-peptides) in $\mathbb{S}_{\mathrm{HyP}}^-$ (or $\mathbb{S}_{\mathrm{HyL}}^-$) was mispredicted, we have $\mathrm{MCC} = 1$; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have $\mathrm{MCC} = 0$, meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$, we have $\mathrm{MCC} = 0$, meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier-to-understand when using Equation (15) to examine a predictor for its four metrics, particularly for its Mathew's correlation coefficient. It is instructive to point out that the metrics as defined in Equation (15) are valid for single-label systems; for multi-label systems [34,46–48], a set of more complicated metrics should be used, as given in [49].

3.2. Jackknife Cross-Validation

How to properly test a predictor for its anticipated success rates is very important in objectively evaluating its quality and potential application value. Generally speaking, the following three cross-validation methods are often used to examine the quality of a predictor and its effectiveness in practical application: the independent dataset test, the subsampling or the *K*-fold (such as 5-, 7- or 10-fold) crossover test and the jackknife test [50]. However, as elaborated by a penetrating analysis in [51], considerable arbitrariness exists in the independent dataset test. Furthermore, as demonstrated in [52], the subsampling (or *K*-fold crossover validation) test cannot avoid arbitrariness either. The jackknife test is the least arbitrary, which can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been widely recognized and increasingly utilized by investigators to examine the quality of various predictors (see, e.g., [32,53–62]). Accordingly, in this study, the jackknife test was also adopted to evaluate the accuracy of the current predictor. Listed in Table 1 are the jackknife test results obtained by iHyd-PseAAC on the benchmark datasets of Supplementary Information S1 and the benchmark dataset of Supplementary Information S2, respectively.

Table 1. The jackknife test results by the new predictor on the benchmark datasets in the Supplementary Information S1 and S2. HyP, hydroxyproline; HyL, hydroxylysine; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews correlation coefficient.

Benchmark dataset [a]	Sn (%)	Sp (%)	Acc (%)	MCC
Supplementary Information S1 for HyP	80.66	80.54	80.57	0.51
Supplementary Information S2 for HyL	87.85	83.01	83.56	0.50

[a] None of the sequences included has more than 80% pairwise sequence identity with any other.

To further demonstrate our predictor, the jackknife test was conducted on two more stringent benchmark datasets given in Supplementary Information S3 and S4, where none of the included sequences has more than 40% pairwise sequence identity with any other. The results thus obtained are listed in Table 2.

Table 2. The jackknife test results by the iHyd-PseAAC predictor on the benchmark datasets in Supplementary Information S3 and S4·

Benchmark dataset [a]	Sn (%)	Sp (%)	Acc (%)	MCC
Supplementary Information S1 for HyP	70.68	89.03	78.42	0.52
Supplementary Information S2 for HyL	79.04	86.37	83.12	0.51

[a] None of sequences included has more than 40% pairwise sequence identity with any other.

It is interesting to see by comparing the two tables that the rates of Acc and MCC are about the same in both cases. Although the rates of Sn in Table 2 are somewhat lower than those in Table 1, the rates of Sp in Table 2 are higher than those in Table 1. Accordingly, the success rates as measured by the four metrics in Equation (15) are basically about the same without dropping down significantly from using an 80% cutoff benchmark dataset to a 40% cutoff one, clearly indicating that iHyd-PseAAC is a useful predictor validated by rigorous cross-validation.

3.3. Test by Public Database

Moreover, from the Swiss-Prot database, we retrieved all those proteins whose hydroxylated sites were experimentally validated. After excluding those with a length less than 50 amino acids, we obtained 156 hydroxyproline proteins and 31 hydroxylysine proteins, respectively. Their codes and hydroxylated sites are given in Supplementary Information S5 and S6, respectively. The predicted results by iHyd-PseAAC on these real proteins are given in Table 3, from which we can see that the overall success rates thus obtained are quite consistent with those derived by the cross-validation on the benchmark datasets, as shown in Tables 1 and 2, fully indicating that iHyd-PseAAC is not only a valid predictor, but also may become a very useful high throughput toll for practical applications in this area.

Table 3. The overall success rates in identifying hydroxylated sites for the proteins retrieved from the Swiss-Prot database.

Hydroxylated type	Sn (%)	Sp (%)	Acc (%)
Proline	71.2	79.3	75.3
Lysine	72.7	80.6	76.8

4. Conclusions

As we can see from Table 1, the overall accuracies for the hydroxyproline and hydroxylysine cases are 80.57% and 83.56%, which are higher than 76.0% and 82.1%, the corresponding rates reported by Hu *et al*. [11]. At first glance, the value of MCC seems relatively low. Actually, as mentioned in Section 3.1, different from Acc, whose score is between 100% and 0%, the score for MCC is between one and −1, with zero meaning no better than random prediction. Accordingly, the MCC rate of 0.50–0.51 is generally deemed as a quite decent result. Particularly, the benchmark dataset in the current system is very imbalanced, which contains 636 hydroxylated peptides and 2669 non-hydroxylated peptides for proline, which also may lower the MCC rate. The same is true for the case of hydroxylation.

Particularly, no web-server was provided for the method in [11], and hence, its application value is quite limited. Actually, so far, no web-server whatsoever has been provided in this area. As pointed out in [63] and emphasized in a series of recent publications (see, e.g., [16–18,20,42,45]), one of the keys in developing a practically more useful prediction method is to establish a user-friendly and publicly accessible web-server. In view of this, the web-server for iHyd-PseAAC has been established, which can be freely accessed at http://app.aporc.org/iHyd-PseAAC/.

Furthermore, for the convenience of the vast majority of biologists and pharmaceutical scientists, below, let us provide a step-by-step guide to show how the users can easily get the desired result by using iHyd-PseAAC without the need to follow the complicated mathematical equations presented in this paper just for its integrity.

5. The User Guide for the Web-Server iHyd-PseAAC

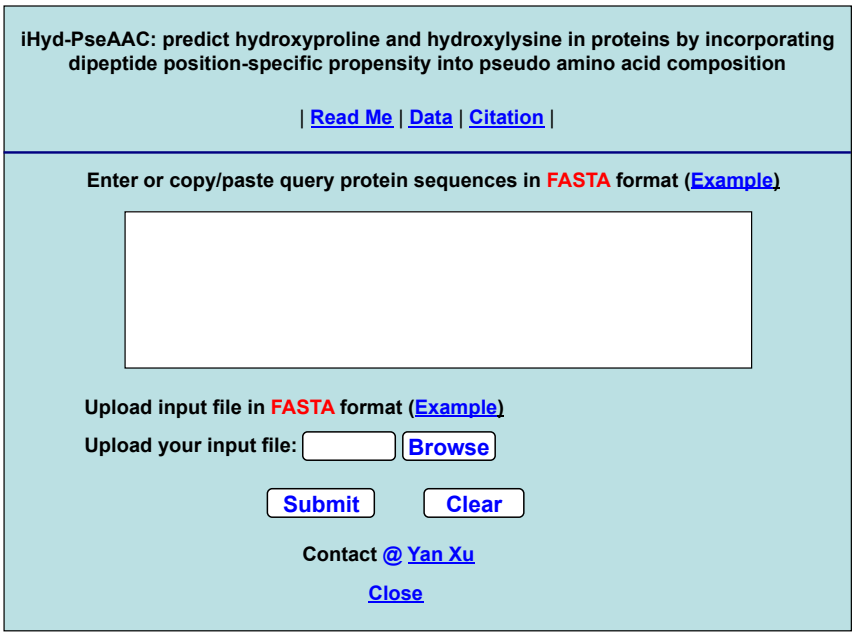
Step 1. Open the web-server at the site at http://app.aporc.org/iHyd-PseAAC/, and you will see the top page of the predictor on your computer screen, as shown in Figure 4. Click on the "Read Me" button to see a brief introduction about the iHyd-PseAAC predictor and the caveat when using it.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of Figure 4. The protein sequences should be in FASTA format. The input examples can be seen by clicking on the "Example" button right above the input box.

Step 4. Click on the "Citation" button to find the relevant paper that documents the detailed development and algorithm of iHyd-PseAAC.

Step 5. Click on the "Data" button to download the benchmark dataset used to train and test the iHyd-PseAAC predictor.

Figure 4. The top-page of the web-server, iHud-PseAAC, at http://app.aporc.org/iHyd-PseAAC/.



Acknowledgments

The authors wish to thank the three anonymous reviewers for their constructive comments, which were very helpful in strengthening the presentation of this study. This work is supported by the National Natural Science Foundation of China (No. 11301024, No. 11371365, No. 11101029, No. 31201002, No. 11071013) and the Fundamental Research Funds for the Central Universities.

Author Contributions

Xu and Chou conceived and designed the experiments. Wen and Shao processed and analyzed data. Xu and Deng wrote the manuscript.

Supplementary Information

Supplementary Information S1. The positive dataset, $\mathbb{S}^{+}_{\mathrm{HyP}}$, and negative dataset, $\mathbb{S}^{-}_{\mathrm{HyP}}$, contain 636 hydroxyproline peptide fragments and 2699 non-hydroxyproline peptide fragments, respectively. None of the sequences included has $\geq 80\%$ pairwise sequence identity with any other in the same subset.

Supplementary Information S2. The positive dataset, $\mathbb{S}^{+}_{\mathrm{HyL}}$, and negative dataset, $\mathbb{S}^{-}_{\mathrm{HyL}}$, contain 107 hydroxylysine peptide fragments and 836 non-hydroxylysine peptide fragments, respectively. None of the sequences included has $\geq 80\%$ pairwise sequence identity with any other in the same subset.

Supplementary Information S3. The positive dataset, $\mathbb{S}^{+}_{\mathrm{Hyp}}$, and negative dataset, $\mathbb{S}^{-}_{\mathrm{Hyp}}$, contain 306 hydroxyproline peptide fragments and 1035 non-hydroxyproline peptide fragments, respectively. None of the sequences included has $\geq 40\%$ pairwise sequence identity with any other in the same subset.

Supplementary Information S4. The positive dataset, $\mathbb{S}^{+}_{\mathrm{Hyl}}$, contains 44 hydroxylysine peptide fragments, and the negative dataset, $\mathbb{S}^{-}_{\mathrm{Hyl}}$, contains 528 non-hydroxylysine peptide fragments. None of the sequences included has $\geq 40\%$ pairwise sequence identity with any other in the same subset.

Supplementary Information S5. The 156 experimentally validated hydroxyproline proteins and their hydroxylated sites were retrieved from the Swiss-Prot database.

Supplementary Information S6. The 31 experimentally validated hydroxylysine proteins and their hydroxylated sites were retrieved from the Swiss-Prot database.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. 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*, 535–546.
2. Yamauchi, M.; Shiiba, M. Lysine hydroxylation and cross-linking of collagen. *Methods Mol. Biol.* **2008**, *446*, 95–108.
3. Chopra, R.K.; Ananthanarayanan, V.S. Conformational implications of enzymatic proline hydroxylation in collagen. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 7180–7184.
4. Krane, S.M. The importance of proline residues in the structure, stability and susceptibility to proteolytic degradation of collagens. *Amino Acids* **2008**, *35*, 703–710.
5. Palfi, V.K.; Perczel, A. How stable is a collagen triple helix? An ab initio study on various collagen and beta-sheet forming sequences. *J. Comput. Chem.* **2008**, *29*, 1374–1386.
6. Guszczyn, T.; Sobolewski, K. Deregulation of collagen metabolism in human stomach cancer. *Pathobiology* **2004**, *71*, 308–313.
7. Sunila, E.S.; Kuttan, G. A preliminary study on antimetastatic activity of Thuja occidentalis L. in mice model. *Immunopharmacol. Immunotoxicol.* **2006**, *28*, 269–280.
8. Guruvayoorappan, C.; Kuttan, G. Anti-metastatic effect of Biophytum sensitivum is exerted through its cytokine and immunomodulatory activity and its regulatory effect on the activation and nuclear translocation of transcription factors in B16F-10 melanoma cells. *J. Exp. Ther. Oncol.* **2008**, *7*, 49–63.
9. Richards, A.A.; Stephens, T.; Charlton, H.K.; Jones, A.; Macdonald, G.A.; Prins, J.B.; Whitehead, J.P. Adiponectin multimerization is dependent on conserved lysines in the collagenous domain: Evidence for regulation of multimerization by alterations in posttranslational modifications. *Mol. Endocrinol.* **2006**, *20*, 1673–1687.
10. Yang, Z.R. Predict collagen hydroxyproline sites using support vector machines. *J. Comput. Biol.* **2009**, *16*, 691–702.
11. Hu, L.L.; Niu, S.; Huang, T.; Wang, K.; Shi, X.H.; Cai, Y.D. Prediction and analysis of protein hydroxyproline and hydroxylysine. *PLoS One* **2010**, *5*, e15917.

12. Garnier, J.; Osguthorpe, D.J.; Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120*, 97–120.
13. Guo, S.H.; Deng, E.Z.; Xu, L.Q.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **2014**, doi:10.1093/bioinformatics/btu083.
14. Liu, B.; Zhang, D.; Xu, R.; Xu, J.; Wang, X.; Chen, Q.; Dong, Q.; Chou, K.C. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **2014**, *30*, 472–479.
15. Fan, Y.N.; Xiao, X.; Min, J.L.; Chou, K.C. iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. *Int. J. Mol. Sci.* **2014**, *15*, 4915–4937.
16. Qiu, W.R.; Xiao, X.; Chou, K.C. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **2014**, *15*, 1746–1766.
17. Min, J.L.; Xiao, X.; Chou, K.C. iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed Res. Int.* **2013**, *2013*, 701317, doi:10.1155/2013/701317.
18. Chen, W.; Feng, P. M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition *Nucleic Acids Res.* **2013**, *41*, e69.
19. Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **2013**, *442*, 118–125.
20. Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol.* **2013**, *337C*, 71–79.
21. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247.
22. Lu, C.T.; Huang, K.Y.; Su, M.G.; Lee, T.Y.; Bretana, N.A.; Chang, W.C.; Chen, Y.J.; 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.* **2013**, *41*, D295–D305.
23. Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **1993**, *268*, 16938–16948.
24. Chou, K.C. Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.* **1996**, *233*, 1–14.
25. Chou, K.C. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.* **1995**, *4*, 1365–1383.
26. Chou, K.C. Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 136–139.
27. Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.* **2001**, *14*, 75–79.
28. Chou, K.C. Prediction of signal peptides using scaled window. *Peptides* **2001**, *22*, 1973–1979.
29. Chou, K.C. Review: Prediction of protein signal sequences. *Curr. Protein Peptide Sci.* **2002**, *3*, 615–622.

30. Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1–16.
31. Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40.
32. Chen, Y.K.; Li, K.B. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *318*, 1–12.
33. Xu, Y.; Ding, J.; Wu, L.Y.; Chou, K.C. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition *PLoS One* **2013**, *8*, e55844.
34. Xiao, X.; Wang, P.; Lin, W.Z.; Jia, J.H.; Chou, K.C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **2013**, *436*, 168–177.
35. Mohabatkar, H.; Beigi, M.M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of allergenic proteins by means of the concept of chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.* **2013**, *9*, 133–137.
36. Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* **2013**, *1*, e171.
37. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct. Funct. Genet.* **2001**, *43*, 246–255.
38. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19.
39. Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **2012**, *425*, 117–119.
40. Georgiou, D.N.; Karakasidis, T.E.; Megaritis, A.C. A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. *Open Bioinforma. J.* **2013**, *7*, 41–48.
41. Liu, B.; Wang, X.; Zou, Q.; Dong, Q.; Chen, Q. Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Inform.* **2013**, *32*, 775–782.
42. Lin, S.X.; Lapointe, J. Theoretical and experimental biology in one. *J. Biomed. Sci. Eng.* **2013**, *6*, 435–442.
43. Cao, D.S.; Xu, Q.S.; Liang, Y.Z. Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, *29*, 960–962.
44. Chou, K.C.; Tomasselli, A.L.; Reardon, I.M.; Heinrikson, R.L. Predicting HIV protease cleavage sites in proteins by a discriminant function method. *Proteins: Struct. Funct. Genet.* **1996**, *24*, 51–72.
45. Chen, W.; Lin, H.; Feng, P.M.; Ding, C.; Zuo, Y.C.; Chou, K.C. iNuc-PhysChem: A sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One* **2012**, *7*, e47843.

46. Chou, K.C.; Shen, H.B. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728–1734.
47. Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **2012**, *8*, 629–641.
48. Shen, H.B.; Chou, K.C. Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun.* **2007**, *355*, 1006–1011.
49. Chou, K.C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Mol. Biosyst.* **2013**, *9*, 1092–1100.
50. Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.
51. Chou, K.C.; Shen, H.B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153–162.
52. Chou, K.C.; Shen, H.B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci.* **2010**, *2*, 1090–1103.
53. Fan, G.L.; Li, Q.Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. *Amino Acids* **2012**, *43*, 545–555.
54. Fan, G.L.; Li, Q.Z. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *334*, 45–51.
55. Huang, C.; Yuan, J.Q. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. *J. Theor. Biol.* **2013**, *335*, 205–212.
56. Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol.* **2008**, *252*, 350–356.
57. Lin, H.; Wang, H.; Ding, H.; Chen, Y.L.; Li, Q.Z. Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. *Acta Biotheor.* **2009**, *57*, 321–330.
58. Qiu, J.D.; Huang, J.H.; Liang, R.P.; Lu, X.Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: An approach from discrete wavelet transform. *Anal. Biochem.* **2009**, *390*, 68–73.
59. Sahu, S.S.; Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **2010**, *34*, 320–327.
60. Sun, X.Y.; Shi, S.P.; Qiu, J.D.; Suo, S.B.; Huang, S.Y.; Liang, R.P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. BioSyst.* **2012**, *8*, 3178–3184.
61. Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z.; Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.* **2009**, *259*, 366–372.

62. Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **2007**, *248*, 546–551.
63. Chou, K.C.; Shen, H.B. Review: Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.* **2009**, *2*, 63–92, doi:10.4236/ns.2009.12011.