

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/264794376>

iNitro-Tyr: Prediction of Nitrotyrosine Sites in Proteins with General Pseudo Amino Acid Composition

Article in PLoS ONE · August 2014

DOI: 10.1371/journal.pone.0105018 · Source: PubMed

CITATIONS

206

READS

805

6 authors, including:



Yan Xu

China University of Petroleum

55 PUBLICATIONS 1,789 CITATIONS

SEE PROFILE



Ling-Yun Wu

Chinese Academy of Sciences

103 PUBLICATIONS 3,389 CITATIONS

SEE PROFILE



Kuo-Chen Chou

Gordon Life Science Institute

827 PUBLICATIONS 75,622 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Calcium binding proteins [View project](#)



prediction [View project](#)



iNitro-Tyr: Prediction of Nitrotyrosine Sites in Proteins with General Pseudo Amino Acid Composition

Yan Xu^{1*}, Xin Wen¹, Li-Shu Wen², Ling-Yun Wu³, Nai-Yang Deng⁴, Kuo-Chen Chou^{5,6}

1 Department of Information and Computer Science, University of Science and Technology Beijing, Beijing, China, **2** College of Sciences, Liaoning Shiyu University, FuShun, China, **3** Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, **4** College of Science, China Agricultural University, Beijing, China, **5** Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia, **6** Gordon Life Science Institute, Boston, Massachusetts, United States of America

Abstract

Nitrotyrosine is one of the post-translational modifications (PTMs) in proteins that occurs when their tyrosine residue is nitrated. Compared with healthy people, a remarkably increased level of nitrotyrosine is detected in those suffering from rheumatoid arthritis, septic shock, and coeliac disease. Given an uncharacterized protein sequence that contains many tyrosine residues, which one of them can be nitrated and which one cannot? This is a challenging problem, not only directly related to in-depth understanding the PTM's mechanism but also to the nitrotyrosine-based drug development. Particularly, with the avalanche of protein sequences generated in the postgenomic age, it is highly desired to develop a high throughput tool in this regard. Here, a new predictor called "iNitro-Tyr" was developed by incorporating the position-specific dipeptide propensity into the general pseudo amino acid composition for discriminating the nitrotyrosine sites from non-nitrotyrosine sites in proteins. It was demonstrated via the rigorous jackknife tests that the new predictor not only can yield higher success rate but also is much more stable and less noisy. A web-server for iNitro-Tyr is accessible to the public at <http://app.aporc.org/iNitro-Tyr/>. For the convenience of most experimental scientists, we have further provided a protocol of step-by-step guide, by which users can easily get their desired results without the need to follow the complicated mathematics that were presented in this paper just for the integrity of its development process. It has not escaped our notice that the approach presented here can be also used to deal with the other PTM sites in proteins.

Citation: Xu Y, Wen X, Wen L-S, Wu L-Y, Deng N-Y, et al. (2014) iNitro-Tyr: Prediction of Nitrotyrosine Sites in Proteins with General Pseudo Amino Acid Composition. PLoS ONE 9(8): e105018. doi:10.1371/journal.pone.0105018

Editor: Reiner Albert Veitia, Institut Jacques Monod, France

Received: April 16, 2014; **Accepted:** July 16, 2014; **Published:** August 14, 2014

Copyright: © 2014 Xu et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work is supported by the National Natural Science Foundation of China (No. 11301024, No. 11371365, No. 11101029, No. 31201002, No. 11071013). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* Email: xuyan@ustb.edu.cn

Introduction

As one of the post-translational modifications (PTMs) of proteins, nitrotyrosine is a product of tyrosine nitration mediated by reactive nitrogen species such as peroxynitrite anion and nitrogen dioxide (Fig. 1). Compared with the fluids from healthy people, a remarkably increased level of nitrotyrosine is detected in those suffering from rheumatoid arthritis, septic shock, and coeliac disease. Accordingly, knowledge of nitrotyrosine sites in proteins is very useful for both basic research and drug development. Although conventional experimental methods did provide useful insight into the biological roles of tyrosine nitration [1–3], it is time-consuming and expensive to determine the nitrotyrosine sites based on the experimental approach alone. Particularly, identification of endogenous 3-N-Tyr modifications remains largely elusive (see, e.g., [4–7]). With the avalanche of protein sequences generated in the postgenomic age, it is highly desired to develop computational methods for identifying the nitrotyrosine sites in proteins. The present study was initiated in an attempt to propose a new method for identifying the nitrotyrosine sites in proteins in hope that it can play a complementary role with the existing methods in this area.

As summarized in [8] and demonstrated in a series of recent publications [9–21], to establish a really useful statistical predictor for a biological system, we need to consider the following

procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly capture their essence and intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy; (v) establish a user-friendly web-server that is accessible to the public. Below, let us describe how to deal with these steps one by one.

Materials and Methods

1. Benchmark Dataset

To develop a statistical predictor, it is fundamentally important to establish a reliable and stringent benchmark dataset to train and test the predictor. If the benchmark dataset contains some errors, the predictor trained by it must be unreliable and the accuracy tested by it would be completely meaningless.

For facilitating description later, let us adopt the Chou's peptide formulation here that was used for studying HIV protease cleavage sites [22,23], specificity of GalNAc-transferase [24], and signal peptide cleavage sites [25]. According to Chou's scheme, a potential nitrotyrosine peptide, i.e., a peptide with Tyr (namely Y) located at its center (Fig. 2), can be expressed as

$$\mathbf{P}_{\xi}(\mathbb{Y}) = \mathbf{R}_{-\xi} \mathbf{R}_{-(\xi-1)} \cdots \mathbf{R}_{-2} \mathbf{R}_{-1} \mathbb{Y} \mathbf{R}_{+1} \mathbf{R}_{+2} \cdots \mathbf{R}_{+(\xi-1)} \mathbf{R}_{+\xi} \quad (1)$$

where the subscript ξ is an integer, $\mathbf{R}_{-\xi}$ represents the ξ -th upstream amino acid residue from the center, \mathbf{R}_{ξ} the ξ -th downstream amino acid residue, and so forth. A $(2\xi+1)$ -tuple peptide $\mathbf{P}_{\xi}(\mathbb{Y})$ can be further classified into the following categories:

$$\mathbf{P}_{\xi}(\mathbb{Y}) \in \begin{cases} \mathbf{P}_{\xi}^{+}(\mathbb{Y}), & \text{if its center is a nitrotyrosine site} \\ \mathbf{P}_{\xi}^{-}(\mathbb{Y}), & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{P}_{\xi}^{+}(\mathbb{Y})$ represents a true nitrotyrosine peptide, $\mathbf{P}_{\xi}^{-}(\mathbb{Y})$ a false nitrotyrosine peptide, and \in represents “a member of” in the set theory.

As pointed out by a comprehensive review [26], 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 (K-fold) cross-validation test. Thus, the benchmark dataset for the current study can be formulated as

$$\mathbb{S}_{\xi} = \mathbb{S}_{\xi}^{+} \cup \mathbb{S}_{\xi}^{-} \quad (3)$$

where \mathbb{S}_{ξ}^{+} only contains the samples of $\mathbf{P}_{\xi}^{+}(\mathbb{Y})$, i.e., the nitrotyrosine peptides; \mathbb{S}_{ξ}^{-} only contains the samples of $\mathbf{P}_{\xi}^{-}(\mathbb{Y})$, i.e., the non-nitrotyrosine peptide (cf. **Eq. 2**); and \cup represents the symbol for “union” in the set theory.

Since the length of the peptide $\mathbf{P}_{\xi}(\mathbb{Y})$ is $2\xi+1$ (**Eq. 1**), the benchmark dataset with different values of ξ will contain peptides of different numbers of amino acid residues, as formulated by

$$\mathbb{S}_{\xi} \text{ contains the peptides of } \begin{cases} 13 \text{ residues,} & \text{when } \xi = 6 \\ 15 \text{ residues,} & \text{when } \xi = 7 \\ 17 \text{ residues,} & \text{when } \xi = 8 \\ 19 \text{ residues,} & \text{when } \xi = 9 \\ 21 \text{ residues,} & \text{when } \xi = 10 \\ \vdots & \vdots \end{cases} \quad (4)$$

The detailed procedures to construct \mathbb{S}_{ξ} are as follows. (i) Its elements were derived based on the same 546 source proteins used in [27] that contain 1,044 nitrotyrosine sites (see columns 1 and 2 of Supporting Information S1). (ii) Slide a flexible window of $2\xi+1$ amino acids (**Fig. 3**) along each of the 546 protein sequences taken from the Uni-Prot database (version 2014_01). (iii) Collect only those peptide segments with Y (tyrosine) at the center. (iv) If the upstream or downstream in a protein was less than ξ , the lacking residue was filled with a dummy residue “X” [28]. (v) Those peptide samples thus obtained were put into the positive subset \mathbb{S}_{ξ}^{+} if their centers have been experimentally confirmed as the nitrotyrosine sites; otherwise, into the negative subset \mathbb{S}_{ξ}^{-} .

By following the aforementioned procedures, five such benchmark datasets ($\mathbb{S}_{\xi=6}, \mathbb{S}_{\xi=7}, \mathbb{S}_{\xi=8}, \mathbb{S}_{\xi=9}$, and $\mathbb{S}_{\xi=10}$) had been constructed. Each of these datasets contained 1,044 nitrotyrosine peptides and 7,669 non-nitrotyrosine peptides. Note that the sample numbers thus obtained have some minor difference with those in [27]. This is because some proteins originally used in [27] have been removed or replaced in the updated version of the Uni-Prot database.

However, it was observed via preliminary trials that when $\xi=9$, i.e., the peptide samples concerned were formed by 19 residues, the corresponding results were most promising (see **Fig. 4** and **Fig. 5**). Accordingly, we choose $\mathbb{S}_{\xi=9}$ as the benchmark dataset for further investigation. Thus, **Eq. 3** can be reduced to

$$\mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \quad (5)$$

where $\mathbb{S} = \mathbb{S}_9$, $\mathbb{S}^{+} = \mathbb{S}_9^{+}$ containing 1,044 nitrotyrosine peptide samples, and $\mathbb{S}^{-} = \mathbb{S}_9^{-}$ containing 7,669 non-nitrotyrosine peptide samples. The detailed 19-tuple peptide sequences and their positions in proteins are given in Supporting Information S1.

2. Feature Vector and Pseudo Amino Acid Composition

One of the most important but also most difficult problems in computational biology today is how to effectively formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. This is because all the existing operation engines, such as correlation angle approach [29], covariance discriminant [30], neural network [31], support vector machine (SVM) [32], random forest [33], conditional random field [28], K-nearest neighbor (KNN) [34], OET-KNN [35], Fuzzy K-nearest neighbor [36], ML-KNN algorithm [37],

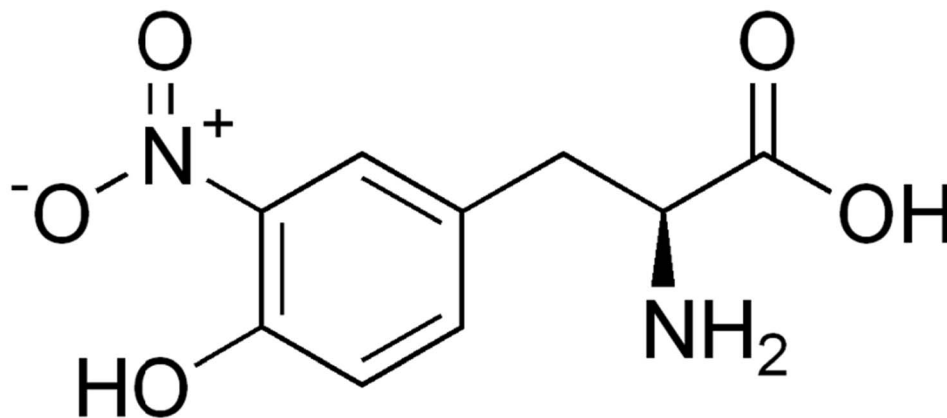


Figure 1. A schematic drawing to show protein nitrotyrosine.
doi:10.1371/journal.pone.0105018.g001

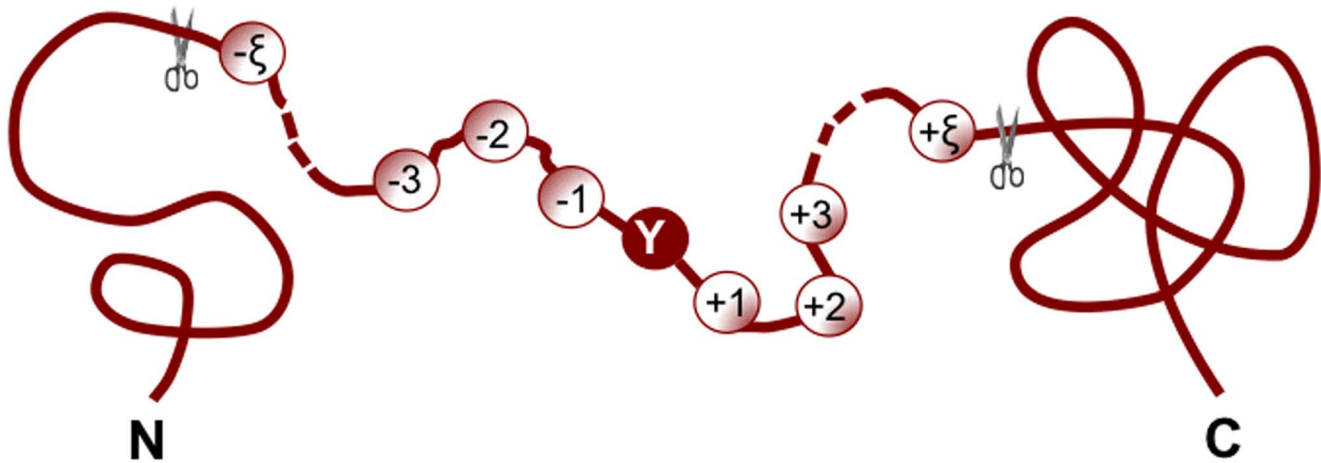


Figure 2. An illustration to show Chou's scheme for a peptide of $(2\xi + 1)$ residues with tyrosine (Y) at the center. Adapted from Chou [55,76] with permission.
doi:10.1371/journal.pone.0105018.g002

and SLLE algorithm [30], can only handle vector but not sequence samples. However, a vector defined in a discrete model may totally miss the sequence-order information. To deal with such a dilemma, the approach of pseudo amino acid composition [38] or Chou's PseAAC [39] was proposed. Ever since it was introduced in 2001 [38], the concept of PseAAC has been rapidly penetrated into almost all the areas of computational proteomics, such as in identifying bacterial virulent proteins [40], predicting anticancer peptides [41], predicting protein subcellular location [42], predicting membrane protein types [43], analyzing genetic sequence [44], predicting GABA(A) receptor proteins [45], identifying antibacterial peptides [46], predicting anticancer peptides [41], identifying allergenic proteins [47], predicting metalloproteinase family [48], identifying GPCRs and their types [49], identifying protein quaternary structural attributes [50], among many others (see a long list of references cited in a 2014 article [51]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [9], as well as other biological samples (see, e.g., [52]). Because it has been widely and increasingly used, recently three types of powerful open access soft-ware, called 'PseAAC-Builder' [53], 'propy' [54], and 'PseAAC-General' [51], were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC.

According to a comprehensive review [8], PseAAC can be generally formulated as

$$\mathbf{P} = [\psi_1 \ \psi_2 \ \cdots \ \psi_u \ \cdots \ \psi_\Omega]^T \quad (6)$$

where \mathbf{T} is the transpose operator, while Ω an integer to reflect the vector's dimension. The value of Ω as well as the components $\psi_u (u = 1, 2, \dots, \Omega)$ in **Eq. 6** will depend on how to extract the desired information from a protein/peptide sequence. Below, let us describe how to extract the useful information from the benchmark datasets to define the peptide samples via **Eq. 6**.

For convenience in formulation, let rewrite **Eq. 1** as follows

$$\mathbf{P}_\xi = \mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_\xi \mathbf{R}_{\xi+1} \cdots \mathbf{R}_{2\xi} \mathbf{R}_{2\xi+1} \quad (7)$$

where $\mathbf{R}_{\xi+1}$, the residue at the center of the peptide, is tyrosine (Y), and all the other residues $\mathbf{R}_i (i \neq \xi + 1)$ can be any of the 20

native amino acids or the dummy code X 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 X. Accordingly, the number of possible different dipeptides will be $21 \times 21 = 441$, and the number of dipeptide subsite positions on the sequence of **Eq. 7** will be $(2\xi + 1 - 1) = 2\xi$.

Now, let us introduce a positive and a negative PSDP (position-specific dipeptide propensity) matrix, as given below

$$\mathbb{Z}^+(\xi) = \begin{bmatrix} z_{1,1}^+ & z_{1,2}^+ & \cdots & z_{1,2\xi}^+ \\ z_{2,1}^+ & z_{2,2}^+ & \cdots & z_{2,2\xi}^+ \\ \vdots & \vdots & \ddots & \vdots \\ z_{441,1}^+ & z_{441,2}^+ & \cdots & z_{441,2\xi}^+ \end{bmatrix} \quad (8a)$$

$$\mathbb{Z}^-(\xi) = \begin{bmatrix} z_{1,1}^- & z_{1,2}^- & \cdots & z_{1,2\xi}^- \\ z_{2,1}^- & z_{2,2}^- & \cdots & z_{2,2\xi}^- \\ \vdots & \vdots & \ddots & \vdots \\ z_{441,1}^- & z_{441,2}^- & \cdots & z_{441,2\xi}^- \end{bmatrix} \quad (8b)$$

where the element

$$\begin{cases} z_{ij}^+ = F^+(D_i|j) \\ z_{ij}^- = F^-(D_i|j) \end{cases} \quad (i = 1, 2, \dots, 441; j = 1, 2, \dots, 2\xi) \quad (9)$$

and

$$D_1 = AA, D_2 = AC, D_3 = AD, \dots, D_{440} = XY, D_{441} = XX \quad (10)$$

In **Eq. 9**, $F^+(D_i|j)$ is the occurrence frequency of the i -th dipeptide ($i = 1, 2, \dots, 441$) at the j -th subsite on the sequence of **Eq. 7** (or the j -th column in the positive subset dataset \mathbb{S}^+) that

can be easily derived using the method described in [55] from the sequences in the Supporting Information S1; while $F^-(D_i|j)$ is the corresponding occurrence frequency but derived from the negative subset dataset \mathbb{S}^- . Thus, for the peptide sequence of **Eq. 7**, its attribute to the positive set \mathbb{S}^+ or negative set \mathbb{S}^- can be formulated by a 2ξ -D (dimension) vector \mathbf{P}^+ or \mathbf{P}^- , as defined by [23]

$$\mathbf{P}_\xi^+ = [\psi_1^+ \quad \psi_2^+ \quad \cdots \quad \psi_u^+ \quad \cdots \quad \psi_{2\xi}^+]^T \quad (11a)$$

$$\mathbf{P}_\xi^- = [\psi_1^- \quad \psi_2^- \quad \cdots \quad \psi_u^- \quad \cdots \quad \psi_{2\xi}^-]^T \quad (11b)$$

where

$$\psi_u^+ = \begin{cases} z_{1,u}^+ & \text{when } R_u R_{u+1} = AA \\ z_{2,u}^+ & \text{when } R_u R_{u+1} = AC \\ \vdots & \vdots \\ z_{21,u}^+ & \text{when } R_u R_{u+1} = AX \\ z_{22,u}^+ & \text{when } R_u R_{u+1} = CA \\ \vdots & \vdots \\ z_{441,u}^+ & \text{when } R_u R_{u+1} = XX \end{cases} \quad [u=1,2,\dots,2\xi=\Omega] \quad (12a)$$

$$\psi_u^- = \begin{cases} z_{1,u}^- & \text{when } R_u R_{u+1} = AA \\ z_{2,u}^- & \text{when } R_u R_{u+1} = AC \\ \vdots & \vdots \\ z_{21,u}^- & \text{when } R_u R_{u+1} = AX \\ z_{22,u}^- & \text{when } R_u R_{u+1} = CA \\ \vdots & \vdots \\ z_{441,u}^- & \text{when } R_u R_{u+1} = XX \end{cases} \quad [u=1,2,\dots,2\xi=\Omega] \quad (12b)$$

where R_u and R_{u+1} represent the residues in the u -th and $(u+1)$ -th positions of the peptide concerned.

3. Discriminant Function Approach

Now in the 2ξ -D space, let us define an ideal nitrotyrosine peptide \mathbf{II}^+ [22] and an ideal non-nitrotyrosine peptide \mathbf{II}^- as expressed by

$$\mathbf{II}_\xi^+ = \begin{bmatrix} \lambda_1^+ \\ \lambda_2^+ \\ \vdots \\ \lambda_u^+ \\ \vdots \\ \lambda_{2\xi}^+ \end{bmatrix} \quad \mathbf{II}_\xi^- = \begin{bmatrix} \lambda_1^- \\ \lambda_2^- \\ \vdots \\ \lambda_u^- \\ \vdots \\ \lambda_{2\xi}^- \end{bmatrix} \quad (13)$$

where λ_i^+ ($i=1,2,\dots,2\xi$) is the upper limit of the corresponding matrix element in **Eq. 12a**, and λ_i^- ($i=1,2,\dots,2\xi$) is the upper limit of the corresponding matrix element in **Eq. 12b**. Theoretically speaking, each of these hypothetical upper limits in **Eq. 13** should be 1 [23]. Thus, the similarity score of \mathbf{P}_ξ^+ with \mathbf{II}^+ and that of \mathbf{P}_ξ^- with \mathbf{II}^- can be defined as

$$\begin{cases} \mathbf{P}_\xi^+ \cdot \mathbf{II}_\xi^+ = \sum_{u=1}^{2\xi} \psi_u^+ \\ \mathbf{P}_\xi^- \cdot \mathbf{II}_\xi^- = \sum_{u=1}^{2\xi} \psi_u^- \end{cases} \quad (14)$$

Similar to the treatment in [23], let us define a discriminant function Δ given by

$$\Delta_\xi = (\mathbf{P}_\xi^+ \cdot \mathbf{II}_\xi^+ - \mathbf{P}_\xi^- \cdot \mathbf{II}_\xi^-) - \mathcal{R} = \sum_{u=1}^{2\xi} (\psi_u^+ - \psi_u^-) - \mathcal{R} \quad (15)$$

where \mathcal{R} is the adjust parameter used to optimize the overall success rate when the positive and negative benchmark datasets are highly imbalanced in size. Now the peptide \mathbf{P}_ξ of **Eq. 7** can be identified according to the following rule

$$\begin{cases} \mathbf{P}_\xi \text{ belongs to nitrotyrosine peptide,} & \text{if } \Delta_\xi > 0 \\ \mathbf{P}_\xi \text{ belongs to non-nitrotyrosine peptide,} & \text{if } \Delta_\xi \leq 0 \end{cases} \quad (16)$$

The predictor obtained via the above procedures is called **iNitro-Tyr**. How to properly and objectively evaluate the anticipated accuracy of a new predictor and how to make it easily accessible and user-friendly are the two key issues that will have important impacts on its application value [56]. Below, let us address these problems.

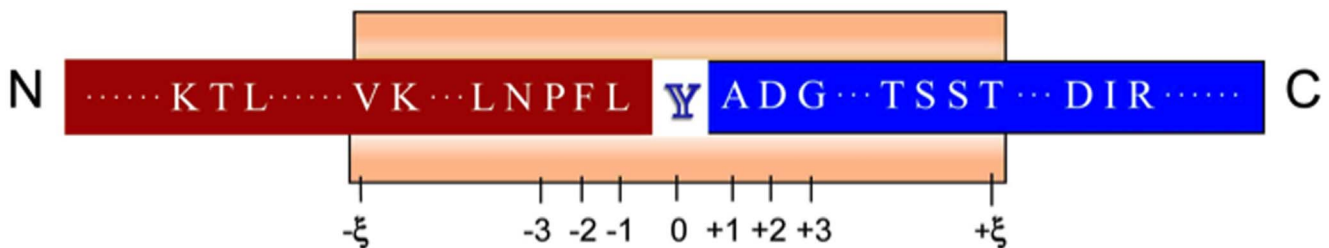


Figure 3. Illustration to show the peptide segment highlighted by sliding the scaled window $[-\xi, +\xi]$ along a protein sequence. During the sliding process, the scales on the window are aligned with different amino acids so as to define different peptide segments. When, and only when, the scale 0 is aligned with Y (tyrosine), is the $(2\xi+1)$ -tuple peptide segment seen within the window regarded as a potential nitrotyrosine peptide. Adapted from Chou [55,77] with permission. doi:10.1371/journal.pone.0105018.g003

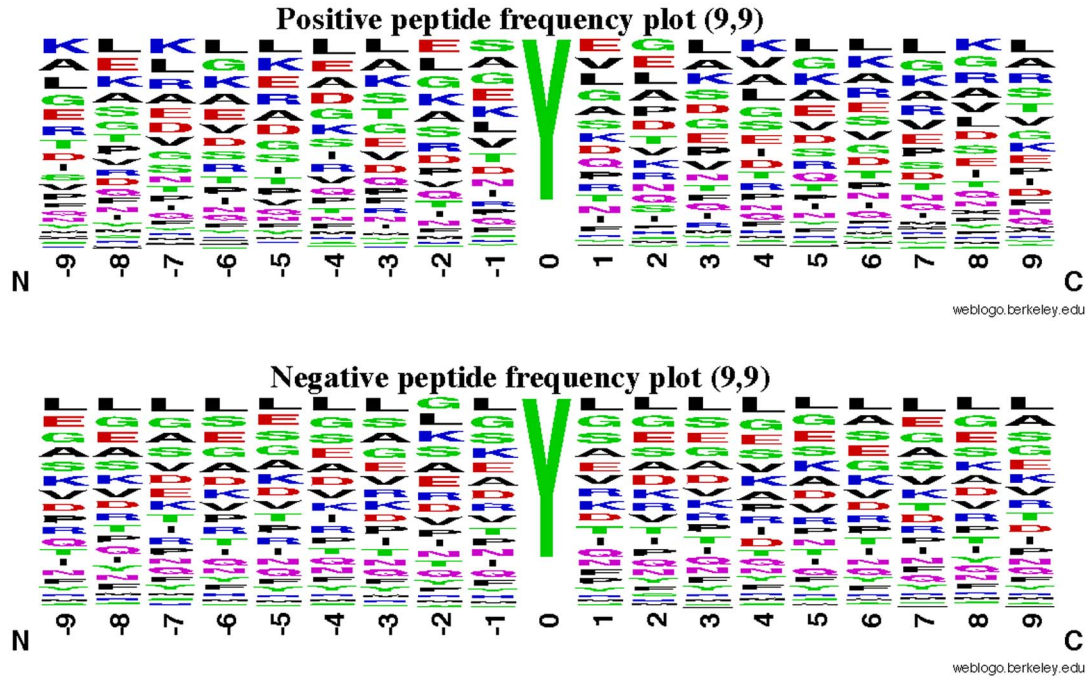


Figure 4. A sequence logo plot to show the difference between the positive and negative peptides. The window's size is 19 when $\xi = 9$. See Eq. 1 and the legend of Fig. 3 for further explanation. doi:10.1371/journal.pone.0105018.g004

Results and Discussion

1. Metrics for Scoring Prediction Quality

In literature the following four metrics are often used to score the quality of a predictor at four different angles

$$\begin{cases} \text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \end{cases} \quad (17)$$

where TP represents the number of the true positive; TN, the number of the true negative; FP, the number of the false positive; FN, the number of the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient. To most biologists, unfortunately, the four metrics as formulated in **Eq. 17** are not quite intuitive and easy-to-understand, particularly the equation for MCC. Here let us adopt the formulation proposed recently in [9,11,28] based on the symbols introduced by Chou [25,55] in predicting signal peptides. According to the formulation, the same four metrics can be expressed as

$$\begin{cases} \text{Sn} = 1 - \frac{N_+^+}{N_+^+}, & 0 \leq \text{Sn} \leq 1 \\ \text{Sp} = 1 - \frac{N_+^-}{N_+^-}, & 0 \leq \text{Sp} \leq 1 \\ \text{Acc} = 1 - \frac{N_+^+ + N_+^-}{N_+^+ + N_+^-}, & 0 \leq \text{Acc} \leq 1 \\ \text{MCC} = \frac{1 - \left(\frac{N_+^+}{N_+^+} + \frac{N_+^-}{N_+^-} \right)}{\sqrt{\left(1 + \frac{N_+^- - N_+^+}{N_+^+} \right) \left(1 + \frac{N_+^- - N_+^+}{N_+^-} \right)}}, & -1 \leq \text{MCC} \leq 1 \end{cases} \quad (18)$$

where N_+^+ is the total number of the nitrotyrosine peptides investigated while N_+^- the number of the nitrotyrosine peptides incorrectly predicted as the non-nitrotyrosine peptides; N_-^+ the total number of the non-nitrotyrosine peptides investigated while N_-^- the number of the non-nitrotyrosine peptides incorrectly predicted as the nitrotyrosine peptides [57].

Now, it is crystal clear from **Eq. 18** that when $N_+^+ = 0$ meaning none of the nitrotyrosine peptides was incorrectly predicted to be a non-nitrotyrosine peptide, we have the sensitivity $\text{Sn} = 1$. When $N_+^+ = N_+^+$ meaning that all the nitrotyrosine peptides were incorrectly predicted as the non-nitrotyrosine peptides, we have the sensitivity $\text{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning none of the non-nitrotyrosine peptides was incorrectly predicted to be the nitrotyrosine peptide, we have the specificity $\text{Sp} = 1$; whereas $N_+^- = N_-^+$ meaning all the non-nitrotyrosine peptides were incorrectly predicted as the nitrotyrosine peptides, we have the specificity $\text{Sp} = 0$. When $N_+^+ = N_+^+ = 0$ meaning that none of nitrotyrosine peptides in the positive dataset \mathbb{S}^+ and none of the non-nitrotyrosine peptides in the negative dataset \mathbb{S}^- was incorrectly predicted, we have the overall accuracy $\text{Acc} = 1$ and $\text{MCC} = 1$; when $N_+^+ = N_+^+$ and $N_+^- = N_-^+$ meaning that all the nitrotyrosine peptides in the positive dataset \mathbb{S}^+ and all the non-nitrotyrosine peptides in the negative dataset \mathbb{S}^- were incorrectly predicted, we have the overall accuracy $\text{Acc} = 0$ and $\text{MCC} = -1$; whereas when $N_+^+ = N_+^+/2$ and $N_+^- = N_-^+/2$ we have $\text{Acc} = 0.5$ and $\text{MCC} = 0$ meaning no better than random prediction. As we can see from the above discussion based on **Eq. 18**, the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier-to-understand.

It is instructive to point out, however, the set of metrics in **Eqs. 17–18** is valid only for the single-label systems. For the multi-label systems, such as those for the subcellular localization of multiplex proteins (see, e.g., [58–62]) where a protein may have two or more

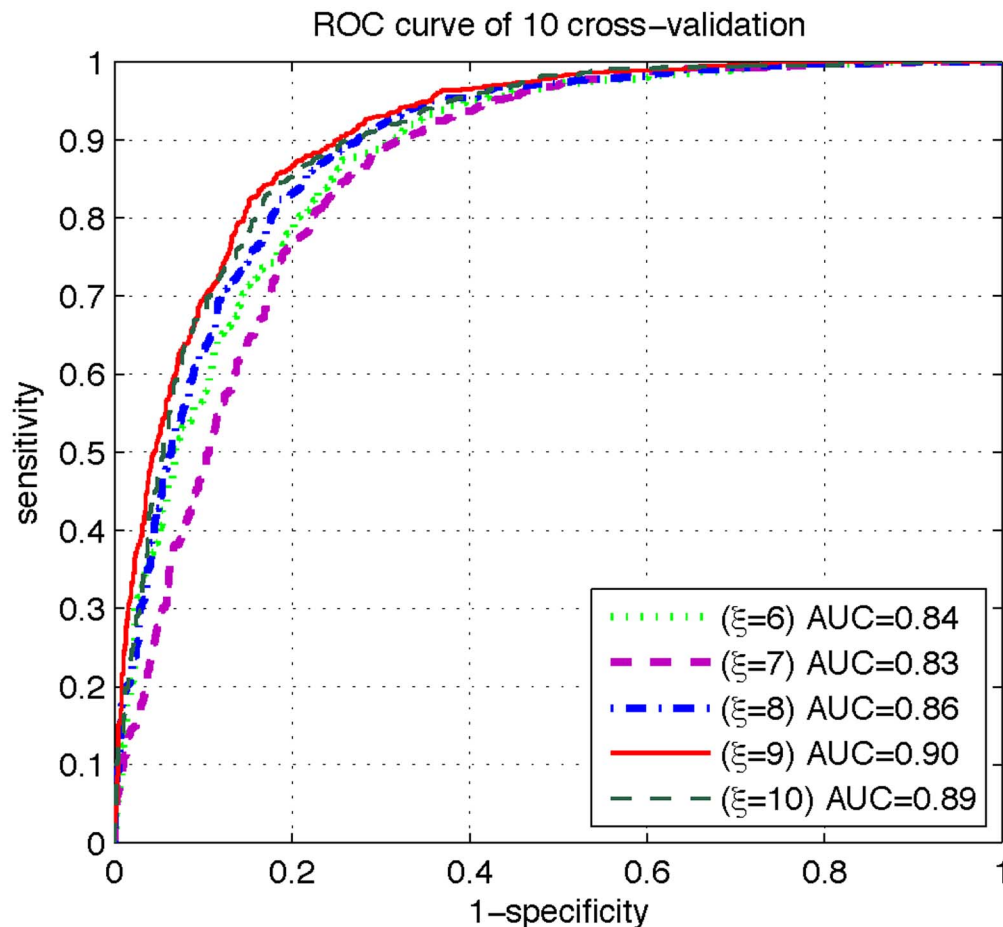


Figure 5. A plot to show the different ROC curves obtained by the 10-fold cross-validation under different ξ values. As we can see, when $\xi=9$, the corresponding AUC (i.e., the area under its curve) is the largest, meaning the most promising compared with the other values of ξ . doi:10.1371/journal.pone.0105018.g005

locations, and those for the functional types of antimicrobial peptides (see, e.g., [63] where a peptide may possess two or more functional types, a completely different set of metrics is needed as elaborated in [37].

2. Jackknife Cross-Validation

With a set of clear and valid metrics as defined in **Eq. 18** to measure the quality of a predictor, the next thing we need to consider is how to objectively derive the values of these metrics for a predictor.

In statistical prediction, the following three cross-validation methods are often used to calculate the metrics of **Eq. 18** for evaluating the quality of a predictor: independent dataset test, subsampling test, and jackknife test [64]. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [65]. The reasons are as follows. (i) For the independent dataset test, although all the samples used to test the predictor are outside the training dataset used to train it so as to exclude the “memory” effect or bias, the way of how to select the independent samples to test the predictor could be quite arbitrary unless the number of independent samples is sufficiently large. This kind of arbitrariness might result in completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent testing dataset might fail to keep so when tested by another independent testing dataset [64]. (ii) For the

subsampling test, the concrete procedure usually used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset, as demonstrated by Eqs.28–30 in [8]. Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as a good one. (iii) In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each sample will be in turn moved between the two. The jackknife test can exclude the “memory” effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine the quality of various predictors (see, e.g., [33,41,43,45–47,66–72]).

Table 1. Comparison of the new iNitro-Tyr predictor with the existing predictors in identifying the nitrotyrosine sites; the rates listed below were derived by the jackknife cross-validation on the 546 source proteins used in [27].

Predictor	Threshold	Acc (%)	MCC	Sn (%)	Sp (%)
GPS-YN02 ^a	High	82.57	0.1884	28.89	90.02
	Medium	79.60	0.2171	40.53	85.02
	Low	76.51	0.2335	50.09	90.18
iNitro-Tyr ^b		84.52	0.4905	81.76	85.89

^aAs reported in [27], where $\xi = 7$, i.e., the length of the potential nitrotyrosine peptides considered is $(2\xi + 1) = 15$.

^bSee Eqs. 15–16, where $\mathfrak{R} = 0.70$ and $\xi = 9$, i.e., the length of the potential nitrotyrosine peptides considered is $(2\xi + 1) = 19$.

doi:10.1371/journal.pone.0105018.t001

Accordingly, in this study we also used the jackknife cross-validation method to calculate the metrics in **Eq. 18** although it would take more computational time.

3. Comparison with Other Methods

The jackknife test results by iNitro-Tyr on the benchmark dataset $S = S^+ \cup S^-$ (cf. Supporting Information S1) for the four metrics defined in **Eq. 18** are listed in **Table 1**, where for facilitating comparison, the corresponding results by GPS-YN02 [27] with different thresholds are also given.

From the table, we can see the following facts. (i) The overall accuracy by the current iNitro-Tyr predictor is $\text{Acc} = 84.52\%$, which is higher than the overall accuracy by GPS-YN02 regardless what threshold is used for the latter. (ii) The Mathew's correlation coefficient obtained by iNitro-Tyr is $\text{MCC} = 0.4905$, which is significantly higher than that by GPS-YN02, indicating that the new predictor is more stable and less noisy. (iii) The sensitivity and specificity obtained by iNitro-Tyr are $\text{Sn} = 81.76\%$ and $\text{Sp} = 85.89\%$, which are much more evenly distributed than those by the GPS-YN02 predictor.

It is instructive to point out that, as shown by **Eqs. 12a** and **b**, the amino acid pairwise coupling effects [11] has been incorporated via the general form of PseAAC [8] to formulate the peptide samples. If, however, we just used the single amino acid specific position occurrence frequency to formulate the peptide samples, the corresponding prediction quality would drop down to $\text{Acc} = 44.88\%$ and $\text{MCC} = 0.1656$, clearly indicating that consideration of the amino acid pairwise coupling effects could significantly enhance the prediction quality, fully consistent with the reports by previous investigators [73,74], where it was observed that the prediction of protein secondary structural contents had been remarkably improved by taking into account the amino acid pairwise coupling effects.

Accordingly, compared with the best of existing predictors for identifying the nitrotyrosine sites in proteins, the new **iNitro-Tyr** predictor not only can yield higher or comparable accuracy, but is also much more stable and less noisy. It is anticipated that **iNitro-Tyr** may become a useful high throughput tool in this area, or at the very least play a complementary role to the existing predictors.

4. Web-Server and User Guide

For the convenience of most experimental scientists, we have established a web-server for the **iNitro-Tyr** predictor, with which users can easily get their desired results according to the steps below without the need to understand the mathematical equations in the method section.

Step 1. Open the web server at <http://app.aporc.org/iNitro-Tyr/> and you will see the top page of the predictor on your computer screen, as shown in **Fig. 6**. Click on the **Read Me**

button to see a brief introduction about **iNitro-Tyr** predictor and the caveat when using it.

Step 2. Either type or copy/paste the sequences of query proteins into the input box shown at the center of **Fig. 6**. All the input sequences should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with the symbol ">" in the first column, followed by lines of sequence data in which amino acids are represented using single-letter codes. Except for the mandatory symbol ">", all the other characters in the single initial line are optional and only used for the purpose of identification and description. The sequence ends if another line starting with the symbol ">" appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the **Example** button right above the open box. Note that if your input protein sequences should be formed by the 20 native amino acid codes (ACDEFGHIKLMNPQRSTVWY).

Step 3. Click on the **Submit** button to see the predicted results. For example, if you use the two query protein sequences in the **Example** window as the input, after clicking the **Submit** button, you will see the following on your screen. (i) The 1st protein (P05181) contains 18 Y residues; of which only those located at the sequence position 71, 318, 349, 381, and 423 are of nitrotyrosine site, while all the others are of non-nitrotyrosine site. (ii) The 2nd protein (P03023) contains 8 Y residues; of which only those located at the sequence positions 7, 12, 17, and 47 belong to the nitrotyrosine site, while all the others belong to non-nitrotyrosine site. All these results are fully consistent with experimental observations except for one Y residue at the position 349 in the 1st protein (P05181) that is actually non-nitrotyrosine site but was overpredicted as nitrotyrosine site.

Step 4. As shown on the lower panel of **Fig. 6**, you may also submit your query proteins in an input file (with FASTA format) via the "Browse" button. To see the sample of input file, click on the **Example** button right under the input box.

Step 5. Click on the **Data** button to download the benchmark dataset used to train and test the iNitro-Tyr predictor.

Conclusions

As one of the important posttranslational modifications (PTMs), nitrotyrosine is a product occurring in proteins when their tyrosine (Tyr or Y) residue is nitrated. Since a remarkably increasing level of nitrotyrosine is detected for those patients who have suffered from rheumatoid arthritis, septic shock, and coeliac disease, knowledge of nitrotyrosine is very useful for developing drugs against these diseases.

A new predictor was developed for identifying the nitrotyrosine sites in proteins based on a set of 19-tuple peptides generated as follows. Sliding a window of 19 amino acids along each of the 546 protein sequences taken from a protein database, collected were

iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition

| [Read Me](#) | [Data](#) | [Citation](#) |

Enter or copy/paste query protein sequences in **FASTA** format ([Example](#))

Upload input file in **FASTA** format ([Example](#))

Upload your input file: [Browse](#)

[Submit](#) [Clear](#)

[Contact @ Yan Xu](#)

[Close](#)

Figure 6. A semi-screenshot to show the top page of the iNitro-Tyr server. Its website address is at <http://app.aporc.org/iNitro-Tyr/>. doi:10.1371/journal.pone.0105018.g006

only those peptide segments with Y (tyrosine) at the center, i.e., the potential nitrotyrosine-site-containing peptides. The benchmark dataset thus obtained contains 1,044 experiment-confirmed nitrotyrosine peptides and 7,669 non-nitrotyrosine peptides.

The new predictor is called iNitro-Tyr, in which each of the potential nitrotyrosine-site-containing peptides was formulated with a 18-D vector formed by incorporating the position-specific dipeptide propensity (PSDP) into the general form [8] of pseudo amino acid composition [38,75] or Chou's PseAAC [39,51,54].

It has been observed by the rigorous cross validations that the iNitro-Tyr not only yields higher success rates but also is more stable and less noisy as reflected by a set of four metrics generally used to measure the quality of a predictor from different angles.

For the convenience of most experimental scientists, the web-server of iNitro-Tyr has been established at <http://app.aporc.org/iNitro-Tyr/>. Furthermore, to maximize their convenience, a step-by-step guide has been provided, by which users can easily get their desired results without the need to follow the complicated mathematics that were presented in this paper just for the integrity of the predictor.

It has not escaped our notice that the current approach can also be used to develop various effective methods for identifying the sites of other PTM sites in proteins.

References

1. Casoni F, Basso M, Massignan T, Gianazza E, Cheroni C, et al. (2005) Protein nitration in a mouse model of familial amyotrophic lateral sclerosis: possible multifunctional role in the pathogenesis. *J Biol Chem* 280: 16295–16304.
2. Ghesquiere B, Colaert N, Helsens K, Dejager L, Vanhaute C, et al. (2009) In vitro and in vivo protein-bound tyrosine nitration characterized by diagonal chromatography. *Mol Cell Proteomics* 8: 2642–2652.

Supporting Information

Supporting Information S1 The benchmark dataset used in this study contains 8,713 peptides formed by 19 amino acid residues with Y (tyrosine) at the center. Of these peptides, 1,044 are of nitrotyrosine and 7,669 of non-nitrotyrosine. Listed are also the codes of the source proteins from which these 19-tuple peptide sequences are derived as well as their corresponding sites in proteins. See the main text for further explanation. (DOC)

Acknowledgments

The authors wish to thank the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper.

Author Contributions

Conceived and designed the experiments: YX. Performed the experiments: YX LSW. Analyzed the data: KCC XW. Contributed reagents/materials/analysis tools: YX XW LSW. Wrote the paper: YX NYD KCC. Collected the data: XW LSW. Revised the manuscript: LYW NYD KCC.

3. Zhan X, Du Y, Crabb JS, Gu X, Kern TS, et al. (2008) Targets of tyrosine nitration in diabetic rat retina. *Mol Cell Proteomics* 7: 864–874.
4. Jarmula A, Rode W (2013) Computational study of the effects of protein tyrosine nitrations on the catalytic activity of human thymidylate synthase. *J Comput Aided Mol Des* 27: 45–66.
5. Abello N, Kerstjens HA, Postma DS, Bischoff R (2009) Protein tyrosine nitration: selectivity, physicochemical and biological consequences, denaturation, and proteomics methods for the identification of tyrosine-nitrated proteins. *J Proteome Res* 8: 3222–3238.
6. Feeney MB, Schoneich C (2013) Proteomic Approaches to Analyze Protein Tyrosine Nitration. *Antioxid Redox Signal*.
7. Dekker F, Abello N, Wisastra R, Bischoff R (2012) Enrichment and detection of tyrosine-nitrated proteins. *Curr Protoc Protein Sci Chapter 14: Unit 14.13*.
8. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *Journal of Theoretical Biology* 273: 236–247.
9. Chen W, Feng PM, Lin H (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research* 41: e69.
10. Min JL, Xiao X (2013) iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed Research International* 2013: 701317.
11. Xu Y, Shao XJ, Wu LY, Deng NY (2013) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* 1: e171.
12. Xiao X, Min JL, Wang P (2013) iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *Journal of Theoretical Biology* 337C: 71–79.
13. Fan YN, Xiao X, Min JL (2014) iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. *International Journal of Molecular Sciences* 15: 4915–4937.
14. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, et al. (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* 30: 1522–1529.
15. Liu B, Zhang D, Xu R, Xu J, Wang X, et al. (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30: 472–479.
16. Qiu WR, Xiao X (2014) iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int J Mol Sci* 15: 1746–1766.
17. Xu Y, Wen X, Shao XJ, Deng NY (2014) iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *International Journal of Molecular Sciences* 15: 7594–7610.
18. Ding H, Deng EZ, Yuan LF, Liu L, Lin H, et al. (2014) iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Research International* 2014: 286419.
19. Qiu WR, Xiao X, Lin WZ (2014) iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach. *BioMed Research International* 2014: ID 947416.
20. Chen W, Feng PM, Lin H (2014) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Research International* 2014: 623149.
21. Chen W, Feng PM, Deng EZ, Lin H (2014) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry* 10.1016/j.ab.2014.1006.1022.
22. Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *Journal of Biological Chemistry* 268: 16938–16948.
23. Chou KC (1996) Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Analytical Biochemistry* 233: 1–14.
24. Chou KC (1995) A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Science* 4: 1365–1383.
25. Chou KC (2001) Prediction of signal peptides using scaled window. *Peptides* 22: 1973–1979.
26. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry* 370: 1–16.
27. Liu Z, Cao J, Ma Q, Gao X, Ren J, et al. (2011) GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. *Mol Biosyst* 7: 1197–1204.
28. Xu Y, Ding J, Wu LY (2013) iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE* 8: e55844.
29. Chou JJ (1993) A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. *Biopolymers* 33: 1405–1414.
30. Wang M, Yang J, Xu ZJ (2005) SLLE for predicting membrane protein types. *Journal of Theoretical Biology* 232: 7–15.
31. Feng KY, Cai YD (2005) Boosting classifier for predicting protein domain structural class. *Biochemical & Biophysical Research Communications* 334: 213–217.
32. Feng PM, Chen W, Lin H (2013) iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Analytical Biochemistry* 442: 118–125.
33. Kandaswamy KK, Martinetz T, Moller S, Suganthan PN, et al. (2011) AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *Journal of Theoretical Biology* 270: 56–62.
34. Chou KC, Shen HB (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *Journal of Proteome Research* 5: 1888–1897.
35. Shen HB (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical Biochemistry* 394: 269–274.
36. Xiao X, Min JL, Wang P (2013) iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS ONE* 8: e72234.
37. Chou KC (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems* 9: 1092–1100.
38. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics* (Erratum: ibid, 2001, Vol44, 60) 43: 246–255.
39. Lin SX, Lapointe J (2013) Theoretical and experimental biology in one. *J Biomedical Science and Engineering (JBSE)* 6: 435–442.
40. Nanni L, Lumini A, Gupta D, Garg A (2012) Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information. *IEEE/ACM Trans Comput Biol Bioinform* 9: 467–475.
41. Hajisharifi Z, Piryaei M, Mohammad Beigi M, Behbahani M, Mohabatkar H (2014) Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *Journal of Theoretical Biology* 341: 34–40.
42. Mei S (2012) Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *Journal of Theoretical Biology* 310: 80–87.
43. Chen YK, Li KB (2013) 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. *Journal of Theoretical Biology* 318: 1–12.
44. Georgiou DN, Karakasis TE, Megaritis AC (2013) A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. *The Open Bioinformatics Journal* 7: 41–48.
45. Mohabatkar H, Mohammad Beigi M, Esmaili A (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *Journal of Theoretical Biology* 281: 18–23.
46. Khosravi M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H (2013) Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods. *Protein & Peptide Letters* 20: 180–186.
47. Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2013) Prediction of Allergic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. *Medicinal Chemistry* 9: 133–137.
48. Mohammad Beigi M, Behjati M, Mohabatkar H (2011) Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *Journal of Structural and Functional Genomics* 12: 191–197.
49. Zia Ur R, Khan A (2012) Identifying GPCRs and their Types with Chou's Pseudo Amino Acid Composition: An Approach from Multi-scale Energy Representation and Position Specific Scoring Matrix. *Protein & Peptide Letters* 19: 890–903.
50. Sun XY, Shi SP, Qiu JD, Suo SB, Huang SY, et al. (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Molecular BioSystems* 8: 3178–3184.
51. Du P, Gu S, Jiao Y (2014) PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *International Journal of Molecular Sciences* 15: 3495–3506.
52. Jiang Y, Huang T, Chen L, Gao YF, Cai Y, et al. (2013) Signal propagation in protein interaction network during colorectal cancer progression. *Biomed Res Int* 2013: 287019.
53. Du P, Wang X, Xu C, Gao Y (2012) PseAAC-BUILDER: A cross-platform standalone program for generating various special Chou's pseudo-amino acid compositions. *Analytical Biochemistry* 425: 117–119.
54. Cao DS, Xu QS, Liang YZ (2013) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29: 960–962.
55. Chou KC (2001) Using subsite coupling to predict signal peptides. *Protein Engineering* 14: 75–79.
56. Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. *Natural Science* 2: 63–92.
57. Chou KC (2001) Prediction of protein signal sequences and their cleavage sites. *PROTEINS: Structure, Function, and Genetics* 42: 136–139.
58. Chou KC, Wu ZC, Xiao X (2012) iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems* 8: 629–641.
59. Xiao X, Wu ZC (2011) iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of Theoretical Biology* 284: 42–51.
60. Wu ZC, Xiao X (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Molecular BioSystems* 7: 3287–3297.

61. Lin WZ, Fang JA, Xiao X (2013) iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems* 9: 634–644.
62. Chou KC, Wu ZC, Xiao X (2011) iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. *PLoS One* 6: e18258.
63. Xiao X, Wang P, Lin WZ, Jia JH (2013) iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry* 436: 168–177.
64. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 30: 275–349.
65. Chou KC, Shen HB (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols* 3: 153–162.
66. Shen HB, Yang J, Liu XJ (2005) Using supervised fuzzy clustering to predict protein structural classes. *Biochem Biophys Res Commun* 334: 577–581.
67. Mei S (2012) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *Journal of Theoretical Biology* 293: 121–130.
68. Chen W, Lin H, Feng PM, Ding C, Zuo YC, et al. (2012) iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE* 7: e47843.
69. Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Computational Biology and Chemistry* 34: 320–327.
70. Huang C, Yuan JQ (2013) Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. *Journal of Theoretical Biology* 335: 205–212.
71. Kong L, Zhang L, Lv J (2014) Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. *J Theor Biol* 344: 12–18.
72. Jia C, Lin X, Wang Z (2014) 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* 15: 10410–10423.
73. Liu W (1999) Protein secondary structural content prediction. *Protein Engineering* 12: 1041–1050.
74. Chou KC (1999) Using pair-coupled amino acid composition to predict protein secondary structure content. *Journal of Protein Chemistry* 18: 473–480.
75. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21: 10–19.
76. Chou KC (2002) Review: Prediction of protein signal sequences. *Current Protein and Peptide Science* 3: 615–622.
77. Chou KC, Shen HB (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Commun* 357: 633–640.