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

Yan Xu1*, Xin Wen1, Li-Shu Wen2, Ling-Yun Wu3, Nai-Yang Deng4, Kuo-Chen Chou5,6

1 Department of Information and Computer Science, University of Science and Technology Beijing, Beijing, China, 2 College of Sciences, Liaoning Shihou 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-NTyr 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...
of different numbers of amino acid residues, as formulated by benchmark dataset with different values of peptide set theory. Thus, the benchmark dataset for the current method if it is tested by the jackknife test or subsampling (K-fold) to separate a benchmark dataset into a training dataset and a categories:

\[ P_\xi(Y) = R_{\xi} \cdots R_{-1} \cdots R_{-5} \cdots R_{-\xi} \cdots R_{+1} \cdots R_{+5} \cdots R_{+\xi} \]  

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

\[ P_\xi(Y) \in \begin{cases} P_\xi^+(Y), & \text{if its center is a nitrotyrosine site} \\ P_\xi^-(Y), & \text{otherwise} \end{cases} \]  

where \( P_\xi^+(Y) \) represents a true nitrotyrosine peptide, \( P_\xi^-(Y) \) a false nitrotyrosine peptide, and \( \epsilon \) 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

\[ S_\xi = S_\xi^+ \cup S_\xi^- \]  

where \( S_\xi^+ \) only contains the samples of \( P_\xi^+(Y) \), i.e., the nitrotyrosine peptides; \( S_\xi^- \) only contains the samples of \( P_\xi^-(Y) \), 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 \( P_\xi(Y) \) is \( 2\xi+1 \) (Eq. 1), the benchmark dataset with different values of \( \xi \) will contain peptides of different numbers of amino acid residues, as formulated by

\[ S_\xi \text{ contains the peptides of } \begin{cases} 13 \text{ residues, when } \xi = 6 \\ 15 \text{ residues, when } \xi = 7 \\ 17 \text{ residues, when } \xi = 8 \\ 19 \text{ residues, when } \xi = 9 \\ 21 \text{ residues, when } \xi = 10 \\ \vdots \end{cases} \]  

The detailed procedures to construct \( S_\xi \) 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 \( \xi \), the lacking residue was filled with a dummy residue “X” [28]. (v) Those peptide samples thus obtained were put into the positive subset \( S_\xi^+ \) if their centers have been experimentally confirmed as the nitrotyrosine sites; otherwise, into the negative subset \( S_\xi^- \).

By following the aforementioned procedures, five such benchmark datasets \( (S_{\xi=6}, S_{\xi=7}, S_{\xi=8}, S_{\xi=9}, \text{ and } S_{\xi=10}) \) 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 UniProt 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 \( S_{\xi=9} \) as the benchmark dataset for further investigation. Thus, Eq. 3 can be reduced to

\[ S = S^+ \cup S^- \]  

where \( S = S_9 \), \( S^+ = S_9^+ \) containing 1,044 nitrotyrosine peptide samples, and \( S^- = S_9^- \) 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
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], while the 3rd one for those of Chou’s general PseAAC.

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

\[ \mathbf{P} = [\psi_1, \psi_2, \ldots, \psi_n, \ldots, \psi_\Omega]^T \]  

(6)

where \( \mathbf{T} \) is the transpose operator, while \( \Omega \) an integer to reflect the vector’s dimension. The value of \( \Omega \) as well as the components \( \psi_\nu (\nu = 1, 2, \ldots, \Omega) \) 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.

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

\[ \mathbf{P} = [\psi_1, \psi_2, \ldots, \psi_n, \ldots, \psi_\Omega]^T \]  

(6)

where \( \mathbf{T} \) is the transpose operator, while \( \Omega \) an integer to reflect the vector’s dimension. The value of \( \Omega \) as well as the components \( \psi_\nu (\nu = 1, 2, \ldots, \Omega) \) 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

\[ \mathbf{P}_s = \mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_s \mathbf{R}_{s+1} \cdots \mathbf{R}_{2s} \mathbf{R}_{2s+1} \]  

(7)

where \( \mathbf{R}_{\xi + 1} \), the residue at the center of the peptide, is tyrosine (Y), and all the other residues \( \mathbf{R}_\nu (\nu \neq \xi + 1) \) 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 dipeptide subsite positions on the sequence of \( \mathbf{P}_s \) will be \((2_{\xi + 1} + 1) = 2_{\xi + 1} \).

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

\[ \mathbf{Z}^+ (\xi) = \begin{bmatrix} z^+_{1,1} & z^+_{1,2} & \cdots & z^+_{1,2_{\xi + 1}} \\ z^+_{2,1} & z^+_{2,2} & \cdots & z^+_{2,2_{\xi + 1}} \\ \vdots & \vdots & \ddots & \vdots \\ z^+_{2_{\xi + 1},1} & z^+_{2_{\xi + 1},2} & \cdots & z^+_{2_{\xi + 1},2_{\xi + 1}} \end{bmatrix} \]  

(8a)

\[ \mathbf{Z}^- (\xi) = \begin{bmatrix} z^-_{1,1} & z^-_{1,2} & \cdots & z^-_{1,2_{\xi + 1}} \\ z^-_{2,1} & z^-_{2,2} & \cdots & z^-_{2,2_{\xi + 1}} \\ \vdots & \vdots & \ddots & \vdots \\ z^-_{2_{\xi + 1},1} & z^-_{2_{\xi + 1},2} & \cdots & z^-_{2_{\xi + 1},2_{\xi + 1}} \end{bmatrix} \]  

(8b)

where the element

\[ \begin{cases} z^+_{ij} = F^+ (\mathbf{D}_s^+ j) \\ z^-_{ij} = F^- (\mathbf{D}_s^- j) \end{cases} \quad (i = 1, 2, \ldots, 441; j = 1, 2, \ldots, 2_{\xi + 1}) \]  

(9)

and

\[ \mathbf{D}_1 = \mathbf{A}_1 \mathbf{A}_2 = \mathbf{A}_2 \mathbf{A}_3 = \mathbf{A}_3 \mathbf{A}_4 = \cdots = \mathbf{D}_{440} = \mathbf{X} \mathbf{D}_{441} = \mathbf{X} \mathbf{X} \]  

(10)

In Eq. 9, \( F^+ (\mathbf{D}_s^+ j) \) is the occurrence frequency of the \( i \)-th dipeptide \((i = 1, 2, \ldots, 441)\) at the \( j \)-th subsite on the sequence of \( \mathbf{D}_s \) (or the \( j \)-th column in the positive subset dataset \( \mathbf{S}_s^+ \)) 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 \( S^- \). Thus, for the peptide sequence of Eq. 7, its attribute to the positive set \( S^+ \) or negative set \( S^- \) can be formulated by a \( 2\xi\cdot D \) (dimension) vector \( P^+ \) or \( P^- \), as defined by [23]

\[
P_{\xi}^+ = \begin{bmatrix} \psi_1^+ & \psi_2^+ & \cdots & \psi_u^+ & \cdots & \psi_{2\xi}^+ \end{bmatrix}^T \tag{11a}
\]

\[
P_{\xi}^- = \begin{bmatrix} \psi_1^- & \psi_2^- & \cdots & \psi_u^- & \cdots & \psi_{2\xi}^- \end{bmatrix}^T \tag{11b}
\]

where

\[
\psi_u^+ = \begin{cases} 
\gamma_u^+ & \text{when } R_u, R_{u+1} = AA \\
\gamma_u^+ & \text{when } R_u, R_{u+1} = AC \\
\vdots & \\
\gamma_u^+ & \text{when } R_u, R_{u+1} = AX \\
\gamma_u^+ & \text{when } R_u, R_{u+1} = CA \\
\vdots & \\
\gamma_u^+ & \text{when } R_u, R_{u+1} = XX
\end{cases} \tag{12a}
\]

\[
\psi_u^- = \begin{cases} 
\gamma_u^- & \text{when } R_u, R_{u+1} = AA \\
\gamma_u^- & \text{when } R_u, R_{u+1} = AC \\
\vdots & \\
\gamma_u^- & \text{when } R_u, R_{u+1} = AX \\
\gamma_u^- & \text{when } R_u, R_{u+1} = CA \\
\vdots & \\
\gamma_u^- & \text{when } R_u, R_{u+1} = XX
\end{cases} \tag{12b}
\]

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 \( 2\xi\cdot D \) space, let us define an ideal nitrotyrosine peptide \( \Pi^+ \) [22] and an ideal non-nitrotyrosine peptide \( \Pi^- \) as expressed by

\[
\Pi^+_\xi = \begin{bmatrix} \lambda_1^+ & \lambda_2^+ & \cdots & \lambda_u^+ & \cdots & \lambda_{2\xi}^+ \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix} \tag{13}
\]

\[
\Pi^-_\xi = \begin{bmatrix} \lambda_1^- & \lambda_2^- & \cdots & \lambda_u^- & \cdots & \lambda_{2\xi}^- \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}
\]

where \( \lambda_i^+ (i=1,2,\cdots,2\xi) \) is the upper limit of the corresponding matrix element in Eq. 12a, and \( \lambda_i^- (i=1,2,\cdots,2\xi) \) 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 \( P_{\xi}^+ \) with \( \Pi^+ \) and that of \( P_{\xi}^- \) with \( \Pi^- \) can be defined as

\[
P_{\xi}^+ \cdot \Pi^+_\xi = \sum_{u=1}^{2\xi} \psi_u^+ \tag{14}
\]

\[
P_{\xi}^- \cdot \Pi^-_\xi = \sum_{u=1}^{2\xi} \psi_u^-
\]

Similar to the treatment in [23], let us define a discriminant function \( \Delta \) given by

\[
\Delta_\xi = \left( P_{\xi}^+ \cdot \Pi^+_\xi - P_{\xi}^- \cdot \Pi^-_\xi \right) - \mathcal{R} = \sum_{u=1}^{2\xi} \left( \psi_u^+ - \psi_u^- \right) - \mathcal{R} \tag{15}
\]

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 \( P_{\xi} \) of Eq. 7 can be identified according to the following rule

\[
\begin{cases} 
P_{\xi} \text{ belongs to nitrotyrosine peptide,} & \text{if } \Delta_\xi > 0 \\
P_{\xi} \text{ belongs to non-nitrotyrosine peptide,} & \text{if } \Delta_\xi \leq 0
\end{cases} \tag{16}
\]

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 \((2\xi+1)\)-tuble 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
According to the formulation proposed recently in [9,11,28] based on the symbols formulated in [25,55] introducing by Chou [25,55] in predicting signal peptides. The same four metrics can be expressed as

\[
\begin{align*}
\text{Sn} &= \frac{TP}{TP+FN} \\
\text{Sp} &= \frac{TN}{TN+FP} \\
\text{Acc} &= \frac{TP+TN}{TP+TN+FP+FN} \\
\text{MCC} &= \frac{(TP \times TN) - (FP \times FN)}{(TP+FN)(TN+FP)(TN+FN)}
\end{align*}
\]

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 intuitive and easier-to-understand, particularly the equation for MCC. Here let us adapt 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

\[
\begin{align*}
\text{Sn} &= 1 - \frac{N_-}{N^+}, \quad 0 \leq Sn \leq 1 \\
\text{Sp} &= 1 - \frac{N_+}{N^-}, \quad 0 \leq Sp \leq 1 \\
\text{Acc} &= 1 - \frac{N_+ + N_-}{N^+ + N^-}, \quad 0 \leq Acc \leq 1 \\
\text{MCC} &= \frac{1 - \left( \frac{N_+ + N_-}{N^+ + N^-} \right)}{\sqrt{\left( 1 + \frac{N_+ + N_-}{N^+ + N^-} \right) \left( 1 + \frac{N_+ + N_-}{N^+ + N^-} \right)}}, \quad -1 \leq MCC \leq 1
\end{align*}
\]

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 \(S_n = 1\). When \(N_- = N^+\) meaning that all the nitrotyrosine peptides were incorrectly predicted as the non-nitrotyrosine peptides, we have the sensitivity \(S_n = 0\). Likewise, when \(N_- = 0\) meaning none of the non-nitrotyrosine peptides was incorrectly predicted to be the nitrotyrosine peptide, we have the specificity \(S_p = 1\); whereas \(N_+ = N^-\) meaning all the non-nitrotyrosine peptides were incorrectly predicted as the nitrotyrosine peptides, we have the specificity \(S_p = 0\). When \(N_+ = N^- = 0\) meaning that none of nitrotyrosine peptides in the positive dataset \(S^+\) and none of the non-nitrotyrosine peptides in the negative dataset \(S^-\) was incorrectly predicted, we have the overall accuracy \(Acc = 1\) and \(MCC = 1\); when \(N_- = N^+\) and \(N_+ = N^-\) meaning that all the nitrotyrosine peptides in the positive dataset \(S^+\) and all the non-nitrotyrosine peptides in the negative dataset \(S^-\) were incorrectly predicted, we have the overall accuracy \(Acc = 0\) and \(MCC = -1\); whereas when \(N_+ = N^- + 2\) and \(N_+ = N^- - 2\) we have \(Acc = 0.5\) and \(MCC = 0\) 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
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 an 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 an 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]).
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-YNO2 [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 Acc = 84.52%, which is higher than the overall accuracy by GPS-YNO2 regardless what threshold is used for the latter. (ii) The Mathew’s correlation coefficient obtained by iNitro-Tyr is MCC = 0.4905, which is significantly higher than that by GPS-YNO2, indicating that the new predictor is more stable and less noisy. (iii) The sensitivity and specificity obtained by iNitro-Tyr are Sn = 81.76% and Sp = 85.89%, which are much more evenly distributed than those by the GPS-YNO2 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 Acc = 44.88% and 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 positions 7, 12, 17, and 47 belong to the nitrotyrosine site, while all the others are of non-nitrotyrosine site. (ii) The 2nd protein (P09029) 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.

4. Conclusions

As one of the important post translational 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...
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.

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.

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.

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, Helens 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.
29. Chou JJ (1993) A formulation for correlating properties of peptides and its
28. Xu Y, Ding J, Wu LY (2013) iSNO-PseAAC: Predict cysteine S-nitrosylation
22. Chou KC (1993) A vectorized sequence-coupling model for predicting HIV
21. Chen W, Feng PM, Lin H (2013) iSpot-PseDNC: identify recombination spots
20. Chen W, Feng PM, Lin H (2014) iTIS-PseTNC: a sequence-based
18. Ding H, Deng EZ, Yuan LF, Liu L, Lin H, et al. (2014) iCTX-Type: A
16. Qiu WR, Xiao X (2014) iRSpot-TNCPseAAC: Identify recombination spots
15. Liu B, Zhang D, Xu R, Xu J, Wang X, et al. (2014) Combining evolutionary
13. Fan YN, Xiao X, Min JL (2014) iNR-Drug: Predicting the interaction of drugs
12. Xiao X, Min JL, Wang P (2013) iCDI-PseFpt: Identify the channel-drug
6. Chou KC (2011) Some remarks on protein attribute prediction and pseudo
5. Abello N, Kerstjens HA, Postma DS, Bischoff R (2009) Protein tyrosine
4. Jarmula A, Rode W (2013) Computational study of the effects of protein tyrosine
213–217.
34. Chou KC, Shen HB (2006) Predicting eukaryotic protein subcellular location by
33. Chou KC (2001) Some Remarks on Predicting Multi-Label Attributes in
32. Liu B, Zhang D, Xu R, Xu J, Wang X, et al. (2014) Combining evolutionary
31. Wu ZC, Xiao X (2011) iLoc-Plant: a multi-label classifier for predicting the
30. Sun XY, Shi SP, Qiu JD, Suo SB, Huang SY, et al. (2012) Identifying protein
29. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino
28. Ding H, Deng EZ, Yuan LF, Liu L, Lin H, et al. (2014) Nuc-PseKNC: a sequence-based predictor for predicting antifreeze proteins from structural class. Biochemical & Biophysical Research Communications 334: 213–217.
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.
26. Xu Y, Ding J, Wu LY (2013) iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating dipole-dipole position-specific propensity into pseudo amino acid composition. International Journal of Molecular Sciences 15: 7594–7610.
25. 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.
24. Qiu WR, Xiao X, Lin WZ (2014) Methy-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach. BioMed Research International 2014: ID 947416.
23. Chen W, Feng PM, Lin H (2014) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Research International 2014: 621149.
22. Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry 268: 16938–16946.
21. Chou KC (1996) Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical Biochemistry 233: 1–14.
20. Chou KC (1995) A sequence-coupled-vector projection-model for predicting the specificity of GalNAc-transferase. Protein Science 4: 1363–1383.
19. Chou KC (2001) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979.
18. Chou KC, Shen HB (2007) Review: Recent progress in protein subcellular location prediction. Analytical Biochemistry 370: 1–16.
17. 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.
16. 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: e55664.
15. Chou JJ (1993) A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. Biochemical and Biophysical Research Communications 180: 137–146.
14. Wang M, Yang J, Xu ZJ (2005) SLE for predicting membrane protein types.
13. Feng KY, Cai YD (2005) Boosting classifier for predicting protein domain structural class. Biochemical and Biophysical Research Communications 334: 213–217.
12. Feng PM, Chen W, Lin H (2013) iHSP-PseRAA: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical Biochemistry 428: 42–46.
11. Kanadasamy KK, Martinez T, Moller S, Suganathan PN, et al. (2011) AFF-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology 270: 56–62.
10. Min JL, Xiao X (2013) iEye-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. BioMed Research International 2013: 703137.
9. Xu Y, Shao XJ, Wu LY, Deng NY (2013) iSONO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. Perif 1: e171.
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.
7. Chen W, Feng PM, Lin H (2013) iSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition Nucleic Acids Research 41: e69.
6. Feeney MB, Schoneich C (2013) Protemic Approaches to Analyze Protein Tyrosine Nitration. Antioxid Redox Signal.
5. Abello N, Kerstjens HA, Postma DS, Bischoff R (2009) Protein tyrosine nitration: selectivity, physicochemical and biological consequences, denitration, and proteomics methods for the identification of tyrosine-nitrated proteins. J Proteome Res 8: 3222–3236.
4. Jarmula A, Rode W (2013) Computational study of the effects of protein tyrosine nitration on the catalytic activity of human tyrosidase synthase. J Comput Chem 34: 459–466.
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, Liu 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: e47043.

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 (2015) 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 N-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 (2002) Review: Prediction of protein subfamily classes. Current Protein and Peptide Science 3: 615–622.

76. Chou KC, Shen HB (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357: 633–640.