OPEN ACCESS Freely available online

PLOS ONE

iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition

Yan Xu[1,3]*, Jun Ding[1], Ling-Yun Wu[2], Kuo-Chen Chou[3]*

1 Department of Information and Computer Science, University of Science and Technology Beijing, Beijing, China, **2** Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, **3** Gordon Life Science Institute, San Diego, California, United States of America

Abstract

Posttranslational modifications (PTMs) of proteins are responsible for sensing and transducing signals to regulate various cellular functions and signaling events. S-nitrosylation (SNO) is one of the most important and universal PTMs. With the avalanche of protein sequences generated in the post-genomic age, it is highly desired to develop computational methods for timely identifying the exact SNO sites in proteins because this kind of information is very useful for both basic research and drug development. Here, a new predictor, called iSNO-PseAAC, was developed for identifying the SNO sites in proteins by incorporating the position-specific amino acid propensity (PSAAP) into the general form of pseudo amino acid composition (PseAAC). The predictor was implemented using the conditional random field (CRF) algorithm. As a demonstration, a benchmark dataset was constructed that contains 731 SNO sites and 810 non-SNO sites. To reduce the homology bias, none of these sites were derived from the proteins that had ≥ 40 pairwise sequence identity to any other. It was observed that the overall cross-validation success rate achieved by iSNO-PseAAC in identifying nitrosylated proteins on an independent dataset was over 90%, indicating that the new predictor is quite promising. Furthermore, a user-friendly web-server for iSNO-PseAAC was established at http://app.aporc.org/iSNO-PseAAC/, by which users can easily obtain the desired results without the need to follow the mathematical equations involved during the process of developing the prediction method. It is anticipated that iSNO-PseAAC may become a useful high throughput tool for identifying the SNO sites, or at the very least play a complementary role to the existing methods in this area.

Citation: Xu Y, Ding J, Wu L-Y, Chou K-C (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(2): e55844. doi:10.1371/journal.pone.0055844

Editor: Eugene A. Permyakov, Russian Academy of Sciences, Institute for Biological Instrumentation, Russian Federation

Received December 4, 2012; **Accepted** January 2, 2013; **Published** February 7, 2013

Funding: This work is partially supported by the National Natural Science Foundation of China (No. 11101029, No. 10971223, No. 60970091, No. 11131009, No. 11071013) and the Fundamental Research Funds for the Central Universities, NCET of China (No. NCET-11-0574), and Knowledge Innovation Program of the Chinese Academy of Sciences (No. kjcx-yw-s7). 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.

* E-mail: yxu@gordonlifescience.org (YX); kcchou@gordonlifescience.org (K-CC)

Introduction

The post-translational modifications (PTMs) play a key role in providing proteins with structural and functional diversity, as well as in regulating cellular plasticity and dynamics. As illustrated in **Fig. 1**, the PTMs are covalent processing events that change the properties of a protein by proteolytic cleavage for adding a modifying group to one or more amino acids [1]. One of the most important and universal PTMs is S-nitrosylation (SNO). Recent reports have indicated that SNO can modulate protein stability and activities [2,3], as well as play an important role in a variety of biological processes, including cell signaling, transcriptional regulation, apoptosis, and chromatin remodeling [4].

Meanwhile, increasing evidences have indicated that SNO also plays an important role in various major diseases [5], such as cancer [6], Parkinson's [7,8], Alzheimer's [9], and Amyotrophic Lateral Sclerosis (ALS) [10].

Therefore, identifying the SNO sites in proteins is very important to both basic science and drug development.

Many experimental methods have been developed for identifying SNO sites, such as BST (biotin switch assay) [11], SNOSID [2,12], and SNO-RAC [13]. These methods have indeed provided very useful information in this area. Unfortunately, as pointed out by Seth and Stamler [14], experimental identification of SNO sites with a site-directed mutagenesis strategy is laborious and low-throughput due to the labile nature and the low-abundance of SNO. Particularly, with the avalanche of protein sequences generated in the postgenomic age, it is highly desired to develop computational method for timely and reliably identifying the SNO sites in proteins.

Actually, some computational methods have been proposed in this regard. For instance, based on a benchmark dataset consisting of 65 positive and 65 negative samples, Gross and co-workers [15] developed a computational method called SNOSID for identifying the SNO sites in proteins. A few years later, based on 549 experimentally verified SNO sites in 363 proteins, Xue et al [16] proposed a different method called GRS-SNO for the same purpose. Shortly afterwards, Li et al. [17] tried to improve the prediction performance by introducing the SVM (support vector machine) algorithm. Recently, Li et al. [18] proposed a predictor by means of the nearest neighbor algorithm (NNA) with the maximum relevance minimum redundancy (mRMR) approach.

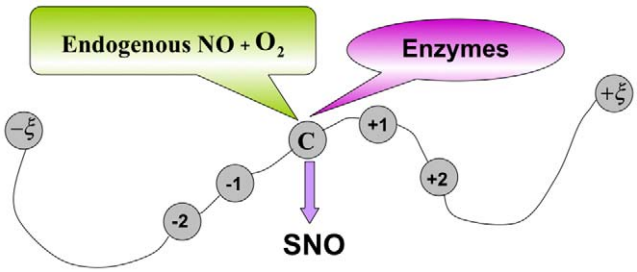

Figure 1. A schematic illustration to show the S-nitrosylation (SNO) site of a protein segment. The protein segment contains $2\xi+1$ residues, where C (cysteine) is located at the center of the peptide and all the other amino acids are depicted as an open circle with a number to indicate their sequential positions, respectively.
doi:10.1371/journal.pone.0055844.g001

Each of the aforementioned methods has its own merit and did play a role in stimulating the development of this area although bearing various limits. For example, no web-server has been provided for the most recent method [18], and hence its usage is quite limited, particularly for the majority of experimental scientists.

The present study was initiated in an attempt to develop a new and more powerful method to identify the SNO sites in proteins in hopes that it may become a useful tool for both basic research and drug development in the relevant areas.

As summarized in [19] and demonstrated in a series of recent publications (see, e.g., [20,21,22,23]), to establish a really useful statistical predictor for a protein or DNA system based on the sequence information, we usually need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein or DNA sequence samples with a feature vector that can truly reflect the 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 prediction accuracy; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these procedures one by one.

Materials and Methods

1. Benchmark Dataset

The benchmark dataset used in this study was derived from the dbSNO (http://dbsno.mbc.nctu.edu.tw/), a database that integrates the experimentally verified cysteine SNO sites in 1,757 proteins from different species [24]. To reduce the redundancy and avoid homology bias, we randomly picked 438 proteins in which none had $\geq 40\%$ pairwise sequence identity to any other. Based on these proteins and the annotations in the dbSNO database, a total of 731 experimentally verified SNO sites were collected. Meanwhile, to construct a corresponding negative dataset, a total of 810 experimentally verified non-SNO sites were randomly collected from the 438 proteins as well. The corresponding peptide fragments for the 731 SNO sites and 810 non-SNO sites were derived from UniProt database (release 2012_08), as can be generally formulated by

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

where the subscript ξ is an integer, $\mathbf{R}_{-\xi}$ is the ξ-th downstream amino acid residue from cysteine (C), $\mathbf{R}_{\xi}$ the ξ-th upstream amino acid residue, and so forth. Hereafter let us call a peptide as SNO or non-SNO peptide if its center is a SNO or non-SNO site, respectively. In the current study, we choose $\xi=10$. If the upstream or downstream in a protein was less than 10, the lacking residues were filled with the dummy code X. Thus, the benchmark dataset $\mathbb{S}$ can be formulated as

$$\mathbb{S}=\mathbb{S}^{+}+\mathbb{S}^{-} \quad (2)$$

where the positive dataset $\mathbb{S}^{+}$ contains $N^{+}=731$ SNO peptide fragments, while the negative dataset $\mathbb{S}^{-}$ contains $N^{-}=810$ non-SNO peptide fragments (cf. **Eq.1**), respectively. For reader's convenience, their sequences as well as the corresponding sites and protein codes are given in Supporting Information S1.

2. Sample Formulation or Feature Vector

To develop a sequence-based predictor for identifying the attribute of a protein or peptide, one of the keys is to formulate its sequence with an effective mathematical expression that can truly reflect the intrinsic correlation with the attribute to be predicted [25]. The most straightforward method to formulate the sample of a protein or peptide is to use its entire amino acid sequence. To identify its attribute, the tools for computing amino acid sequence similarity, such as BLAST [26,27], were utilized to search the database for those targets that have high sequence similarity to the query protein or peptide. Subsequently, the attribute annotations of the target proteins or peptides thus found were used to infer the attribute for the query protein or peptide. Unfortunately, this kind of straightforward sequential model, although containing the entire sequence information, failed to work when the query protein or peptide did not have any significant sequence similarity to the attribute-known proteins or peptides.

To avoid the above difficulty, which is inherent to the sequential model, various non-sequential or discrete models to formulate protein or peptide samples were proposed in hopes to enhance the prediction power.

Among the discrete models, the simplest one is the amino acid (AA) composition or AAC [28]. However, if using AAC to represent a peptide sample, its sequence-order or position-specific information would be totally lost, and hence might considerably limit the prediction quality.

To avoid completely losing the sequence-order information, the pseudo amino acid composition (PseAAC) was proposed to represent the sample of a protein or peptide [29,30]. The idea of PseAAC has been widely used in bioinformatics, proteomics, and system biology [25], such as predicting protein structural class [31], predicting metalloproteinase family [32], predicting protein subcellular localization [33], predicting DNA-binding proteins [21], identifying allergenic proteins [34], identify recombination spots [35], identifying bacterial virulent proteins [36], predicting protein folding rate [37], predicting GABA(A) receptor proteins [38], predicting protein supersecondary structure [39], predicting cyclin proteins [40], classifying amino acids [41], predicting enzyme family class [42], identifying risk type of human papillomaviruses [43], identifying protein quaternary structural attributes [44], identifying GPCRs and their types [45], and discriminating outer membrane proteins [46], among many others (see a long list of references cited in [19]). Because of its wide and increasing usage, in 2012 a powerful software called "PseAAC-Builder" (http://www.pseb.sf.net) [47] was established for generating various special modes of PseAAC for protein or peptide sequences.

According to a recent review [19], the general form of PseAAC for a protein or peptide **P** is formulated by

$$\mathbf{P}=[\psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_\Omega]^{\mathrm{T}} \tag{3}$$

where the subscript Ω is an integer, and its value as well as the components ψ_1, ψ_2, … will depend on how to extract the desired information from the amino acid sequence of **P** (cf. **Eq.1**). Below, let us describe how to extract useful information from the benchmark dataset $\mathbb{S}$ to define the peptide samples concerned via **Eq.3**.

It is obvious from **Eq.1** that when $\xi=10$, the corresponding peptide contains $(2\xi+1)=21$ amino acid residues. Since the residue at the center of the sequence is always C, we can omit it. Thus, for the convenience of formulation, **Eq.1** can be reduced to

$$\mathbf{P}=\mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_9\mathrm{R}_{10}\mathrm{R}_{11}\mathrm{R}_{12}\cdots\mathrm{R}_{19}\mathrm{R}_{20} \tag{4}$$

Also, as mentioned above, besides the 20 native amino acids, the sequence may also contain a dummy amino acid X. Here, 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.

Thus, we can introduce the following 21×20 matrix, the so-called "Position Specific Amino Acid Propensity" (PSAAP) matrix [48], to define the components of **Eq.3**

$$\mathbb{Z}=\begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,20} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ z_{20,1} & z_{20,2} & \cdots & z_{20,20} \\ z_{20+1,1} & z_{20+1,2} & \cdots & z_{20+1,20} \end{bmatrix} \tag{5}$$

where the element

$$z_{i,j}=F^{+}(\mathrm{R}_i|j)-F^{-}(\mathrm{R}_i|j) \quad (i=1,2,\cdots,21; j=1,2,\cdots,20) \tag{6}$$

where $F^{+}(\mathrm{R}_i|j)$ is the occurrence frequency of the *i*-th amino acid $(i=1,2,\cdots,21)$ in the*j*-th column in the positive benchmark dataset $\mathbb{S}^{+}$ that can be easily derived using the method described in [49] from the sequences in the Supporting Information S1, while $F^{-}(\mathrm{R}_i|j)$ is the corresponding occurrence frequency but derived from the negative benchmark dataset $\mathbb{S}^{-}$.

Thus, the components in **Eq.3** can be uniquely defined by

$$\psi_u=\begin{cases} z_{1,u} & \text{when } \mathrm{R}_i=\mathrm{A} \\ z_{2,u} & \text{when } \mathrm{R}_i=\mathrm{C} \\ \vdots & \vdots \quad [u=1,2,\cdots,\Omega\,(=20)] \\ z_{20,u} & \text{when } \mathrm{R}_i=\mathrm{Y} \\ z_{21,u} & \text{when } \mathrm{R}_i=\mathrm{X} \end{cases} \tag{7}$$

Since the components of the feature vector in **Eq.3** are now derived from the benchmark dataset $\mathbb{S}=\mathbb{S}^{+}+\mathbb{S}^{-}$, its correlation with SNO sites and non-SNO sites are self-evident.

3. Operation Engine

In this study, the "Conditional Random Field" (CRF) algorithm [50] was adopted to operate the prediction. It is a discriminative probabilistic model that inherits the advantages of "Maximum Entropy Markov Models" (MEMMs), often used for labeling and segmenting sequence data. The CRF operation engine has been quite successfully utilized in various areas of bioinformatics and computational proteomics, such as gene prediction [51], SNP array analysis [52], and protein structure [53].

In this study, the CRF software was downloaded from the website at http://www.di.ens.fr/~mschmidt/Software/crfChain.html. When used in the current study, the input of CRF is the query peptide fragment **P** as formulated by the feature vector of **Eq.3** as well as Eqs. 5–7, and the output is **Q(P)**, thus the query peptide is identified as

$$\mathbf{P}\in\begin{cases} \text{SNO peptide, if } \mathrm{Q}(\mathbf{P})>\Theta \\ \text{non-SNO peptide, otherwise} \end{cases} \tag{8}$$

where $\Theta=0.58$ is a threshold obtained by optimizing the overall success rate for the peptides in the benchmark dataset $\mathbb{S}$ as done in [54].

The predictor thus established via the above procedures is called **iSNO-PseAAC**, which can be used to identify the nitrosylated proteins and their SNO sites. To provide an intuitive picture, a flowchart is provided in **Fig. 2** to illustrate the prediction process of **iSNO-PseAAC**.

Results and Discussion

1. Four Different Metrics for Measuring the Prediction Quality

One of the important procedures in developing a useful statistical predictor [19] is to objectively evaluate its performance or anticipated success rate. To provide a more intuitive and easier-to-understand method to measure the prediction quality, here the criteria proposed in [55] was adopted. According to those criteria, the rates of correct predictions for the SNO peptides in dataset $\mathbb{S}^{+}$ and the non-SNO peptides in dataset $\mathbb{S}^{-}$ are respectively defined by

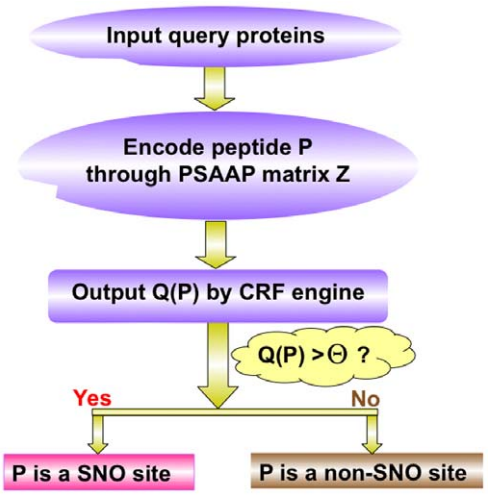

Figure 2. A flowchart to show the prediction process of iSNO-PseAAC.
doi:10.1371/journal.pone.0055844.g002

$$\begin{cases} \Lambda^+ = \dfrac{N^+ - N_-^+}{N^+}, & \text{for the SNO-peptides} \\ \Lambda^- = \dfrac{N^- - N_-^-}{N^-}, & \text{for the non-SNO peptides} \end{cases} \tag{9}$$

where N^+ is the total number of the SNO peptides investigated while N_-^+ the number of the SNO peptides incorrectly predicted as the non-SNO peptides; N^- the total number of the non-SNO peptides investigated while N_+^- the number of the non-SNO peptides incorrectly predicted as the SNO peptides. The overall success prediction rate is given by [56]

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{10}$$

It is obvious from **Eqs. 9–10** that, if and only if none of the SNO peptides and the non-SNO peptides are mispredicted, i.e., $N_-^+ = N_+^- = 0$ and $\Lambda^+ = \Lambda^- = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be smaller than 1.

On the other hand, it is instructive to point out that the following equation is often used in literatures for examining the performance quality of a predictor

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{MCC} = \dfrac{(\mathrm{TP}\times\mathrm{TN}) - (\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{cases} \tag{11}$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient.

The relations between the symbols in **Eq.10** and those in **Eq.11** are given by

$$\begin{cases} \mathrm{TP} = N^+ - N_-^+ \\ \mathrm{TN} = N^- - N_+^- \\ \mathrm{FP} = N_+^- \\ \mathrm{FN} = N_-^+ \end{cases} \tag{12}$$

Substituting **Eq.12** into **Eq.11** and also considering **Eq.10**, we obtain

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ \mathrm{MCC} = \dfrac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N^\pm}{N^+}\right)\left(1 + \frac{N^\pm - N_+^-}{N^-}\right)}} \end{cases} \tag{13}$$

From the above equation, we can see: when $N_-^+ = 0$ meaning none of the SNO peptides was mispredicted to be a non-SNO peptide, we have the sensitivity $\mathrm{Sn} = 1$; while $N_-^+ = N^+$ meaning that all the SNO peptides were mispredicted to be the non-SNO peptides, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning none of the non-SNO peptides was mispredicted, we have the specificity $\mathrm{Sp} = 1$; while $N_+^- = N^-$ meaning all the non-SNO peptides were incorrectly predicted as the SNO peptides, we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of SNO peptides in the dataset $\mathbb{S}^+$ and none of the non-SNO peptides in $\mathbb{S}^-$ was incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = \Lambda = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the SNO peptides in the dataset $\mathbb{S}^+$ and all the non-SNO peptides in $\mathbb{S}^-$ were mispredicted, we have the overall accuracy $\mathrm{Acc} = \Lambda = 0$. The MCC correlation coefficient is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$ meaning that none of the SNO peptides in the dataset $\mathbb{S}^+$ and none of non-SNO peptides in $\mathbb{S}^-$ was mispredicted, we have $\mathrm{Mcc} = 1$; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have $\mathrm{Mcc} = 0$ meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$ we have $\mathrm{MCC} = -1$ meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier-to-understand when using **Eq.13** to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient.

2. Cross-Validation to Evaluate Success Rates

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling (K-fold cross-validation) test, and jackknife test. However, as elaborated in [57] and demonstrated by Eqs.28–32 of [19], among the three cross-validation methods, the jackknife test is deemed the least arbitrary and most objective because it can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictor (see, e.g., [36,45,58,59,60,61,62]). However, to reduce computational time, here let us adopt the 10-fold cross-validation to examine the prediction quality as done by many investigators for PTM sites prediction [63,64,65,66]. The cross-validations were performed 50 times for different subsampling combinations, followed by averaging their outcomes.

The results thus obtained on the benchmark dataset $\mathbb{S}$ for the four metrics as defined in **Eq.13** are given in **Table 1**, where for facilitating comparison the corresponding results obtained by GPS-SNO [16] are also given. As can be seen from the table, the overall success, sensitivity and MCC rates achieved by **iSNO-PseAAC** are all significantly higher than those by the GPS-SNO predictor [16] regardless its threshold was set at "high", "medium", or "low". As for the method proposed in [17] and the method recently proposed in [18], the former web-server was not working, while the latter had no web-server at all, and hence no corresponding data can be given in **Table 1** for comparison.

3. Large-Scale Prediction in Identifying Nitrosylated Proteins

Listed in Supporting Information S2 are the predicted results by **iSNO-PseAAC** for a set of 461 independent nitrosylated proteins, none of which occurs in the 438 proteins used to train the current predictor. They were taken from Xue et al. [16] and known belonging to nitrosylated proteins as verified by experiments. As

Table 1. The performance comparison of iSNO-PseAAC with other existing prediction methods[a] in this area.

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC
iSNO-PseAAC	67.01	68.15	67.62	0.3515
GPS-SNO[b]	18.88	89.63	56.07	0.1210
GPS-SNO[c]	28.04	81.98	56.39	0.1193
GPS-SNO[d]	45.01	73.33	59.90	0.1915

[a]The method proposed in [18] has no web-server provided, and the web-server in [17] did not work. Therefore, the rates for the two methods are unavailable.
[b]The method proposed in [16] when the threshold parameter was set "high".
[c]The method proposed in [16] when the threshold parameter was set "medium".
[d]The method proposed in [16] when the threshold parameter was set "low".
doi:10.1371/journal.pone.0055844.t001

we can see from Supporting Information S3, of the 461 proteins, 416 were predicted containing at least one SNO sites meaning belonging nitrosylated proteins. The overall success rate was $416/461 = 90.24\%$.

4. Web-Server Guide

For the convenience of the vast majority of experimental scientists, a web-server for **iSNO-PseAAC** was established. Below, let us give a step-by-step guide on how to use the web-server to get the desired results without the need to follow the mathematic equations that were presented just for the integrity in developing the predictor.

Step 1. Open the web server at at http://app.aporc.org/iSNO-PseAAC/ and you will see the top page of the predictor on your computer screen, as show in **Fig. 3**. Click on the Read Me button to see a brief introduction about **iSNO-PseAAC** predictor and the caveat when using it.

Step 2. Either type or copy/paste the query protein sequences into the input box shown at the center of **Fig. 3**. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol (">") in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a ">" 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 input box.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the query protein sequences in the Example window as the input, after clicking the Submit button, you will see on your screen the predicted SNO site positions and the corresponding sequences segments as formulated by **Eq.1**. All these results are fully consistent with the experimentally verified results. It takes about a few seconds for the above computation before the predicted results appear on the computer screen; the more number of query proteins and longer of each sequence, the more time it is usually needed.

Step 4. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of **iSNO-PseAAC**.

Step 5. Click on the Data button to download the benchmark datasets used to train and test the **iSNO-PseAAC** predictor.

Caveat. To obtain the predicted result with the expected success rate, the entire sequence of the query protein rather than its fragment should be used as an input. A sequence with less than 50 amino acid residues is generally deemed as a fragment. Also, the size of your input for each submission should be less than 100K; if greater than 100K, please contact Yan Xu at xuyan@ustb.edu.cn.

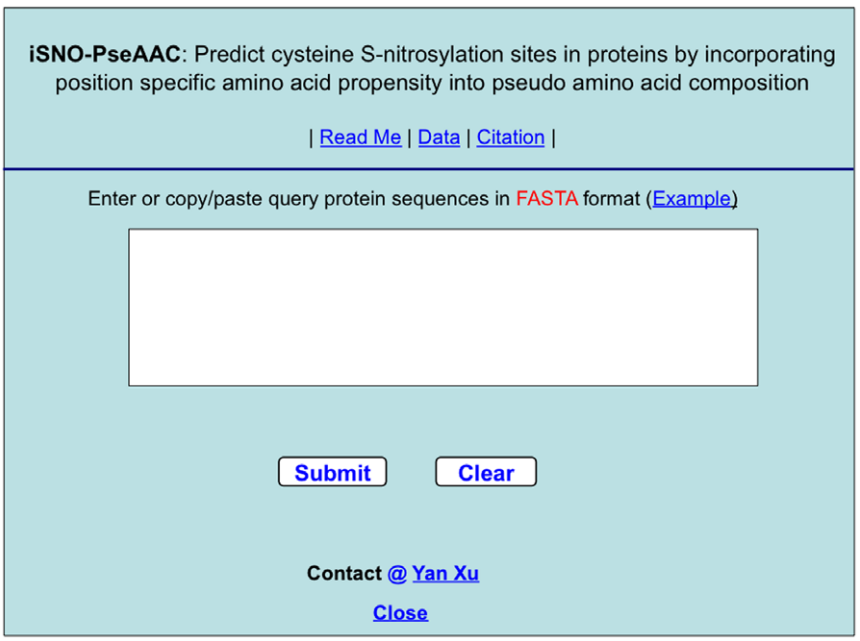

Figure 3. A semi-screenshot to show the top page of the iSNO-PseAAC web-server. Its website address is at http://app.aporc.org/iSNO-PseAAC/.
doi:10.1371/journal.pone.0055844.g003

Supporting Information

Supporting Information S1 The benchmark dataset $\mathbb{S}=\mathbb{S}^+\cup\mathbb{S}^-$, where the positive dataset $\mathbb{S}^+$ contains $N^+=731$ SNO sites while the negative dataset $\mathbb{S}^-$ contains $N^-=810$ non-SNO sites.
(PDF)

Supporting Information S2 Predicted results by iSNO-PseAAC on an independent dataset of 461 proteins, which have been verified by experiments as nitrosylated proteins but none of which occurs in the 438 proteins used to train the current predictor. The overall success rate was $416/461=92.24\%$.
(PDF)

Supporting Information S3 The detailed SNO sites detected by iSNO-PseAAC on an independent dataset with 461 nitrosylated proteins, of which 416 were predicted containing at least one SNO site.
(PDF)

Acknowledgments

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

Author Contributions

Conceived and designed the experiments: YX K-CC. Performed the experiments: YX JD L-YW. Analyzed the data: YX K-CC. Contributed reagents/materials/analysis tools: JD. Wrote the paper: YX K-CC.

References

1. Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21: 255–261.
2. Derakhshan B, Wille PC, Gross SS (2007) Unbiased identification of cysteine S-nitrosylation sites on proteins. Nat Protoc 2: 1685–1691.
3. Tsang AH, Lee YI, Ko HS, Savitt JM, Pletnikova O, et al. (2009) S-nitrosylation of XIAP compromises neuronal survival in Parkinson's disease. Proc Natl Acad Sci U S A 106: 4900–4905.
4. Nott A, Watson PM, Robinson JD, Crepaldi L, Riccio A (2008) S-Nitrosylation of histone deacetylase 2 induces chromatin remodelling in neurons. Nature 455: 411–415.
5. Foster MW, Hess DT, Stamler JS (2009) Protein S-nitrosylation in health and disease: a current perspective. Trends Mol Med 15: 391–404.
6. Aranda E, Lopez-Pedrera C, De La Haba-Rodriguez JR, Rodriguez-Ariza A (2012) Nitric oxide and cancer: the emerging role of S-nitrosylation. Curr Mol Med 12: 50–67.
7. Yao D, Gu Z, Nakamura T, Shi ZQ, Ma Y, et al. (2004) Nitrosative stress linked to sporadic Parkinson's disease: S-nitrosylation of parkin regulates its E3 ubiquitin ligase activity. Proc Natl Acad Sci U S A 101: 10810–10814.
8. Uehara T, Nakamura T, Yao D, Shi ZQ, Gu Z, et al. (2006) S-nitrosylated protein-disulphide isomerase links protein misfolding to neurodegeneration. Nature 441: 513–517.
9. Cho DH, Nakamura T, Fang J, Cieplak P, Godzik A, et al. (2009) S-nitrosylation of Drp1 mediates beta-amyloid-related mitochondrial fission and neuronal injury. Science 324: 102–105.
10. Schonhoff CM, Matsuoka M, Tummala H, Johnson MA, Estevez AG, et al. (2006) S-nitrosothiol depletion in amyotrophic lateral sclerosis. Proc Natl Acad Sci U S A 103: 2404–2409.
11. Jaffrey SR, Erdjument-Bromage H, Ferris CD, Tempst P, Snyder SH (2001) Protein S-nitrosylation: a physiological signal for neuronal nitric oxide. Nat Cell Biol 3: 193–197.
12. Greco TM, Hodara R, Parastatidis I, Heijnen HF, Dennehy MK, et al. (2006) Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. Proc Natl Acad Sci U S A 103: 7420–7425.
13. Forrester MT, Thompson JW, Foster MW, Nogueira L, Moseley MA, et al. (2009) Proteomic analysis of S-nitrosylation and denitrosylation by resin-assisted capture. Nat Biotechnol 27: 557–559.
14. Seth D, Stamler JS (2011) The SNO-proteome: causation and classifications. Curr Opin Chem Biol 15: 129–136.
15. Hao G, Derakhshan B, Shi L, Campagne F, Gross SS (2006) SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures. Proc Natl Acad Sci U S A 103: 1012–1017.
16. Xue Y, Liu Z, Gao X, Jin C, Wen L, et al. (2010) GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. PLoS One 5: e11290.
17. Li YX, Shao YH, Jing L, Deng NY (2011) An efficient support vector machine approach for identifying protein S-nitrosylation sites. Protein Pept Lett 18: 573–587.
18. Li BQ, Hu LL, Niu S, Cai YD, Chou KC (2012) Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. Journal of Proteomics 75: 1654–1665.
19. 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.
20. Xiao X, Wang P, Chou KC (2012) iNR-PhysChem: A Sequence-Based Predictor for Identifying Nuclear Receptors and Their Subfamilies via Physical-Chemical Property Matrix. PLoS ONE 7: e30869.
21. Lin WZ, Fang JA, Xiao X, Chou KC (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE 6: e24756.
22. 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.
23. 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.
24. Chen YJ, Ku WC, Lin PY, Chou HC, Khoo KH (2010) S-alkylating labeling strategy for site-specific identification of the s-nitrosoproteome. J Proteome Res 9: 6417–6439.
25. Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6: 262–274.
26. Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. In: Suhai S, editor. Theoretical and Computational Methods in Genome Research. New York: Plenum. pp. 1–14.
27. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 17: 149–163.
28. Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99: 152–162.
29. 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.
30. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19.
31. 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.
32. 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.
33. Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34: 565–572.
34. Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2013) Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. Medicinal Chemistry 9: 133–137.
35. Chen W, Feng PM, Lin H, Chou KC (2012) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Research doi:101093/nar/gks1450.
36. 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.
37. Guo J, Rao N, Liu G, Yang Y, Wang G (2011) Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. Journal of Computational Chemistry 32: 1612–1617.
38. Mohabatkar H, Mohammad Beigi M, Esmaeili 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.
39. Zou D, He Z, He J, Xia Y (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. Journal of Computational Chemistry 32: 271–278.
40. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
41. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. Journal of Theoretical Biology 257: 17–26.

42. Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology 248: 546–551.
43. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263: 203–209.
44. 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.
45. 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.
46. Hayat M, Khan A (2012) Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. Protein & Peptide Letters 19: 411–421.
47. Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Analytical Biochemistry 425: 117–119.
48. Tang YR, Chen YZ, Canchaya CA, Zhang Z (2007) GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. Protein Eng Des Sel 20: 405–412.
49. Chou KC (2001) Using subsite coupling to predict signal peptides. Protein Engineering 14: 75–79.
50. Lafferty W, Andrew, M Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth ternational Conference on Machine Learning San Francisco, CA, ,USA: Morgan Kaufmann Publishers Inc. pp. 282–289.
51. DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, et al. (2007) Conrad: gene prediction using conditional random fields. Genome Res 17: 1389–1398.
52. Wu L, Shen Y, Liu X, Ma X, Xi B, et al. (2009) The 1425G/A SNP in PRKCH is associated with ischemic stroke and cerebral hemorrhage in a Chinese population. Stroke 40: 2973–2976.
53. Li F, Sonveaux P, Rabbani ZN, Liu S, Yan B, et al. (2007) Regulation of HIF-1alpha stability through S-nitrosylation. Mol Cell 26: 63–74.
54. Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry 268: 16938–16948.
55. Chou KC (2001) Prediction of protein signal sequences and their cleavage sites. PROTEINS: Structure, Function, and Genetics 42: 136–139.
56. Chou KC (2001) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979.
57. Chou KC, Shen HB (2010) Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms (doi:10.4236/ns.2010.210136). Natural Science 2: 1090–1103 (openly accessible at http://www.scirp.org/journal/NS/).
58. Hayat M, Khan A (2012) MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM. J ournal of Theoretical Biology 292: 93–102.
59. Jahandideh S, Srinivasasainagendra V, Zhi D (2012) Comprehensive comparative analysis and identification of RNA-binding protein domains: Multi-class classification and feature selection. J Theor Biol 312: 65–75.
60. Nanni L, Brahnam S, Lumini A (2012) Wavelet images and Chou's pseudo amino acid composition for protein classification. Amino Acids 43: 657–665.
61. Niu XH, Hu XH, Shi F, Xia JB (2012) Predicting Protein Solubility by the General Form of Chou's Pseudo Amino Acid Composition: Approached from Chaos Game Representation and Fractal Dimension. Protein & Peptide Letters 19: 940–948.
62. Lin WZ, Fang JA, Xiao X, Chou KC (2012) Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model. PLoS One 7: e49040.
63. Kim JH, Lee J, Oh B, Kimm K, Koh I (2004) Prediction of phosphorylation sites using SVMs. Bioinformatics 20: 3179–3184.
64. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, et al. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res 35: W588–594.
65. Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, et al. (2009) Incorporating support vector machine for identifying protein tyrosine sulfation sites. J Comput Chem 30: 2526–2537.
66. Shao JL, Xu D, Tsai S, Wang YF, Ngar S. (2009) Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction. PLoS One 4: e4920.