Int. J. Mol. Sci. **2014**, *15*, 10410-10423; doi:10.3390/ijms150610410

OPEN ACCESS

International Journal of

Molecular Sciences

ISSN 1422-0067

www.mdpi.com/journal/ijms

Article

Prediction of Protein *S*-Nitrosylation Sites Based on Adapted Normal Distribution Bi-Profile Bayes and Chou's Pseudo Amino Acid Composition

Cangzhi Jia *, Xin Lin and Zhiping Wang *

Department of Mathematics, Dalian Maritime University, Dalian 116026, China;
E-Mail: xilinmath@163.com

* Authors to whom correspondence should be addressed; E-Mails: cangzhijia@dlmu.edu.cn (C.J.); zpwangdlmu@163.com (Z.W.); Tel.: +86-139-4093-3510 (C.J.); Fax: +86-411-8470-7245 (C.J.).

Received: 14 February 2014; in revised form: 12 May 2014 / Accepted: 20 May 2014 / Published: 10 June 2014

Abstract: Protein *S*-nitrosylation is a reversible post-translational modification by covalent modification on the thiol group of cysteine residues by nitric oxide. Growing evidence shows that protein *S*-nitrosylation plays an important role in normal cellular function as well as in various pathophysiologic conditions. Because of the inherent chemical instability of the *S*-NO bond and the low abundance of endogenous *S*-nitrosylated proteins, the unambiguous identification of *S*-nitrosylation sites by commonly used proteomic approaches remains challenging. Therefore, computational prediction of *S*-nitrosylation sites has been considered as a powerful auxiliary tool. In this work, we mainly adopted an adapted normal distribution bi-profile Bayes (ANBPB) feature extraction model to characterize the distinction of position-specific amino acids in 784 *S*-nitrosylated and 1568 non-*S*-nitrosylated peptide sequences. We developed a support vector machine prediction model, iSNO-ANBPB, by incorporating ANBPB with the Chou's pseudo amino acid composition. In jackknife cross-validation experiments, iSNO-ANBPB yielded an accuracy of 65.39% and a Matthew's correlation coefficient (*MCC*) of 0.3014. When tested on an independent dataset, iSNO-ANBPB achieved an accuracy of 63.41% and a *MCC* of 0.2984, which are much higher than the values achieved by the existing predictors SNOSite, iSNO-PseAAC, the Li *et al*. algorithm, and iSNO-AAPair. On another training dataset, iSNO-ANBPB also outperformed GPS-SNO and iSNO-PseAAC in the 10-fold crossvalidation test.

Keywords: *S*-nitrosylation; post-translational modification; bi-profile Bayes; amino acid physicochemical properties

1. Introduction

Protein *S*-nitrosylation, the covalent attachment of a nitric oxide (NO) moiety to cysteine residues of proteins resulting in the formation of *S*-nitrosothiols (SNO), is a typical redox-dependent posttranslational modification that is associated with redox-based cellular signaling [1–3]. Protein *S*-nitrosylation has been reported to play roles in the *in vitro*/*in vivo* regulation of a variety of metabolic enzymes, oxidoreductases, proteases, protein kinases, and protein phosphatases, as well as in the function of regulatory factors (including G protein) [4,5]. Many studies have shown that *S*-nitrosylated proteins exhibit abnormal increases or decreases in a variety of diseases [6]. For example, protein *S*-nitrosylation products were significantly increased compared with normal levels in diabetes, tuberculosis and other diseases; while protein *S*-nitrosylation products were significantly decreased compared with normal levels in asthma, neonatal oxygen deficiency, emphysema, and other diseases. Therefore, the regulation of protein *S*-nitrosylation modification may be a new and effective way for health protection. In addition, deregulation of *S*-nitrosylation has been implicated in tumor initiation and progression [4,7]. The increasing prominence of the roles of *S*-nitrosylation in diseases indicates a need for improved analytical methods to identify and quantify *S*-nitrosylated proteins under various physiological and pathophysiological conditions for investigative studies and clinical diagnosis [1,6,7]. The use of traditional mass spectrometry-based proteomics has been challenging because of the inherent chemical instability of the *S*-NO bond [4,8]. Currently, the biotin switch technique (BST), which was designed to purify and detect *S*-nitrosylated proteins, has become a widely used method for studying protein *S*-nitrosylation [9]. However, some researchers have suggested that the ascorbic acid signal enhancement as necessary and sufficient conditions of BST has led to a high number of false positives. A further study has shown that BST cannot be used to determine *S*-nitrosylated sites when the proportion of *S*-nitrosylated sites is less than 1% [10]. Hence, the computational prediction of protein *S*-nitrosylation sites may provide useful and experimentally testable information about potential protein *S*-nitrosylation sites. In recent years, several computational approaches have been developed to predict protein *S*-nitrosylated sites.

Hao *et al.* [11] developed the earliest prediction tool for *S*-nitrosylation called SNOSID, which is a support vector machine (SVM) system trained on the limited 65 *S*-nitrosylation sites and 65 non-*S*-nitrosylation sites that were available at the time. Xue *et al.* [12] constructed the first online server GPS-SNO for *S*-nitrosylation site prediction based on the modified group-based prediction system (GPS) version 3.0 algorithm. Trained on a large dataset of 504 experimentally verified *S*-nitrosylation sites in 327 unique proteins, GPS-SNO achieved an accuracy of 75.80%, a sensitivity of 53.57%, and a specificity of 80.14% in the jackknife cross-validation test. However, the independent predictive performance of GPS-SNO was tested on 485 *S*-nitrosylated substrates that were not identified by experimental verification; suggesting that further validation of the predictive capability of GPS-SNO is needed. In 2011, Lee *et al.* [13] and Li *et al.* [14] used different approaches to try to improve the prediction of protein *S*-nitrosylation. Lee *et al.* [13] incorporated information about amino acid composition, accessible surface area, and physicochemical properties into the maximal dependence decomposition (MDD) algorithm to obtain conserved *S*-nitrosylation motifs. Then, by combining the MDD-clustered motifs with a SVM, they built the online server SNOSite for predicting *S*-nitrosylation sites, which achieved an accuracy of 67.5% and a Matthew's correlation coefficient (*MCC*) of 0.245. Li *et al.* [14] established the prediction model CPR-SNO, using a SVM to improve the prediction performance. Instead of a SVM, Li *et al.* [15] proposed a nearest neighbor algorithm model that incorporated maximum relevance minimum redundancy and incremental feature selection techniques; however, the prediction results were not very satisfactory. On a newly created training dataset and an independent testing dataset, the *MCC*s were only 0.1381 and 0.1886, respectively. Xu *et al.* [16] proposed a web server called iSNO-PseAAc, which incorporated position-specific amino acid propensity into pseudo amino acid composition. The iSNO-PseAAc predictor achieved a *MCC* of 0.3515, which is substantially higher than the best *MCC* of 0.1915 obtained by GPS-SNO. More recently, Xu *et al.* [17] developed a new predictor called iSNO-AAPair by taking into account the coupling effects for all the pairs formed by the nearest residues and the pairs formed by the next nearest residues along protein chains. Despite the many *S*-nitrosylation predictors that have been developed, the *MCC* prediction values that they achieve are relatively lower than the values achieved by predictors of other post-translational modifications. Therefore, the discovery of new features will help in the development of more effective tools for protein *S*-nitrosylation site identification.

Bi-profile feature extraction has been applied in the prediction of many types of protein post-translational modification and has provided significant improvements in prediction performance [18–25]. The theoretical basis of this approach is that positive and negative peptide sequences should exhibit different features or characteristics [18]. In this study, we propose a computational model iSNO-ANBPB based on an adapted normal distribution bi-profile Bayes (ANBPB) feature extraction model and Chou's pseudo amino acid compositions for protein *S*-nitrosylation site prediction. We performed jackknife and 10-fold cross-validation experiments on two recently constructed training datasets in [15,16] and tested iSNO-ANBPB on an independent dataset constructed in [15], to comprehensively compare iSNO-ANBPB with four recently developed competing predictors. Three kinds of comparative results consistently indicated that iSNO-ANBPB achieved higher *MCC*s and outperformed other current approaches.

According to a recent comprehensive review [26] and demonstrated by a series of recent publications (see, e.g., [27–30]), to establish a really useful statistical predictor for a protein 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 protein samples with an effective mathematical expression that can truly reflect their 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 of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we describe how to deal with these steps.

2. Results and Discussion

2.1. Results

2.1.1. Sequence Analysis of *S*-Nitrosylation Sites

To explore the distinction between *S*-nitrosylation peptide sequences and non-*S*-nitrosylation peptide sequences, we conducted sequence analysis on the Li training dataset [15]. We calculated the relative position-specific propensities of each amino acid at each position (r_{xj}) in the sequence to obtain the relative frequency of a particular amino acid in the *S*-nitrosylation dataset over the frequency of the same amino acid in the non-*S*-nitrosylation dataset. As shown in Table S1, several amino acids at specific positions revealed some distinctive r_{xj} scores. Amino acids H, K, and N were found to be relatively enriched in the *S*-nitrosylation peptides with average r_{xj} scores of 1.23, 1.25, and 1.13 respectively. On the other hand, amino acids C, F, and W were found to be relatively depleted in the *S*-nitrosylation peptides with average r_{xj} scores of 0.64, 0.86, and 0.74 respectively. However, the independent distinct r_{xj} scores are not sufficient for defining a sequence motif for *S*-nitrosylation sites and more complex patterns of position-specific residue propensities in peptide sequences should be exploited to further improve the computational performance of *S*-nitrosylation site predictors.

2.1.2. Performance of the BPB, BRABSB, ANBPB and RANS Prediction Models

The weight parameters (*W*1 and *W*-1) in a SVM were adapted to increase the precision of sensitivity. For each training process, the initial *W*1 values were set to 1, 1.5, 2, and 2.5, until the *MCC*s reached their maximum. Notably, the performances of all these models significantly improved after the optimization of the *W*1 parameter (Tables S2–S5).

To find the best prediction model to identify potential protein *S*-nitrosylation sites, bi-profile Bayes (BPB) [18], bi-relative adapted binomial score Bayes (BRABSB) [23], adapted normal distribution bi-profile Bayes (ANBPB) [24], and the relative adapted normal score (RANS) [24] feature extraction combined with Chou's pseudo amino acid composition were developed on the same Li training datasets. The performances of the BPB, BRABSB, ANBPB, and RANS models for predicting protein *S*-nitrosylation and non-*S*-nitrosylation sites were examined by jackknife tests. The weight parameter *W*1 was optimized separately for the BPB, BRABSB, ANBPB and RANS models and the detailed results are available in Tables S1–S4. The best results obtained by each model are listed in Table 1. The BPB and ANBPB models reached their highest *MCC* values of 0.2933 and 0.3014, respectively, for *W*1 = 2, while the BRABSB and RANS models reached their highest *MCC* values of 0.2949 and 0.2391, respectively, for *W*1 = 2.5. The ANBPB model achieved the best *MCC* value.

Table 1. Best predictive performances of four sequence encoding schemes.

Sequence Encoding Scheme	*W*1	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
BPB + Ecomposition [a] + Scomposition [b]	2	65.31	65.63	65.52	0.2933
BRABSB + Ecomposition + Scomposition	2.5	73.09	58.16	63.14	0.2949
ANBPB + Ecomposition + Scomposition	**2**	**67.60**	**64.29**	**65.39**	**0.3014**
RANS + Ecomposition + Scomposition	2.5	63.90	61.42	62.24	0.2391

[a] Ecomposition denotes the composition of positively charged amino acids; [b] Scomposition denotes the composition of α-helix propensities of amino acids.

2.1.3. Comparison of the Performance of iSNO-ANBPB with Current Computational Approaches

The classification performances of iSNO-ANBPB, the Li *et al.* method [15], SNOSite [13], iSNO-PseAAC [16], and iSNO-AAPair [17] were compared directly. Because there is no online server for the work done by Li *et al.* [15], iSNO-ANBPB and the Li *et al.* approach were both tested on the training dataset that was constructed in [15]. The results in Table 2 clearly show that iSNO-ANBPB outperformed the Li *et al.* method in the jackknife test. The *Acc* and *MCC* values achieved by iSNO-ANBPB are better by 3.78% and 0.163, respectively, than the *Acc* and *MCC* values achieved by the Li *et al.* method [15]. Further, using an independent Li test dataset, we tested the predictive power of iSNO-ANBPB to recognize novel *S*-nitrosylation sites and compared it with the power of the Li *et al.* method [15], iSNO-PseAAC [16], iSNO-AAPair [17], and SNOSite [13]. As shown in Table 2, the iSNO-ANBPB model achieved an overall accuracy of 63.41% and a *MCC* of 0.2984, which is better than the overall accuracies achieved by the other four methods. We also compared iSNO-ANBPB indirectly with the GPS-SNO predictor proposed by Xue *et al.* [12]. Xu *et al.* [16] reported that iSNO-PseAAC outperformed GPS-SNO when tested on the same benchmark dataset. Therefore, to make a fair comparison, we tested the performances of iSNO-ANBPB, GPS-SNO, and iSNO-PseAAC on the Xu training dataset. The iSNO-ANBPB model again achieved the best prediction performance, with an average accuracy of 70.77% and *MCC* of 0.4146, for the 50 times it was run in the 10-fold crossvalidation. The iSNO-PseAAC model achieved an average accuracy of 67.01% and a *MCC* of 0.3515, and GPS-SNO achieved the best average accuracy of 45.01% and *MCC* of 0.1915 with the threshold set at “low”.

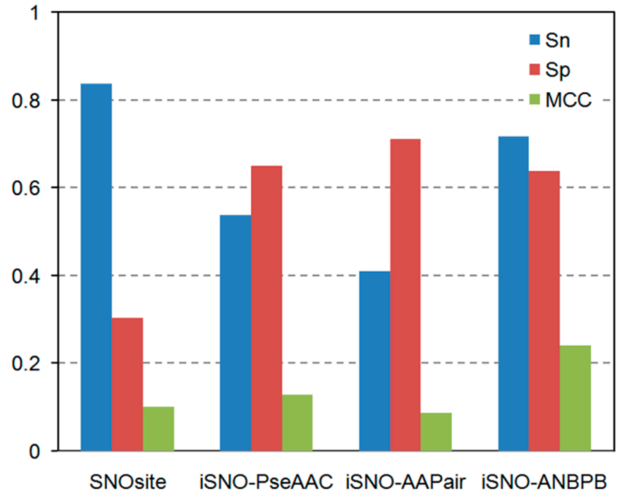
To demonstrate the performance of our iSNO-ANBPB predictor, 37 experimentally-verified *S*-nitrosylated proteins which were not included in the training data set were studied. The sequences of such 37 proteins as well as *S*-nitrosylation site position are given in Supplementary Information. The detailed performances of SNOsite, iSNO-PseAAC, iSNO-AAPair, and iSNO-ANBPB against the 37 independent proteins are summarized in Figure 1. As can be seen from the table, iSNO-ANBPB outperformed the other three predictor in *MCC*, verifying the generalization ability of iSNO-ANBPB.

Table 2. Performance comparison of different computational approaches on different datasets.

Dataset	Methods	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
Li training dataset	Li *et al.* [15]	42.86	70.98	61.61	0.1381
	iSNO-ANBPB	67.60	64.29	65.39	0.3014
Xu dataset	GPS-SNO [a]	18.88	89.63	56.07	0.1210
	GPS-SNO [b]	28.04	81.98	56.39	0.1193
	GPS-SNO [c]	45.01	73.33	59.90	0.1915
	iSNO-PseAAC	67.01	68.15	67.62	0.3515
	iSNO-ANBPB	67.33	73.78	70.77	0.4146
Li test dataset	SNOSite	74.42	28.10	40.24	0.0248
	iSNO-AAPair	27.91	80.17	66.46	0.0858
	Li *et al.* [15]	51.16	69.42	64.63	0.1886
	iSNO-PseAAC	58.14	63.64	62.20	0.1940
	iSNO-ANBPB	74.12	59.50	63.41	0.2984

[a] The data was derived from Table 1 in Xu *et al.* [16] and the threshold of GPS-SNO was set at "high"; [b] The data was derived from Table 1 in Xu *et al.* [16] and the threshold of GPS-SNO was set at "medium"; [c] The data was derived from Table 1 in Xu *et al.* [16] and the threshold of GPS-SNO was set at "low".

Figure 1. Potential *S*-nitrosylation sites predicted on 37 proteins through *S*-nitrosothiols (SNO)site, iSNO-PseAAC, iSNO-AAPair and iSNO-adapted normal distribution bi-profile Bayes (ANBPB) predictor.



2.2. Discussion

Protein *S*-nitrosylation plays a central role in regulatory mechanisms by fine-tuning protein activities associated with diverse cellular processes and biochemical pathway [1,3]. In addition, *S*-nitrosylation appears to have major roles in the etiology of a broad range of human diseases. However, the direct experimental identification of protein *S*-nitrosylation has been challenging, primarily because of the inherent chemical instability of the *S*-NO bond and low abundance of endogenous *S*-nitrosylated proteins [4,5]. Experimental identification of protein *S*-nitrosylation sites has other drawbacks such as expensive experimental costs, time-consuming experiments, and low specificity. Computational techniques have been developed to help overcome these drawbacks.

Moreover, the recent experimental identification of hundreds of *S*-nitrosylation sites opens up the prospect of identifying *S*-nitrosylation sites by combining the experimental data with computer-based screening of peptide sequences.

In this study, we carefully examined the relative position specificity of each amino acid at each position, and identified distinctive amino acid enrichment/depletion profiles for peptide sequences in positive and negative datasets. To encapsulate these complex patterns of residue position-specific propensities for computational prediction, we constructed SVM prediction models using the ANBPB feature extraction approach combined with Chou's PseAAC. ANBPB has been applied to predict protein *O*-GlaNAcylation sites and was shown to significantly improve prediction performance. The theory behind this approach is that the positive and negative profiles for encoding peptide sequences originate from an approximation of the binomial distribution, which can capture and exhibit the relative deviation of frequency of amino acids that surround the *O*-GlaNAcylation sites [24]. Apart from the ANBPB feature extraction, the physicochemical information of the amino acids in the peptide sequence was also considered because it has been demonstrated that the electrostatic charge of amino acids distantly located to a cysteine residue and amino acid propensities for secondary structure are critical for *S*-nitrosylation [15]. The resulting 42 features that we obtained were combined with the SVM classifier to construct our iSNO-ANBPB prediction model.

As described in the above sections, we also established BPB, BRABSB and RANS models to find the most appropriate predictor for protein *S*-nitrosylation. The theoretical distinctions among the four models have been discussed in [24] and the choice of models is determined by the sequence characteristics. For protein *S*-nitrosylation prediction, the ANBPB model gave the best performance, indicating that the ANBPB feature extraction approach may be more suitable than the BPB, BRABSB and RANS approaches for recognizing differences between *S*-nitrosylated and non-*S*-nitrosylated peptide sequences. We suspect that this finding may be because there is a degree of overrepresentation/depletion of certain features in *S*-nitrosylated and non-*S*-nitrosylated peptide sequences. The definition of BPB and BRABSB does not reflect enough the overrepresentation/depletion distinction, so they cannot detect *S*-nitrosylation sites as effectively as the ANBPB model.

We tested our iSNO-ANBPB model against GPS-SNO [12], SNOSite [13], the algorithm developed by Li *et al.* [15], iSNO-PseAAC [16], and iSNO-AAPair [17], because they are among the best *S*-nitrosylation prediction models that are currently available. We could not compare our iSNO-ANBPB model directly with the CPR-SNO predictor [14] because the web-server was not working. Using the Li training dataset, the iSNO-ANBPB model achieved an *Acc* of 65.39%, which is 3.78% higher than the *Acc* for the algorithm developed by Li *et al.* [15]. Using the Xu training dataset, the iSNO-ANBPB model achieved an *Acc* of 70.77%, which is 3.15% higher than the *Acc* achieved by the iSNO-PseAAC method and 11.27% higher than the best *Acc* achieved by GPS-SNO. Notably, the *Acc* achieved by iSNO-ANBPB using the Xu training dataset is about 5.38% higher than of the *Acc* using the Li training dataset, perhaps because the proportion of positive and negative samples in the Xu training dataset is close to 1. Using the Li test dataset, iSNO-ANBPB achieved a *MCC* of 0.2984, which is 0.1044 higher than the previous best-performing predictor iSNO-PseAAC [17], 0.1098 higher than method of Li *et al.* [15], 0.2126 higher than iSNO-AAPair, and 0.2736 higher than SNOSite. The results show that iSNO-ANBPB outperformed previous algorithms in term of precision, especially on independent testing datasets. These datasets are the most likely datasets to be selected for further experimental validation. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors [30,31], we shall make efforts in our future work to provide a web-server for the method presented in this paper.

3. Experimental Section

3.1. Datasets

To objectively and comprehensively compare our approach with other available approaches, we used two recently constructed datasets reported by Li *et al.* [14] and Xu *et al.* [15] (henceforth named the Li and Xu datasets, respectively). The Li training dataset contains 784 positive samples and 1568 negative samples from 499 proteins with <40% sequence similarity, while the Li test dataset contains 43 positive samples and 121 negative samples from 30 proteins with <40% sequence similarity. The Xu training dataset includes 731 positive samples and 810 negative samples from 438 proteins with <40% sequence similarity. Finally, we combined two of the training datasets and removed the redundant samples by by clustering program such as BLASTclust (http://toolkit.tuebingen.mpg.de/blastclust) [32]. The final 1229 positive samples and 1223 negative samples were used to construct the prediction model. After some preliminary trials and in the consideration of the previous works [14,15], we extracted 21-mer *S*-nitrosylation and non-*S*-nitrosylation peptides from both datasets for our analyses. If a possible *S*-nitrosylation site was located at the *N*- or *C*-terminus of the protein and the length of the peptide was less than 21 amino acids, the missing positions were filled with residue "X"s in this study.

3.2. Adapted Normal Distribution Bi-Profile Bayes Features Extraction (ANBPB)

Let $S = s_1, s_2, \cdots, s_n$ denotes a peptide sequence, where s represents an amino acid, *i* (*i* = 1, 2, …, *n*) represents its position, and *n* = 21 represents the length of the peptide sequence in this study. According to bi-profile Bayes method [18], each of the training peptides can be encoded as $(p_1, p_2, \cdots, p_n, p_{n+1}, \cdots, p_{2n})$, where $(p_1, p_2, \cdots, p_n)$ represents the posterior probability of each amino acid at each position in the positive dataset and $(p_{n+1}, p_{n+2}, \cdots, p_{2n})$ represents the posterior probability of each amino acid at each position in the negative dataset. In this study, the frequency of each amino acid at each position was encoded as random variables X_{ij}, *i* (*i* = 1, 2, …, 20) represents the *i*th amino acid $\{\mathrm{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}\}$, and *j* = 1, 2, …, 21 represents the *j*th position. The random variables X_{ij}, (*i* = 1, 2, …, 20; *j* = 1, 2, …, 21) are independent and obey the same binomial distribution *b(n*, *p)*, where *n* = 784/1568 is the number of peptide sequences in positive/negative set, *p* = 1/20 is the probability of each amino acid occurs in each position. Then the normal form variable $\frac{X_{ij} - np}{\sqrt{np(1-p)}}$ has a limiting cumulative distribution function which approximates a normal distribution *N*(0,1). Here, we modified the way of standard variable normalization to highlight and emphasize the distinction of each amino acid at the same position. We let V_j denote the standard variance of $\mathrm{X}_{i,j}$ (*i* = 1, 2, …, 20), *i.e.*, the deviation of frequencies of each at the same *j*th position. And then we $X_{ij}^{'}=\frac{X_{ij}-np}{V_j}$ as the new normalization of X_{ij} and deemed it obeys the standard normal distribution. The posterior probability p_{j} (j = 1, 2, …, 2n) was coded by the adapted normal distribution as follows:

$$p_j=P(X \le X_{i,j})=\varphi(X_{i,j}^{'}) \tag{1}$$

where $\varphi(x)$ is the standard normal distribution function given by $\varphi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$. For more details about this method, please refer to the original paper [24].

3.3. Pseudo Amino Acid Composition Based on Electrostatic Charge and Secondary Structure

To avoid losing many important information hidden in protein or peptide sequences, the pseudo amino acid composition [30,33] or Chou's PseAAC [34] was proposed to replace the simple amino acid composition (AAC) for representing the sample of a protein or peptide. For a brief introduction about Chou's PseAAC, and its recent development and applications, see a comprehensive review [26]. Since the concept of Chou's PseAAC was proposed in 2001, it has rapidly penetrated into almost all the fields of computational proteomics, such as predicting protein submitochondrial localization [35], predicting protein structural class [36], identifying bacterial virulent proteins [37], predicting metalloproteinase family [38], predicting GABA(A) receptor proteins [39], predicting protein supersecondary structure [40], predicting cyclin proteins [41], classifying amino acids [42], identifying risk type of human papillomaviruses [43], identifying GPCRs and their types [44], predicting protein subcellular localization [45], and discriminating outer membrane proteins [46], among many others [26]. Because it has been widely and increasingly used, recently two powerful soft-wares, called "PseAAC-Builder" [47] and "propy" [48], were established recently for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server "PseAAC" (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/) [49] built in 2008.

As indicated by Lee *et al.* [13], Li *et al.* [15], and Marino *et al.* [50], the physicochemical properties of amino acids around cysteine residues can affect the occurrence of cysteine *S*-nitrosylation. Among these properties, electrostatic charge and propensity of secondary structure are critical for protein *S*-nitrosylation [15]. Accordingly, the 20 amino acids were divided into two different classes based on their electrostatic charge: positively charged amino acids (A): {A, C, D, E, H, L, P, Q, S, V, W} and negatively charged amino acids (G): {F, G, I, K, M, N, R, T, Y}. Similarly, based on their secondary structure, the 20 amino acids were divided into two other classes: α-helix propensities of amino acids (H): {C, D, G, N, P, S, T, W, Y} and other amino acids (E): {A, E, F, H, I, K, L, M, Q, R, V}. Owing to the summation of composition of pseudo amino acids (A) and composition of (G) is equal to 1, only one is independent. The same cases for the composition of pseudo amino acids (H) and composition of (E). So in practical calculations, the composition of positively charged amino acids (A) and α-helix propensities of amino acids (H) are adopted to construct the feature vectors.

3.4. Feature Space

According to the recent review [26], a peptide segment in our positive and negative datasets is formulated by

$$P=[\psi_1,\psi_2,\cdots,\psi_{42}] \tag{2}$$

where $\psi_i(i=1,2,\cdots,20)$ was defined by the posterior probability p_i of each amino acid at each position in positive peptide sequences datasets; $\psi_i(i=21,22,\cdots,40)$ was defined based on the posterior probability p_i of each amino acid at each position in negative peptide sequences datasets; ψ_{41},ψ_{42} were the composition of pseudo amino acids (A), and (H), respectively.

3.5. Support Vector Machine Implementation and Parameter Selection

An SVM is a set of related supervised learning methods used for classification and regression based on statistical learning theory. The SVM method has proven to be powerful in many fields of bioinformatics [18–20,51,52]. In this study, the SVM was trained with the LIBSVM package [53] to build the model and perform the predictions. The radial basis kernel function $k(x_i,x_j)=\exp\{-\gamma \| x_i-x_j \|^2\}$ was used for our SVM method. For different input features, the penalty parameter C and kernel parameter γ were optimized using the SVMcgForClass program [53] in the LIBSVM package based on a 15-fold cross-validation. The final parameters that we obtained were C = 22.6274 and $\gamma=0.03125$. Optimized weight parameters (*W*1 and *W*-1) were set as 2 and 1 by looking for the best jackknife test results.

3.6. Performance Assessments

The jackknife test was used in this study to evaluate our method because it is considered as the most objective cross-validation method [31]. Sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*) and *MCC* were used to quantify the performance of our method. They are defined as follows:

$$Sn=\frac{TP}{TP+FN} \tag{3}$$

$$Sp=\frac{TN}{TN+FP} \tag{4}$$

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

where *TP*, *TN*, *FP* and *FN* denote the number of true positives (correctly predicted *S*-nitrosylation sites), true negatives (correctly predicted non-*S*-nitrosylation sites), false positives (falsely predicted *S*-nitrosylation sites), and false negatives (falsely predicted non-*S*-nitrosylation sites), respectively.

Acknowledgments

We thank Yudong Cai (Shanghai University) and co-workers for sharing their datasets. This work was supported by the Fundamental Research Funds for the Central Universities, China (3132013332, 3132013093).

Author Contributions

Cangzhi Jia developed the statistical results and algorithms and carried out their implementation and application with iSNO-ANBPB model. Xin Lin contributed to the tissue sample collection, data analysis. Zhiping Wang conceived of the study, and participated in its design and coordination and helped to draft the manuscript. Zhiping Wang financed the project.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Liu, M.; Talmadge, J.E.; Ding, S.J. Development and application of site-specific proteomic approach for study protein *S*-nitrosylation. *Amino Acids* **2012**, *42*, 1541–1551.
2. Tuteja, N.; Tuteja, M.R.; Misra, M.K. Nitric oxide as a unique bioactive signaling messenger in physiology and pathophysiology. *J. Biomed. Biotechnol.* **2004**, *4*, 227–237.
3. Lane, P.; Hao, G.; Gross, S.S. *S*-nitrosylation is emerging as a specific and fundamental posttranslational protein modification: Head-to-head comparison with *O*-phosphorylation. *Sci. STKE* **2001**, *86*, doi: 10.1126/stke.2001.86.re1.
4. Forrester, M.T.; Foster, M.W.; Benhar, M.; Stamler, J.S. Detection of protein *S*-nitrosylation with the biotin-switch technique. *Free Radic. Biol. Med.* **2009**, *46*, 119–126.
5. Forrester, M.T.; Thompson, J.W.; Foster, M.W.; Nogueira, L.; Moseley, M.A.; Stamler, J.S. Proteomic analysis of *S*-nitrosylation and denitrosylation by resin-assisted capture. *Nat. Biotechnol.* **2009**, *27*, 557–559.
6. Foster, M.W.; McMahon, T.J.; Stamler, J.S. *S*-nitrosylation in health and disease. *Trends Mol. Med.* **2003**, *9*, 160–168.
7. Lim, K.H.; Ancrile, B.B.; Kashatus, D.F.; Counter, C.M. Tumour maintenance is mediated by eNOS. *Nature* **2008**, *452*, 646–649.
8. Mannick, J.B.; Schonhoff, C.M. Measurement of protein *S*-nitrosylation during cell signaling. *Methods Enzymol.* **2008**, *440*, 231–242.
9. Jaffrey, S.R.; Snyder, S.H. The biotin switch method for the detection of *S*-nitrosylated proteins. *Sci. STKE* **2001**, *86*, doi: 10.1126/stke.2001.86.pl1.
10. Huang, B.; Chen, C. An ascorbate-dependent artifact that interferes with the interpretation of the biotin switch assay. *Free Radic. Biol. Med.* **2006**, *41*, 562–567.
11. Hao, G.; Derakhshan, B.; Shi, L.; Campagne, F.; Gross, S.S. SNOSID, a proteomic method for identification of cysteine *S*-nitrosylation sites in complex protein mixtures. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1012–1017.

12. Xue, Y.; Liu, Z.; Gao, X.; Jin, C.; Wen, L.; Yao, X.B.; Ren, J. GPS-SNO: Computational prediction of protein *S*-nitrosylation sites with a modified GPS algorithm. *PLoS One* **2010**, *5*, e11290.
13. Lee, T.Y.; Chen, Y.J.; Lu, T.C.; Huang, H.D.; Chen, Y.J. SNOSite: Exploiting maximal dependence decomposition to identify cysteine *S*-Nitrosylation with substrate site specificity. *PLoS One* **2011**, *6*, e21849.
14. Li, Y.X.; Shao, Y.H.; Jing, L.; Deng, N.Y. An efficient support vector machine approach for identifying protein *S*-nitrosylation sites. *Protein Pept. Lett.* **2011**, *18*, 573–587.
15. Li, B.Q.; Hu, L.L.; Niu, S.; Cai, Y.D.; Chou, K.C. Predict and analyze *S*-nitrosylation modification sites with the mRMR and IFS approaches. *J. Proteome Res.* **2012**, *75*, 1654–1665.
16. Xu, Y.; Ding, J.; Wu, L.Y.; Chou, K.C. iSNO-PseAAC: Predict cysteine *S*-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* **2013**, *8*, e55844.
17. Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine *S*-nitrosylation sites in proteins. *Peer J.* **2013**, *1*, e171.
18. Shao, J.; Xu, D.; Tsai, S.N.; Wang, Y.; Ngai, S.M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One* **2009**, *4*, e4920.
19. Song, J.; Tan, H.; Shen, H.; Mahmood, K.; Boyd, S.E.; Webb, G.I.; Akutsu, T.; Whisstock, J.C. Cascleave: Towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* **2010**, *26*, 752–760.
20. Wee, J.K.; Simarmata, D.; Kam, Y.W. SVM-based prediction of linear B-cell epitopes using Bayes feature extraction. *BMC Genomics* **2010**, *11*, S21.
21. Jia, C.Z.; Liu, T.; Chang A.K.; Zhai, Y. Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. *Biochimie* **2011**, *93*, 778–782.
22. Wang, Y.; Zhang, Q.; Sun, M.; Guo, D. High accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* **2011**, *27*, 777–784.
23. Shao, J.; Xu, D.; Hu, L.; Kwan, Y.W.; Wang, Y.; Kong, X.; Ngai, S.M. Systematic analysis of human lysine acetylation proteins and accurate prediction of human lysine acetylation through bi-relative adapted binomial score Bayes feature representation. *Mol. Biosyst.* **2012**, *8*, 2964–2973.
24. Jia, C.Z.; Liu, T.; Wang, Z.P. *O*-GlcNAcPRED: A sensitive predictor to capture protein *O*-GlcNAcylation sites. *Mol. BioSyst.* **2013**, *9*, 2909–2913.
25. Jia, C.Z.; Zhang, Y.S.; Wang, Z.P. SulfoTyrP: A high accuracy predictor of protein sulfotyrosine sites. *Match Commun. Math. Comput. Chem.* **2014**, *71*, 227–240.
26. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247.
27. Chen W.; Feng P.M.; Lin H.; Chou K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68.

28. Feng P.M.; Chen W.; Lin H.; Chou K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **2013**, *442*, 118–125.
29. Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints, *J. Theor. Biol.* **2013**, *337*, 71–79.
30. Lin, S.X.; Lapointe, J. Theoretical and experimental biology in one. *J. Biomed. Sci. Eng.* **2013**, *6*, 435–442.
31. Chou, K.C.; Shen, H.B. Recent progress in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1–16.
32. *BLASTclust*. Available online: http://toolkit.tuebingen.mpg.de/blastclust (accessed on 1 December 2010).
33. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct. Funct. Genet.* **2001**, *43*, 246–255.
34. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19.
35. Nanni, L.; Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* **2008**, *34*, 653–660.
36. Sahu, S.S.; Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **2010**, *34*, 320–327.
37. Nanni, L.; Lumini, A; Gupta, D.; Garg, A. 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.* **2012**, *9*, 467–475.
38. Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics* **2011**, *12*, 191–197.
39. Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2011**, *281*, 18–23.
40. Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.* **2011**, *32*, 271–278.
41. Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Peptide Lett.* **2010**, *17*, 1207–1214.
42. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* **2009**, *257*, 17–26.
43. Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* **2010**, *263*, 203–209.
44. Zia Ur, R.; Khan, A. 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 Lett.* **2012**, *19*, 890–903.

45. Zhang, S.W.; Zhang, Y.L.; Yang, H.F.; Zhao, C.H.; Pan, Q. 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* **2008**, *34*, 565–572.
46. Hayat, M.; Khan, A. Discriminating outer membrane proteins with fuzzy *K*-nearest neighbor algorithms based on the general form of Chou's PseAAC. *Protein Peptide Lett.* **2012**, *19*, 411–421.
47. Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **2012**, *425*, 117–119.
48. Cao, D.S.; Xu, Q.S.; Liang, Y.Z. Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, *29*, 960–962.
49. Shen, H.B.; Chou, K.C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochems.* **2008**, *373*, 386–388.
50. Marino, S.M.; Gladyshev, V.N. Structural analysis of cysteine *S*-nitrosylation: A modified acid-based motif and the emerging role of trans-nitrosylation. *J. Mol. Biol.* **2009**, *395*, 844–859.
51. Song, X.; Wang, M.; Chen, Y.P.; Wang, H.; Han P.; Sun, H. Prediction of pre-miRNA with multiple stem-loops using pruning algorithm. *Comput. Biol. Med.* **2013**, *43*, 409–416.
52. Kazemian, H.B.; White, K.; Brown, D.P. Applications of evolutionary SVM to prediction of membrane alpha-helices. *Expert Syst. Appl.* **2013**, *40*, 3412–3420.
53. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. Available online: http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed on 1 April 2014).