SCIENTIFIC REPORTS

OPEN

Received: 13 August 2018
Accepted: 1 February 2019
Published online: 28 February 2019

Identification of S-nitrosylation sites based on multiple features combination

Taoying Li, Runyu Song, Qian Yin, Mingyue Gao & Yan Chen

Protein S-nitrosylation (SNO) is a typical reversible, redox-dependent and post-translational modification that involves covalent modification of cysteine residues with nitric oxide (NO) for the thiol group. Numerous experiments have shown that SNO plays a major role in cell function and pathophysiology. In order to rapidly analysis the big sets of data, the computing methods for identifying the SNO sites are being considered as necessary auxiliary tools. In this study, multiple features including Parallel correlation pseudo amino acid composition (PC-PseAAC), Basic kmer1 (kmer1), Basic kmer2 (kmer2), General parallel correlation pseudo amino acid composition (PC-PseAAC_G), Adapted Normal distribution Bi-Profile Bayes (ANBPB), Double Bi-Profile Bayes (DBPB), Bi-Profile Bayes (BPB), Incorporating Amino Acid Pairwise (IAAPair) and Position-specific Tri-Amino Acid Propensity(PSTAAP) were employed to extract the sequence information. To remove information redundancy, information gain (IG) was applied to evaluate the importance of amino acids, which is the information entropy of class after subtracting the conditional entropy for the given amino acid. The prediction performance of the SNO sites was found to be best by using the cross-validation and independent tests. In addition, we also calculated four commonly used performance measurements, i.e. Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), and the Matthew's Correlation Coefficient (MCC). For the training dataset, the overall Acc was 83.11%, the MCC was 0.6617. For an independent test dataset, Acc was 73.17%, and MCC was 0.3788. The results indicate that our method is likely to complement the existing prediction methods and is a useful tool for effective identification of the SNO sites.

Protein post-translational modifications play a very important role in the processing of protein, protein maturation, as well as altering the physical and chemical properties of proteins. As a result, the space conformation, three-dimensional location and the stability of the proteins are likely change, which can lead to the function alteration. Moreover, the structural features of the modified groups can produce a far-reaching impact on the properties, as well as, the functions of proteins. In 1998[1], the Nobel Prize for Physiology or Medicine was rewarded for breakthrough discoveries that showed nitric oxide to be a freely-diffusible signaling molecule and a secondary messenger. NO plays a vital role in the cardiovascular system[2]. It is noticed that S-nitrosylation (SNO) is the covalent interaction of nitric oxide with the thiol group of cysteine residues[1,3] and is well characterized as a major source of NO bioactivity[4]. Many experimental methods have been applied for distinguishing the SNO sites, such as the biotin-switch technique (BST)[5,6], SNO-Cys site identification (SNOSID)[7–9], and the resin-associated capture (RAC)[10]. These experimental methods have successfully provided a very effective information in identifying the SNO sites. The BST was designed to purify and detect the SNO proteins, mainly composed of three principal steps: (i) The methylthiolation of free cysteine thiols with methyl methanethiosulfonate (MMTS); (ii) Reduction of SNOs to thiols with ascorbate; (iii) Ligation of the nascent thiols with N-[6-(Biotinamido)hexyl]-3′-(2′-Pyridyldithio)-propionamide (biotin-HPDP)[11]. In combination with the traditional mass spectrometry (MS), BST has indeed contributed to discovering a lot of potential protein SNO sites[12–15]. A proteomic method called SNOSID, that identified the endogenous and chemically-induced SNOs in the proteins from tissues or cells, was also developed to determine the potential SNO sites on the cysteine residues in complex protein mixtures. Furthermore, RAC based method was also developed to detect the SNO proteins[10]. In 2009, Foster *et al.*[16], explored a protein microarray-based approach to screen the SNO sites. These methods made great contributions to the development of the prediction of SNOs, however, to a certain degree, they were considered to be time-consuming and also had a relatively low throughput data. Recently, several machine learning approaches

Department of Maritime Economics and Management, Dalian Maritime University, No. 1 Linghai Road, Dalian, 116026, China. Correspondence and requests for materials should be addressed to T.L. (email: ytaoli@126.com)

Dimension	Sequence encoding schemes	Sn (%)	Sp (%)	Acc (%)	MCC
25	PC-PseAAC	72.33	54.32	62.82	0.2699
45	PC-PseAAC + kmer1	68.67	61.36	64.83	0.3004
445	PC-PseAAC + kmer1 + kmer2	64.3	65.43	64.89	0.297
470	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G	64.57	65.43	65.02	0.2997
510	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G + ANBPB	65.53	72.84	69.37	0.3848
548	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G + ANBPB + DBPB	65.94	72.84	69.57	0.3888
588	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G + ANBPB + DBPB + BPB	65.94	72.96	69.63	0.3901
607	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G + ANBPB + DBPB + BPB + IAAPair	66.48	73.21	70.02	0.3979
645	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G + ANBPB + DBPB + BPB + IAAPair + PSTAAP	73.46	74.94	74.24	0.4837`

Table 1. Performance of the combination feature with different sequence encoding schemes in jackknife test.

have been proposed and have provided helpful information being used for further experimental verification of the protein SNO sites. Hao *et al.*[7] developed a prediction tool for the SNO sites, which was based on the support vector machine (SVM)[17] algorithm, and used a training dataset that consisted of 65 positive SNO sites and 65 non-SNO sites. A few years later, Xue *et al.*[2] proposed a method called GRS-SNO by using a group-based predicting system based on 504 experimentally verified SNO sites in 327 unique proteins. Shortly afterward, Li *et al.*[18] established the predictor CPR-SNO and built a web server based on a coupling pattern encoding scheme. Xu *et al.*[19,20] developed the iSON-AApair that takes into account, the effects of sequence correlation. More recently, Jia *et al.*[21] used an Adapted Normal Distribution Bi-Profile Bayes (ANBPB) and Chou's PseAAC composition constituting the feature vector. The composition of Zhang *et al.*[22] were also based on the Chou's PseAAC, by incorporating the various sequences derived feature.

Each of the above mentioned methods had their own advantage, as well as, played an important role in the research for prediction of protein S-nitrosylation sites. However, it is noted that the prediction performance is not really satisfactory. Therefore, there is necessity to discover more efficient methods for the SNO sites identification.

In this study, we extracted nine types of features, including PC-PseAAC (25), kmer1 (20), kmer2 (400), PC-PseAAC_G (25), ANBPB (40), DBPB (38), BPB (40), IAAPair (39) and PSTAAP (18). In order to remove the redundant information, the information gain (IG) method was applied to select the features. Finally, the optimization of 425D feature vector (PC-PseAAC (25), kmer1 (20), kmer2 (180), PC-PseAAC_G (25), ANBPB (40), DBPB (38), BPB (40), IAAPair (39), and PSTAAP (18) was used to construct our prediction model. Our results suggest that IG can provide an improved performance, which is comparable to the one without the use of the IG method. The results indicated that selecting the IG feature is a promising method to predict the features with high dimension with the SNO sites.

Results and Discussion

Combination of different features.

To evaluate the performances of the combined feature sets for sorting SNO sites and non-SNO sites, we tested the prediction performances on the Jackknife test[23], which is considered as the most objective and always yields a unique result for a given dataset[21,24]. The combined features were composed of the PC-PseAAC, kmer1, kmer2, PC-PseAAC_G[25], ANBPB[21], DBPB, BPB[26], IAAPair[19], and PSTAAP[20,27] models and the detailed results are shown in Supplementary Table S1. The results show that the prediction performance was enhanced through the combined features. As shown in Table 1, PC-PseAAC with the Acc of 62.82% was regarded as the basic features, and was then incorporated to kmer1 to improve the prediction performance, which reached the Acc of 64.83%. Secondly, combination of features PC-PseAAC + kmer1 were further incorporated with the component of kmer2 one by one, and new combined features PC-PseAAC, kmer1 and kmer2 reached Acc of 64.89%. This process was terminated at feature combination PC-PseAAC, kmer1, kmer2, PC-PseAAC_G, ANBPB, DBPB, BPB, IAAPair, and PSTAAP, which increased the Acc to 74.24% and MCC[17,28–30] to 0.4837. From the above, it can be concluded that the combined features can improve Acc of 11.42%. The parameters λ and the weight factor w were found to offer the best results for the features PC-PseAAC and PC-PseAAC_G and the optimized values were $\lambda = 5$ and $w = 0.5$.

Features selection via IG.

To further improve the prediction performance, these features were optimized based on the above-mentioned IG optimization method. The four types of features PC-PseAAC, kmer1, kmer2, and PC-PseAAC_G are mainly related to the frequency of amino acids but are independent of the position of protein sequences. Hence, we optimized these four types of features based on the IG score of the amino acid residues. Firstly, we sorted the importance of amino acid composition (AAC) and the amino acid pair composition (i.e. kmer2) by IG score, and then applied the incremental feature selection to find out the best feature subset for maximizing prediction performance. According to the final performance evaluation, the application of IG score on kmer2 was especially distinguishable. The detailed prediction performances for different number of features combination on 10-fold cross-validation were shown in Fig. 1. It can be seen that when the dimension for the feature vector selected to be 180, the predictive performance achieved the highest value with Sn of 72.79%, Sp of 74.64%, Acc of 73.71%, and MCC of 0.4741. However, there was no obvious improvement for the other three types features PC-PseAAC, kmer1, and PC-PseAAC_G. This could be due to the low dimensions of these three types of features (less than 50). On the contrary, the dimension of kmer2 was 400, and the feature matrix was an

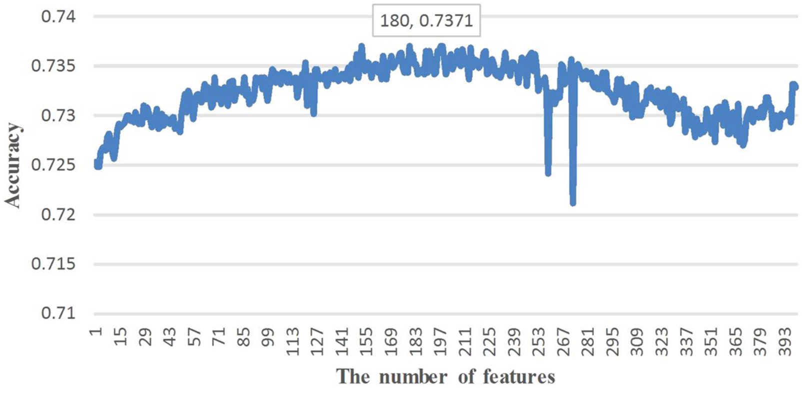

Figure 1. The predictive performance of different models based on incremental feature selection of features sorted by IG.

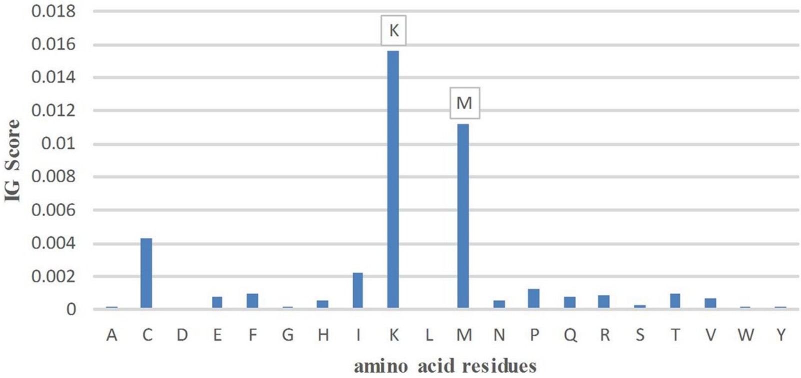

Figure 2. The IG score of each amino acid residues.

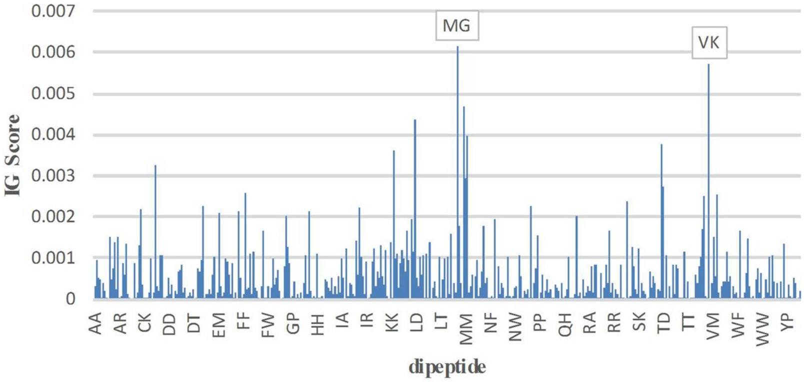

Figure 3. The IG score of each dipeptide.

extremely sparse matrix and hence having IG reflecting a good performance. The results of the IG score ranking importance of amino acid residues and dipeptide are displayed in Fig. 2 and Fig. 3, respectively and the detailed results are shown in Supplementary Table S2. It is noteworthy that the amino acid residues K, M, and C and the dipeptides MG, VK, and ML exhibited a great contribution to the prediction performance. Fig. 2 and Fig. 3 showed that the highest IG score reached 0.0156, 0.0112 and 0.0043 for the amino acid residues K, M, and C, respectively, while the highest IG score reached 0.0062, 0.0057 and 0.0047 for the amino acid dipeptides MG, VK, and ML, respectively.

Before the features selection, the prediction performance with the Sn of 73.46%, the Sp of 74.94%, and the Acc of 74.28%. After removing the irrelevant feature and then determining the optimal combination of features, we then obtained the best prediction performance with the Sn of 73.60%, the Sp of 75.93% and the Acc of 74.82, respectively. As can be seen, all of the three measurements have been improved slightly. But the prediction

IG Dimension	Sequence encoding schemes	Sn (%)	Sp (%)	Acc (%)	MCC
645	PC-PseAAC + kmer1 + kmer2 + PC-PseAAC_G + ANBPB + DBPB + BPB + IAAPair + PSTAAP	73.46	74.94	74.24	0.4837
425	PC-PseAAC + kmer1 + **kmer2** + PC-PseAAC_G + ANBPB + DBPB + BPB + IAAPair + PSTAAP	73.60	75.93	74.82	0.4952
425	**PC-PseAAC** + kmer1 + kmer2 + PC-PseAAC_G + ANBPB + DBPB + BPB + IAAPair + PSTAAP	73.60	75.93	74.82	0.4952
425	PC-PseAAC + **kmer1** + kmer2 + PC-PseAAC_G + ANBPB + DBPB + BPB + IAAPair + PSTAAP	73.60	75.93	74.82	0.4952
425	PC-PseAAC + kmer1 + kmer2 + **PC-PseAAC_G** + ANBPB + DBPB + BPB + IAAPair + PSTAAP	73.60	75.93	74.82	0.4952

Table 2. Features optimization based on IG on Jackknife test. Bold blackbody is the feature extracted by IG method.

performance was not satisfied, it need us to make improvements on this work in the future. The results for the best predictive performance are shown in Table 2. An improved predictive Acc for the models that were trained with the optimized features was being seen when compared with the model with non-optimized features. As given in Supplementary Table S2, such 425 [PC-PseAAC(25) + kmer1(20) + kmer2(180) + PC-PseAAC_G(25) + ANBPB(40) + DBPB(38) + BPB(40) + IAAPair(39) + PSTAAP(18)] features regarded as the optimal feature set for the selected model. Based on the 425 features, the predictive Sn, Sp, and Acc were 73.60%, 75.93% and 74.82%, respectively. These results indicate that the key amino acid residues and the key dipeptide used in optimizing the models can enhance the prediction performance of the SNO sites. Consequently, the features combined with key amino acid residues were applied to implement a novel and high-performance tool for identifying cysteine S-nitrosylated sites.

Comparison with other feature selection methods. In this paper, different feature selection methods were exploited for comparison. We made several comparisons for evaluating the performance of IG with Max-Relevance-Max-Distance[31] (MRMD), a method for feature selection. MRMD contains two components, max distance and maximal relevance. The max distance selects a new feature which has the least redundancy in the residual of features, while the maximal relevance selects feature that has the strongest relevance to the target class.

We used four distance methods ED, COS, TC and Mean of MRMD to find out the best feature vectors combination through using 10-fold cross-validation. The detailed predictive performances are listed in Fig. 4. When the distance function ED was adopted, its best Acc achieved 73.32% with 356 features. And when the distance functions are COS, TC and Mean, the predictive performance is the highest with 397, 398 and 43 features, respectively, whose corresponding predictive performance is 73.14%, 73.35% and 73.34%. Suppose that the total dimension of feature vector is 400, the influence of dimension reduction is not obvious when the distance function ED, COS and TC are used (the predictive performance is the best with 356, 397 and 398 features, respectively). However, the influence of dimension reduction is prominent when the Mean distance function is used (the predictive performance is the best with 43 features), which causes a lot of information lost in the feature vector. The best performances for different feature selection methods are listed in Supplementary Tables S8–11.

From Fig. 4A, we can see that although the performances of two methods, IG and four types of MRMD, are almost identical on the same datasets, Acc of IG has better advantageous. Meanwhile, its Acc is generally higher than that of MRMD method, including ED, COS, TC and Mean. Moreover, it has more advantages to achieve the dimensionality reduction of high-dimensional eigenvectors and unsure high Acc. From Fig. 4B–E, show the predictive performance of different dimensions eigenvectors are shown when MD is ED, COS, TC and Mean, respectively.

Comparison with other methods. To make a fair and fast comparison, we compared the prediction performance of our predictor with GPS-SNO[2], iSNO-PseAAC[20], iSNO-ANBPB[21], PSNO[22], iSNO-AAPair[19] on the Xu training dataset by running 10-fold cross-validation test 10 times. The results were shown in Table 3. Our constructed model exhibits the best prediction performance with Acc of 83.11%, which was 1.41% higher than the previous best-performing predictor iSNO-AAPair, and 7.44% higher than Acc achieved by PSNO. Our predictor also gave a MCC of 0.6617, which was 0.0317 higher than the method of iSNO-AAPair, and 0.1498 higher than PSNO. Furthermore, Sn of our predictor was 83.33%, which was 3.73% higher than Sn of iSNO-AAPair, and 9.18% higher than PSNO. This comparison indicates that the proposed model is indeed promising and could at least play a role that complements the existing state-of-the art methods in this field. In addition, we tested the predictive power of our model with the powers of the SNOSite[32], iSNO-AAPair[19], iSNO-PseAAC[20], iSNO-ANBPB[21] on the Li test dataset; and we also compared our model with the GPS-SNO[2], iSNO-PseAAC[20], iSNO-AAPair[19], and PSNO[22] methods on Xu test dataset. The performances of the above-mentioned models against two test datasets are summarized in Supplementary Tables S3 and S4. On the Li independent test dataset, our model captured proteins O00429 (site 367), P13221 (site 83), P43235 (site 139) as S-nitrosylation sites, while methods iSNO-AAPair and iSNO-PseAAC incorrectly predicted them as non-S-nitrosylation sites. On the Xu independent test dataset, our model captured proteins O70572 (site 176), P51174 (site 342), Q8VDG5 (site 308), Q9WVQ5 (site 146), P55060 (site 344) as S-nitrosylation sites, while models iSNO-PseAAC and GPS-SNO incorrectly predicted S-nitrosylation sites as non- S-nitrosylation sites. To show the prediction results clearly, we summarized Sn, Sp, ACC and MCC that was achieved by each model in Table 4. As it can be seen that our predictor achieved

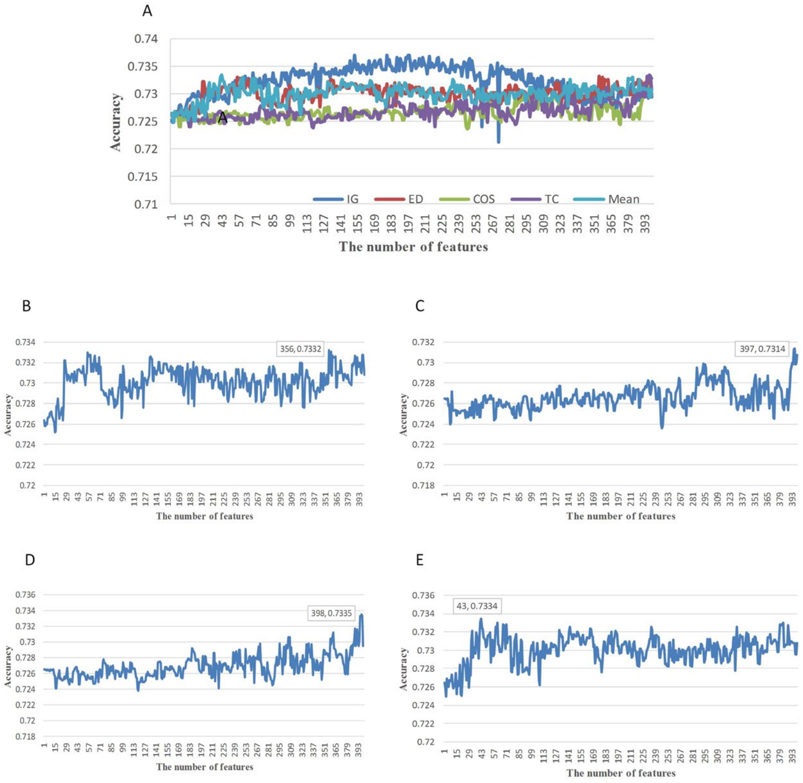

Figure 4. The predictive performance of different models and the comparison of their. (**A**) Comparison on the predictive performance of different feature selection methods. (**B**) The predictive performance of different models based on incremental feature selection of features sorted by distance ED of MRMD. (**C**) The predictive performance of different models based on incremental feature selection of features sorted by distance COS of MRMD. (**D**) The predictive performance of different models based on incremental feature selection of features sorted by distance TD of MRMD. (**E**) The predictive performance of different models based on incremental feature selection of features sorted by distance Mean of MRMD.

Dataset	Test Method	Methods	Sn (%)	Sp (%)	Acc (%)	MCC
Xu training dataset	the 10_fold cross-validation test	GPS-SNO	45.01	73.33	59.9	0.1915
		iSNO-PseAAC	67.01	68.15	67.62	0.3515
		iSNO-ANBPB	67.33	73.78	70.77	0.4146
		PSNO	74.15	77.04	75.67	0.5119
		iSNO-AAPair	79.60	84.10	81.70	0.6300
		Our predictor (maximum)	83.33	82.92	83.11	0.6617
		Our predictor (average)	72.16	74.90	73.54	0.4704
	jackknife test	IG-SCORE	73.60	75.93	74.82	0.4952

Table 3. Compare with other methods performance on the training dataset.

the performance with Sn of 60.47%, Sp of 77.69% and Acc of 73.17% on the Li test dataset. Among the other five methods, the best prediction performance was achieved by the method of Li *et al.*, with Sn of 51.16%, Sp of 69.42% and Acc of 64.63%. Our method is obviously superior to other methods. However on Xu test dataset, our predictor achieved the prediction performance with the Sn of 64.20%, the Sp of 75.00%, and the Acc of 70.17%, which is only better than iSNO-PseAAC with Sn of 50.2%, Sp of 75.1% and Acc of 62.8%. The results show that our predictor outperformed previous methods in terms of precision. But on the Xu test set, the results are not ideal, which may be caused as a result of not considering the physical chemistry properties. In the future work, we will consider more compressive features and further optimize the feature combination approaches.

Dataset	Methods	Sn (%)	Sp (%)	Acc (%)	MCC
Li test dataset	SNOSite	74.42	28.1	40.24	0.0248
	iSNO-AAPair	27.91	80.17	66.46	0.0858
	Li *et al.*	51.16	69.42	64.63	0.1886
	iSNO-PseAAC	58.14	63.64	62.2	0.1940
	iSNO-ANBPB	74.12	59.5	63.41	0.2984
	Our predictor	60.47	77.69	73.17	0.3588
Xu test dataset	GPS-SNO	44.5	81.0	64.7	0.2800
	iSNO-PseAAC	50.2	75.2	62.8	0.3000
	iSNO-AAPair	79.6	84.1	81.7	0.6300
	PSNO	87.7	85.0	86.2	0.7200
	Our predictor	64.20	75.00	70.17	0.3942

Table 4. Compare with other methods performance on the test datasets.

Conclusion

The prediction of SNO sites is essential for better understanding of the basic biological theory, clinical diagnosis as well as the pharmaceuticals. In this study, we introduce the IG which is a tool for the analysis of the importance of amino acid and its position used in feature extraction. Here, we focus on the characteristics of the amino acids with its sequence. Four out of the nine characteristics of the combination are related to the amino acid residues. The four types of features were screened using the IG method, and the best dimension of the feature vector was selected. Among these features, 180 important features were screened from the feature kmer-2 whose dimension was reduced from 400 to 180, with the best prediction performance, and Sn and Acc are reached 83.33% and 83.11%, respectively. Theoretically, there is a lot of information in the protein structure when compared with the simple sequences, and this will be considered in the future scope of the work. With the development of internet and big data era coming, constructing databases[33–40] and establishing powerful webserver are the direction of bioinformatics. Thus, making it convenient to most experimental scientists

Material and Methods

Datasets.

The datasets were constructed using those of Li *et al.*[22] and Xu *et al.*[19,20] (henceforth named the Li dataset and Xu dataset, respectively). As described previously[19,20,24], these datasets were derived on the basis of the experimental verification of the protein S-nitrosylation sites. Xu training dataset consisted of 731 positive SNO sites as positive samples and 810 non-SNO cysteine sites as negative samples from the 438 proteins with <=40% sequence similarity. These samples were used for training our prediction model. The Xu test dataset consisted of 81 SNO sites and 100 non-SNO sites, and the Li test dataset included 43 SNO sites and 121 non-SNO sites. In this study, Xu and Li test datasets were applied to test the prediction performance of our model.

Considering that we have a protein peptide sample P in our datasets, which can be generally formulated by:

$$\mathrm{P} = R_{-t}R_{-(t-1)}\ldots R_{-2}R_{-1}(C)R_{+1}R_{+2}\ldots R_{+(t-1)}R_{+t} \tag{1}$$

where the subscript t is an integer, R_{-t} is the t-th downstream amino acid residue from cysteine(C), R_t the t-th upstream amino acid residue, and so forth. The peptide was termed as SNO or non-SNO peptide depending on whether its center is a SNO or non-SNO sites, respectively. P belonged to one of two categories *viz.* the SNO sites (positive data) or non-SNO sites (negative data). In the current study, we selected t = 10. If the upstream or downstream in a protein was less than 10, the lacking residues were filled using the dummy code X. Thus, the training dataset S was formulated as (⋃: in the set theory to formulate the union of datasets):

$$\mathrm{S} = S^{+} \bigcup S^{-} \tag{2}$$

where the positive dataset S^{+}consisted of 731 SNO cysteine sites, while the negative dataset S^{-} contained 810 non-SNO cysteine sites; The test dataset T_{Li} and T_{Xu} was formulated as:

$$\mathrm{T}_{Li} = \mathrm{T}_{Li}^{+} \bigcup \mathrm{T}_{Li}^{-} \tag{3}$$

$$\mathrm{T}_{Xu} = \mathrm{T}_{Xu}^{+} \bigcup \mathrm{T}_{Xu}^{-} \tag{4}$$

where the positive dataset T_{Li}^{+} and T_{Xu}^{-} contained 43 and 81 SNO peptide fragments, respectively; while the negative dataset T_{Li}^{-} and T_{Xu}^{-} contained 121 and 100 non-SNO peptide fragments, respectively. For the reader's convenience, the three datasets used in this study are given in Supplementary Tables S5–7. The schematic flowchart of our work is being shown in Fig. 5.

Features extraction.

Parallel correlation pseudo amino acid composition (PC-PseAAC). PC-PseAAC[41] is the feature extraction approach that incorporates the contiguous local and the global sequence-order information to obtain the feature vector for the protein sequence. Given a protein peptide P (Eq. 1), the PC-PseAAC feature vector for P is given by:

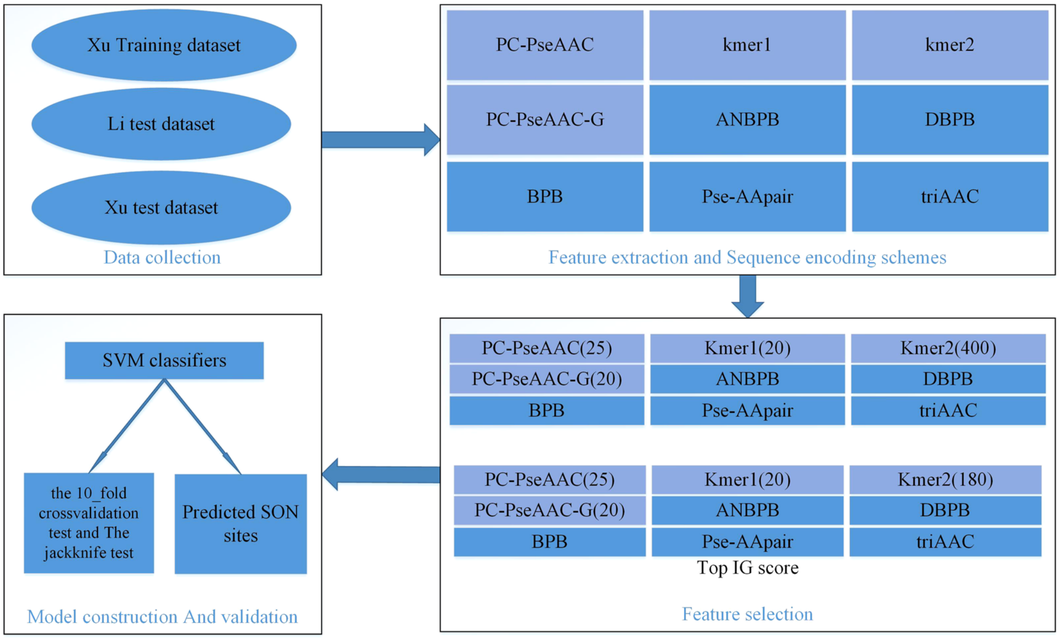

Figure 5. Flowchart of our predictor methodology.

$$\mathrm{V} = [x_1, x_2, \ldots, x_{20}, x_{21}, \ldots, x_{20+\lambda}]^T \tag{5}$$

where

$$\mathrm{x}_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\Theta_j}(1 \le u \le 20) \\ \dfrac{w\Theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\Theta_j}(20 + 1 \le u \le 20 + \lambda) \end{cases} \tag{6}$$

where f_i (i = 1,2, …, 20) is the normalized occurrence frequency of the 20 amino acids in the protein P; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; w is the weight factor ranging from 0 to 1; and Θ_j (j = 1,2, …, 20) is the j-tier correlation factor reflecting the sequence-order correlation between all the j-th most contiguous residues along a protein chain, which is defined as:

$$\Theta_\lambda = \frac{1}{L - \lambda}\sum_{i=1}^{L-\lambda}\Theta(R_i, R_{i+\lambda})\,(0 < \lambda < L) \tag{7}$$

where the correlation function is given by:

$$\begin{aligned}\Theta(R_i, R_j) &= \frac{1}{3}\{[H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) - H_2(R_i)]^2 \\ &\quad + [M(R_j) - M(R_i)]^2\}\end{aligned} \tag{8}$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ are the hydrophobicity value, hydrophilicity value, and side-chain mass, respectively, of the amino acid R_i. It should be noted that before substituting the values of hydrophobicity, hydrophilicity, and side-chain mass into Eq. 7, they are all subjected to a standard conversion as described by the following equation:

$$H_1(i) = \frac{H_1^0(i) - \sum_{i=1}^{20}\frac{H_1^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[H_1^0(i) - \sum_{i=1}^{20}\frac{H_1^0(i)}{20}\right]^2}{20}}} \tag{9}$$

$$H_2(i) = \frac{H_2^0(i) - \sum_{i=1}^{20}\frac{H_2^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[H_2^0(i) - \sum_{i=1}^{20}\frac{H_2^0(i)}{20}\right]^2}{20}}} \tag{10}$$

$$M(i) = \frac{M^0(i) - \sum_{i=1}^{20}\frac{M^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[M^0(i) - \sum_{i=1}^{20}\frac{M^0(i)}{20}\right]^2}{20}}} \tag{11}$$

where $H_1^0(i)$ and $H_2^0(i)$ represent the original hydrophobicity value and the original hydrophilicity value of the i-th amino acid respectively; and $M^0(i)$ is the mass of the i-th amino acid side chain.

General parallel correlation pseudo amino acid composition (PC-PseAAC_G). The PC-PseAAC_G approach[42], not only incorporates the comprehensive built-in indices extracted from the AAindex[43], but also allows the users to upload their own indices to generate the PC-PseAAC_G feature vector. For a given a protein peptide P (Eq. 1), the PC-PseAAC_G feature vector of P is defined as:

$$\mathrm{V} = [x_1, x_2, \ldots, x_{20}, x_{21}, \ldots, x_{20+\lambda}]^T \tag{12}$$

where

$$\mathrm{x}_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\Theta_j}(1 \le u \le 20) \\ \dfrac{w\Theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\Theta_j}(20 + 1 \le u \le 20 + \lambda) \end{cases} \tag{13}$$

where f_i (i = 1,2, …, 20) is the normalized occurrence frequency of the 20 amino acids in the protein P; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; w is the weight factor ranging from 0 to 1; and Θ_j (j = 1,2,…,20) is called the j-tier correlation factor reflecting the sequence-order correlation between all the j-th most contiguous residues along a protein chain, which is defined as:

$$\Theta_\lambda = \frac{1}{L - \lambda}\sum_{i=1}^{L-\lambda}\Theta(R_i, R_{i+\lambda})(0 < \lambda < L) \tag{14}$$

In this case, the correlation function is given by:

$$\Theta(R_i, R_j) = \frac{1}{\mu}\sum_{u=1}^{\mu}[H_u(R_i) - H_u(R_j)]^2 \tag{15}$$

where μ is the number of physicochemical indices considered; $H_u(R_i)$ is the u-th physicochemical index value of the amino acid R_i; $H_u(R_j)$ is the u-th physicochemical index value for the amino acid R_j. It should be noted that before substituting the physicochemical indices values into Eq. 14, they were also all subjected to a standard conversion as described by the following equation:

$$H_u(i) = \frac{H_u^0(i) - \sum_{i=1}^{20}\frac{H_u^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[H_u^0(i) - \sum_{i=1}^{20}\frac{H_u^0(i)}{20}\right]^2}{20}}} \tag{16}$$

where $H_u^0(i)$ is the u-th original physicochemical value of the i-th amino acid.

Basic kmer (kmer). Basic kmer[44] is the simplest approach to represent the proteins by a numerical vector, in which the protein sequences are represented as the occurrence frequencies of k neighboring amino acids[45]. Given a protein sequence P (Eq. 1), the kmer feature vector of P is formulated as follows:

$$\mathrm{V}(\mathrm{kmer} - 1) = [x_1, x_2, \ldots, x_i, \ldots, x_{20}]^T(0 < i \le 20) \tag{17}$$

$$\mathrm{V}(\mathrm{kmer} - 2) = [y_1, y_2, \ldots, y_i, \ldots, y_{400}]^T(0 < i \le 400) \tag{18}$$

where x_i and y_i are the normalized occurrence frequency of the 20 amino acid residues and 400 dipeptides in the protein P, respectively.

Bi-Profile Bayes (BPB). BPB[26] comprehensively considers the information contained in the two aspects of positive and negative samples that have been successfully applied in many fields of bioinformatics and has made effective predictions[26,46–48]. Given a protein peptide P (Eq. 1), the BPB feature vector of P is defined as:

$$\mathrm{V} = [x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{2n}]^T \tag{19}$$

where V is the posterior probability vector; $x_1, x_2, \ldots, x_n$ represents the posterior probability of each amino acid at each position in positive peptide sequence datasets; $x_{n+1}, \ldots, x_{2n}$ represents the posterior probability of each amino acid at each position in negative peptide sequence datasets. Two position-specific profiles for final model training, positive position-specific profiles and negative position-specific profiles, were generated by calculating the frequency of each amino acid at each position in the positive datasets and negative datasets, respectively.

Double Bi-Profile Bayes (DBPB). DBPB is an improvement of BPB that was proposed by Shao *et al.*[24]. As mentioned above, BPB is the posterior probability of each single amino acid at each position in the positive and negative datasets, while DBPB is the posterior probability of each two adjacent amino acids at each position in the positive and negative datasets. Given a protein sequence P (Eq. 1), the DBPB feature vector of P is defined as:

$$\mathrm{V} = [x_1, x_2, \ldots, x_{n-1}, x_{(n-1)+1}, \ldots, x_{2(n-1)}]^T \tag{20}$$

where V is the posterior probability vector; $x_1, x_2, \ldots, x_{n-1}$ that represents the posterior probability of each dipeptide at each position in positive peptide sequence datasets; $x_{(n-1)+1}, \ldots, x_{2(n-1)}$ represents the posterior probability of each dipeptide at each position in the negative peptide sequence datasets. Two position-specific profiles for the final model training, positive position-specific profile and negative position-specific profile were generated by calculating the frequency of each amino acid pair at each position in the positive datasets and negative datasets, respectively.

Adapted Normal distribution Bi-Profile Bayes (ANBPB). ANBPB[21,49] is the improvement of BPB in another aspect. Given a protein sequence P (Eq. 1), the ANBPB feature vector of P is defined as:

$$\mathrm{V} = [p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n}]^T \tag{21}$$

where $p_1, p_2, \ldots, p_n$ is the posterior probability of each amino acid at each position in positive peptide sequences datasets; $p_{n+1}, \ldots, p_{2n}$ is defined based on the posterior probability of each amino acid at each position in negative peptide sequences datasets. The posterior probability $p_1, p_2, \ldots, p_{2n}$ was coded by the adapted normal distribution as follows:

$$p_i = \varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{dt} \tag{22}$$

where $\varphi(x)$ is the standard normal distribution function and the detailed description of the formula is given[21,49].

Incorporating Amino Acid Pairwise (IAAPair). The posterior probability of every two adjacent amino acids and each two next nearest amino acids at each position in the positive peptide sequence datasets is subtracted from in the negative peptide sequence datasets[19]. Given a protein sequence P (Eq. 1), the IAAPair feature vector of P is defined as:

$$\mathrm{V} = [p_1, p_2, \ldots, p_j, \ldots, p_{[(n-1)+(n-2)]}]^T \tag{23}$$

$$+\mathrm{V} = \begin{bmatrix} +p^0{}_{1,1}, + p^0{}_{1,2}, & \cdots & , + p^0{}_{1,n-1}, & +p^1{}_{1,1}, +p^1{}_{1,2}, & \cdots & , + p^1{}_{1,n-2}, \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ +p^0{}_{21^2,1}, + p^0{}_{21^2,2}, & \cdots & , + p^0{}_{21^2,n-1}, & +p^1{}_{21^2,1}, + p^1{}_{21^2,2}, & \cdots & , + p^1{}_{21^2,n-2}, \end{bmatrix}^T \tag{24}$$

$$-\mathrm{V} = \begin{bmatrix} -p^0{}_{1,1}, - p^0{}_{1,2}, & \cdots & , - p^0{}_{1,n-1}, & -p^1{}_{1,1}, - p^1{}_{1,2}, & \cdots & , - p^1{}_{1,n-2}, \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -p^0{}_{21^2,1}, - p^0{}_{21^2,2}, & \cdots & , - p^0{}_{21^2,n-1}, & -p^1{}_{21^2,1}, - p^1{}_{21^2,2}, & \cdots & , - p^1{}_{21^2,n-2}, \end{bmatrix}^T \tag{25}$$

$$\mathrm{V} = (+\mathrm{V}) - (-\mathrm{V}) \tag{26}$$

$$p_j = \begin{cases} \pm {p^0}_{1,j} & when\ R_tR_{t+1} = AA\ and\ 1 \le j \le n - 1 \\ \pm {p^0}_{2,j} & when\ R_tR_{t+1} = AC\ and\ 1 \le j \le n - 1 \\ \vdots & \qquad\qquad \vdots \\ \pm {p^0}_{(21^2-1),j} & when\ R_tR_{t+1} = XY\ and\ 1 \le j \le n - 1 \\ \pm {p^0}_{21^2,j} & when\ R_tR_{t+1} = XX\ and\ 1 \le j \le n - 1 \\ \pm {p^1}_{1,j} & when\ R_tR_{t+1} = AA\ and\ n \le j \le [(n-1) + (n-2)] \\ \pm {p^1}_{2,j} & when\ R_tR_{t+1} = AC\ and\ n \le j \le [(n-1) + (n-2)] \\ \quad\vdots & \qquad\qquad \vdots \\ \pm {p^1}_{(21^2-1),j} & when\ R_tR_{t+1} = XY\ and\ n \le j \le [(n-1) + (n-2)] \\ \pm {p^1}_{21^2,j} & when\ R_tR_{t+1} = XX\ and\ n \le j \le [(n-1) + (n-2)] \end{cases} \tag{27}$$

where V is the posterior probability vector (in this feature, the C in the middle of peptide sequence must not be omitted). When $1 \le j \le n - 1$ p_j is the representative posterior probability of every two nearest amino acids, and when $n \le j \le [(n-1) + (n-2)]$ p_j is the representative posterior probability of each two next nearest amino acids. $+p_{i,j}$ and $-p_{i,j}$ represent the posterior probability of every two nearest amino acids and each two next nearest amino acids at each position in positive and negative peptide sequence datasets, respectively. $\pm p_{i,j} = (+p_{i,j}) - (-p_{i,j})$ is the feature vector.

Position-specific Tri-Amino Acid Propensity (PSTAAP). The posterior probability of every three adjacent amino acids at each position in the positive peptide sequence datasets is subtracted from in the negative peptide sequence datasets[20,27]. Given a cysteine peptide fragment P (Eq. 1), the feature vector of PSTAAP for P is defined as follows:

$$\mathrm{V} = [p_1, p_2, \ldots, p_j, \ldots, p_{n-2}]^T \tag{28}$$

$$+\mathrm{V} = \begin{bmatrix} +p_{1,1}, +p_{1,2}, & \cdots & , +p_{1,n-2} \\ \vdots & \ddots & \vdots \\ +p_{21^3,1}, +p_{21^3,2}, & \cdots & +p_{21^3,n-2} \end{bmatrix}^T \tag{29}$$

$$-\mathrm{V} = \begin{bmatrix} -p_{1,1}, -p_{1,2}, & \cdots & , -p_{1,n-2} \\ \vdots & \ddots & \vdots \\ -p_{21^3,1}, -p_{21^3,2}, & \cdots & , -p_{21^3,n-2} \end{bmatrix}^{T^T} \tag{30}$$

$$\mathrm{V} = (+\mathrm{V}) - (-\mathrm{V}) \tag{31}$$

$$p_j = \begin{cases} \pm p_{1,j} & when\ R_t = AAA \\ \pm p_{2,j} & when\ R_t = AAC \\ \vdots & \qquad\qquad \vdots \\ \pm p_{(21^3-1),j} & when\ R_t = XXY \\ \pm p_{21^3,j} & when\ R_t = XXX \end{cases} \tag{32}$$

where V is the posterior probability vector; $+p_{i,j}$ represents the posterior probability of each tri-amino acids at each position in the positive dataset; $-p_{i,j}$ represents the posterior probability of each tri-amino acids at each position in the negative dataset; $\pm p_j = (+p_{i,j}) - (-p_{i,j})$ is the feature vector.

It should be indicated that recently a very powerful web-server called 'Pse-in-One'[25], and its updated version 'Pse-in-One2.0'[45] have been established and can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the user study needs or desires. In the current study, the feature vectors PC-PseAAC, PC-PseAAC_General, and basic kmer are obtained from the web-server.

Information Gain (IG). The IG[44,46–48] method is usually used to rank the importance of positions and amino acid residues. IG measures the decrease in entropy when a given feature is used to group values of another (class) feature. The entropy of a feature X is defined by:

where $\{x_i\}$ is a set of values of X and $P(x_i)$ is the prior probability of x_i. If Y is considered as another feature, the conditional entropy of X is defined as:

$$\mathrm{H(X|Y)} = -\sum_{j} P(y_j) \sum_{i} P(x_i|y_j)\, log_2\left((x_i|y_j)\right) \tag{33}$$

where $P(x_i|y_j)$ is the posterior probability of X with the value y_j of Y. The amount by which the entropy of X decreases reflects the additional information about X provided by Y and is called the information gain:

$$\mathrm{IG(X|Y)} = \mathrm{H(X)} - \mathrm{H(X|Y)} \tag{34}$$

According to this measure, Y has a stronger correlation with X than with Z, if IG (X|Y) > IG (Z |Y). It is obvious that Y represents the amino acid type, when extracting the IG score for positions. On the other hand, Y represents the amino acid frequency, when extracting IG score for amino acids.

Calculating IG score of positions and amino acid residues:

(1) The importance of positions: The 20 amino acid residues (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y) were coded into digits from 1 to 20. The query sequences segments were coded into an X-dimension digital sequence (protein).
(2) The importance of amino acid residues: The amino acid frequency in the surrounding sequence query site (the site itself is not counted) was calculated. The query sequences were also coded into a 20-dimension feature.
(3) Calculation of the IG score for positions by (1) and the IG procedure was performed. Subsequently, the calculation of the IG score for amino acid residues by (2) and the IG procedure was done. Then, we ranked the corresponding positions and amino acid residues by their IG score and selected the key positions and key amino acid residues.

In this work, we used the IG score to calculate the importance of amino acid residues:

$$\begin{aligned} \mathrm{IG}(\mathrm{X}|\mathrm{Y}_i) &= \mathrm{H(X)} - \mathrm{H}(\mathrm{X}|\mathrm{Y}_i) \\ &= -\sum_{x\in(0,1)} P(x)\, log_2((x)) \sum_{y_j\in(0,1)} P(y_i) \\ &\quad \times \sum_{x\in(0,1)} P(x|y_j)\, log_2((x_i|y_j)) \end{aligned} \tag{35}$$

Equation (36) is divided into two parts, the former is the entropy H(X) of class X, and the latter part is the conditional entropy $\mathrm{H}(\mathrm{X}|\mathrm{Y}_i)$ of X a given amino acid Y.

Suppose the number of training samples is N. Initially, we count each training sample. Subsequently, if each characteristic y_j is added in the training sample x, two times will be counted once:

$$P(y_j = 1) = \frac{count(y_j = 1)}{N} \tag{36}$$

$$P(y_j = 0) = 1 - P(y_j = 1) \tag{37}$$

$$P(X = 0|y_j = 1) = \frac{count(X = 0,\, y_j = 1)}{count(y_j = 1)} \tag{38}$$

$$P(X = 1|y_j = 1) = \frac{count(X = 1,\, y_j = 1)}{count(y_j = 1)} \tag{39}$$

$$P(X = 0|y_j = 0) = \frac{count(X = 0,\, y_j = 0)}{count(y_j = 0)} \tag{40}$$

$$P(X = 1|y_j = 0) = \frac{count(X = 1,\, y_j = 0)}{count(y_j = 0)}. \tag{41}$$

Max-Relevance-Max-Distance (MRMD). MDMR[31] is a feature selection method for reducing dimensionalities, which can be further divided into two aspects.

(1) One is the relevance between sub-feature set and target class. Here, Pearson's correlation coefficient is exploited to measure the relevance. With the increase of Pearson's correlation coefficient, the relevant between feature and target class also increases.
(2) The other is redundancy of sub-feature set. Three kinds of distance functions are utilized to calculate the redundancy. The larger the feature distance, the lower the redundancy for sub-feature set becomes.

The features with large sum of relevance and distance would be chosen as the ultimate sub-feature set. Finally, the sub-feature set generated by MRMD has low redundancy and strong relevance with the target class.

In order to describe the algorithm clearly, we listed some functions in following section. Given the input datasets tabled as N instances, M features $F = \{f_i, i = 1, \ldots, M\}$ and the target class C, the aim is to find a subspace of M features, which is selected from the M dipeptides original space, and makes the greatest contribution to classify the target class C.

Max-relevance (MR). Making the greatest contribution for classifying the target class condition and this often requires the maximal relevance for the target class C on the subspace, which needs us to select a feature set with the highest relevance to target class C. We use the Pearson's correlation coefficient to measure positive correlation and negative correlation. Because it is suitable for calculating continuous variables and easy to implement, Pearson's correlation coefficient is adopted as the measure of relevance between feature and target class C.

The value of MR for feature i can be defined as follows.

$$\mathrm{mac\,MR}_i = \left|\mathrm{PCC}\left(\underset{F_i}{\rightarrow}\underset{C_i}{\rightarrow}\right)\right|(1 \le i \le M) \tag{42}$$

where $\underset{F_i}{\rightarrow}$ is a vector composed from ith features from each instance, and $\underset{C_i}{\rightarrow}$ is also a vector whose every element comes from the target class C of each instance. Their Pearson's correlation coefficient is defined as $\mathrm{PCC}\left(\underset{F_i}{\rightarrow}\underset{C_i}{\rightarrow}\right)$.

Max-Distance (MD). MDMR proposed a new approach to realize Max-Redundancy based on distance function, namely maximal distance, to measure the level of similarity between two feature vectors. There are three types of distance functions that can be chosen, which are Euclidean distance, cosine similarity and Tanimoto coefficient. Compared with the commonly used methods, Euclidean distance is easier to calculate. As compared to the Euclidean distance, cosine similarity focuses on the angle between two vectors. The last one, Tanimoto coefficient, is also called Jaccard coefficient in the broad sense. Under the binary condition, it is similar to Jaccard coefficient. For each feature, its value of distance defined as follows is based on t three types of distance functions mentioned above. According to the following formula, we can obtain their values for the feature i $(1 \le i \le M)$ ED_i, COS_i and TC_i, respectively.

$$\mathrm{ED}_i = \frac{1}{M-1}\sum ED\left(\underset{F_i}{\rightarrow}\underset{F_k}{\rightarrow}\right)(1 \le k \le M, k \ne i) \tag{43}$$

$$\mathrm{COS}_i = \frac{1}{M-1}\sum COS\left(\underset{F_i}{\rightarrow}\underset{F_k}{\rightarrow}\right)(1 \le k \le M, k \ne i) \tag{44}$$

$$\mathrm{TD}_i = \frac{1}{M-1}\sum TD\left(\underset{F_i}{\rightarrow}\underset{F_k}{\rightarrow}\right)(1 \le k \le M, k \ne i) \tag{45}$$

From three formulas above, we have four ways to obtain the final value of MD.

$$max\ \mathrm{MD}_i = \mathrm{ED}_i(1 \le i \le M) \tag{46}$$

$$max\ \mathrm{MD}_i = \mathrm{COS}_i(1 \le i \le M) \tag{47}$$

$$max\ \mathrm{MD}_i = \mathrm{TC}_i(1 \le i \le M) \tag{48}$$

$$mean\ \mathrm{MD}_i = \frac{1}{3}(\mathrm{ED}_i + \mathrm{COS}_i + \mathrm{TC}_i)(1 \le i \le M) \tag{49}$$

We can obtain top m features which are considered to be the sub-feature set with minimal redundancy by MD.

MRMD. The criterion used for combining the two constraints above is called "Max-Relevance-Max-Distance" (MRMD). After having done all the above preparations, we could start to select the features subspace. The algorithm optimizes the following condition.

For a specific problem, the condition for feature selection take into consideration that the MR is not as important as MD. Therefore, the variables w_r $(1 \le w_r \le M)$ and w_d $(1 \le w_d \le M)$ are the weights of MR and MD, respectively.

References

1. Foster, M. W., Hess, D. T. & Stamler, J. S. Protein S-nitrosylation in health and disease: a current perspective. *Trends Mol Med* **15**, 391–404, https://doi.org/10.1016/j.molmed.2009.06.007 (2009).
2. Xue, Y. *et al.* GPS-SNO: Computational Prediction of Protein S-Nitrosylation Sites with a Modified GPS Algorithm. *PloS one* **5**, ARTN e11290, https://doi.org/10.1371/journal.pone.0011290 (2010).
3. Lim, K. H., Ancrile, B. B., Kashatus, D. F. & Counter, C. M. Tumour maintenance is mediated by eNOS. *Nature* **452**, 646–U611, https://doi.org/10.1038/nature06778 (2008).

4. Li, F. *et al*. Regulation of HIF-1 alpha stability through S-nitrosylation. *Mol Cell* **26**, 63–74, https://doi.org/10.1016/j.molcel.2007.02.024 (2007).
5. Burgoyne, J. R. & Eaton, P. A Rapid Approach for the Detection, Quantification, And Discovery Of Novel Sulfenic Acid Or S-Nitrosothiol Modified Proteins Using a Biotin-Switch Method. *Method Enzymol* **473**, 281–303, https://doi.org/10.1016/S0076-6879(10)73015-9 (2010).
6. Jaffrey, S. R., Erdjument-Bromage, H., Ferris, C. D., Tempst, P. & Snyder, S. H. Protein S-nitrosylation: a physiological signal for neuronal nitric oxide. *Nat Cell Biol* **3**, 193–197, https://doi.org/10.1038/35055104 (2001).
7. Gross, S. S. *et al*. SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures. *Nitric Oxide-Biol Ch* **14**, A43–A44, https://doi.org/10.1016/j.niox.2006.04.148 (2006).
8. Greco, T. M. *et al*. Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. *P Natl Acad Sci USA* **103**, 7420–7425, https://doi.org/10.1073/pnas.0600729103 (2006).
9. Derakhshan, B., Wille, P. C. & Gross, S. S. Unbiased identification of cysteine S-nitrosylation sites on proteins. *Nat Protoc* **2**, 1685–1691, https://doi.org/10.1038/nprot.2007.210 (2007).
10. Forrester, M. T. *et al*. Proteomic analysis of S-nitrosylation and denitrosylation by resin-assisted capture. *Nat Biotechnol* **27**, 557–559, https://doi.org/10.1038/nbt.1545 (2009).
11. Hess, D. T., Matsumoto, A., Nudelman, R. & Stamler, J. S. S-nitrosylation: spectrum and specificity. *Nat Cell Biol* **3**, E46–E49, https://doi.org/10.1038/35055152 (2001).
12. Lindermayr, C., Saalbach, G. & Durner, J. Proteomic identification of S-nitrosylated proteins in Arabidopsis thaliana. *Comp Biochem Phys A* **141**, S241–S241 (2005).
13. Kuncewicz, T., Sheta, E. A., Goldknopf, I. L. & Kone, B. C. Proteomic analysis of S-nitrosylated proteins in mesangial cells. *Mol Cell Proteomics* **2**, 156–163, https://doi.org/10.1074/mcp.M300003-MCP200 (2003).
14. Huang, B., Chen, S. C. & Wang, D. L. Shear flow increases S-nitrosylation of proteins in endothelial cells. *Cardiovasc Res* **83**, 536–546, https://doi.org/10.1093/cvr/cvp154 (2009).
15. Lefievre, L. *et al*. Human spermatozoa contain multiple targets for protein S-nitrosylation: An alternative mechanism of the modulation of sperm function by nitric oxide? *Proteomics* **7**, 3066–3084, https://doi.org/10.1002/pmic.200700254 (2007).
16. Foster, M. W., Forrester, M. T. & Stamler, J. S. A protein microarray-based analysis of S-nitrosylation. *P Natl Acad Sci USA* **106**, 18948–18953, https://doi.org/10.1073/pnas.0900729106 (2009).
17. Chen, W., Yang, H., Feng, P., Ding, H. & Lin, H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **33**, 3518–3523, https://doi.org/10.1093/bioinformatics/btx479 (2017).
18. Li, Y. X., Shao, Y. H., Jing, L. & Deng, N. Y. An Efficient Support Vector Machine Approach for Identifying Protein S-Nitrosylation Sites. *Protein Peptide Lett* **18**, 573–587, https://doi.org/10.2174/092986611795222731 (2011).
19. Xu, Y., Shao, X. J., Wu, L. Y., Deng, N. Y. & Chou, K. C. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *Peerj* **1**, UNSP e171, https://doi.org/10.7717/peerj.171 (2013).
20. 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* **8**, ARTN e55844, https://doi.org/10.1371/journal.pone.0055844 (2013).
21. Jia, C. Z., Lin, X. & Wang, Z. P. Prediction of Protein S-Nitrosylation Sites Based on Adapted Normal Distribution Bi-Profile Bayes and Chou's Pseudo Amino Acid Composition. *Int J Mol Sci* **15**, 10410–10423, https://doi.org/10.3390/ijms150610410 (2014).
22. Zhang, J., Zhao, X. W., Sun, P. P. & Ma, Z. Q. PSNO: Predicting Cysteine S-Nitrosylation Sites by Incorporating Various Sequence-Derived Features into the General Form of Chou's PseAAC. *Int J Mol Sci* **15**, 11204–11219, https://doi.org/10.3390/ijms150711204 (2014).
23. Lin, H. *et al*. The prediction of protein structural class using averaged chemical shifts. *J Biomol Struct Dyn* **29**, 643–649, https://doi.org/10.1080/07391102.2011.672628 (2012).
24. 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 Proteomics* **75**, 1654–1665, https://doi.org/10.1016/j.jprot.2011.12.003 (2012).
25. Liu, B. *et al*. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* **43**, W65–W71, https://doi.org/10.1093/nar/gkv458 (2015).
26. Shao, J. L., Xu, D., Tsai, S. N., Wang, Y. F. & Ngai, S. M. Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction. *PloS one* **4**, ARTN e4920, https://doi.org/10.1371/journal.pone.0004920 (2009).
27. Xu, Y., Wen, X., Shao, X. J., Deng, N. Y. & Chou, K. C. iHyd-PseAAC: Predicting Hydroxyproline and Hydroxylysine in Proteins by Incorporating Dipeptide Position-Specific Propensity into Pseudo Amino Acid Composition. *Int J Mol Sci* **15**, 7594–7610, https://doi.org/10.3390/ijms15057594 (2014).
28. Chen, X. X. *et al*. Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *BioMed research international* **2016**, 1654623, https://doi.org/10.1155/2016/1654623 (2016).
29. He, W., Jia, C., Duan, Y. & Zou, Q. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC systems biology* **12**, 44, https://doi.org/10.1186/s12918-018-0570-1 (2018).
30. Jia, C., Yang, Q. & Zou, Q. NucPosPred: Predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. *Journal of theoretical biology* **450**, 15–21, https://doi.org/10.1016/j.jtbi.2018.04.025 (2018).
31. Zou, Q., Zeng, J. C., Cao, L. J. & Ji, R. R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **173**, 346–354 (2016).
32. 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* **6**, ARTN e21849, https://doi.org/10.1371/journal.pone.0021849 (2011).
33. Cui, T. Y. *et al*. MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res* **46**, D371–D374, https://doi.org/10.1093/nar/gkx1025 (2018).
34. Jiang, J., Xing, F., Zeng, X. X. & Zou, Q. RicyerDB: A Database For Collecting Rice Yield-related Genes with Biological Analysis. *Int J Biol Sci* **14**, 965–970, https://doi.org/10.7150/ijbs.23328 (2018).
35. Liang, Z. Y. *et al*. Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* **33**, 467–469, https://doi.org/10.1093/bioinformatics/btw630 (2017).
36. Yi, Y. *et al*. RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res* **45**, D115–D118, https://doi.org/10.1093/nar/gkw1052 (2017).
37. Zhang, T. *et al*. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res* **45**, D135–D138, https://doi.org/10.1093/nar/gkw728 (2017).
38. Manavalan, B., Shin, T. H. & Lee, G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front Microbiol* **9**, Artn 476, https://doi.org/10.3389/Fmicb.2018.00476 (2018).
39. Cheng, J. H. *et al*. Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr Intell Lab* **180**, 64–69, https://doi.org/10.1016/j.chemolab.2018.07.006 (2018).
40. He, W. Y., Ju, Y., Zeng, X. X., Liu, X. R. & Zou, Q. Sc-ncDNAPred: A Sequence-Based Predictor for Identifying Non-coding DNA in Saccharomyces cerevisiae. *Front Microbiol* **9**, Artn 2174, https://doi.org/10.3389/Fmicb.2018.02174 (2018).
41. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43**, 246–255, https://doi.org/10.1002/Prot.1035 (2001).

42. Cao, D. S., Xu, Q. S. & Liang, Y. Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* **29**, 960–962, https://doi.org/10.1093/bioinformatics/btt072 (2013).
43. Kawashima, S. & Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res* **28**, 374–374, https://doi.org/10.1093/Nar/28.1.374 (2000).
44. Liu, B., Wang, X. L., Lin, L., Dong, Q. W. & Wang, X. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *Bmc Bioinformatics* **9**, Artn 510, https://doi.org/10.1186/1471-2105-9-510 (2008).
45. Liu, B., Wu, H. & Chou, K.-C. Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Natural Science* **09**, 67–91, https://doi.org/10.4236/ns.2017.94007 (2017).
46. Zhao, X. W. *et al*. Identification of Protein Pupylation Sites Using Bi-Profile Bayes Feature Extraction and Ensemble Learning. *Math Probl Eng*, Artn 283129, https://doi.org/10.1155/2013/283129 (2013).
47. Jia, C. Z., Liu, T. A., Chang, A. K. & Zhai, Y. Y. Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. *Biochimie* **93**, 778–782, https://doi.org/10.1016/j.biochi.2011.01.013 (2011).
48. Jia, C. Z., He, W. Y. & Yao, Y. H. OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids. *J Biomol Struct Dyn* **35**, 829–835, https://doi.org/10.1080/07391102.2016.1163294 (2017).
49. Jia, C. Z., Liu, T. & Wang, Z. P. O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Mol Biosyst* **9**, 2909–2913, https://doi.org/10.1039/c3mb70326f (2013).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant (number 71271034), the National Social Science Foundation of China under grant (15CGL031), the Fundamental Research Funds for the Central Universities under grant (number 3132016306, 3132018160), the Program for Dalian High Level Talent Innovation Support under grant (2015R063), and the National Scholarship Fund of China for Studying Abroad.

Author Contributions

T.-Y.L. conceived and designed the experiments; R.-Y.S. implemented IG method to optimize features and designed programs; Q.Y. carried out feature extraction; M.-Y.G. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-39743-9.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.