Prediction of S-nitrosylation Sites by Integrating Support Vector Machine and Random Forest

doi: https://doi.org/10.1039/C9MO00098D

Prediction of S-nitrosylation Sites by Integrating Support Vector Machine and Random Forest

Md. Mehedi Hasan[a], Balachandran Manavalan[b], Mst. Shamima Khatun[a], and Hiroyuki Kurata[a,c*]

[a]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan.
[b]Department of Physiology, Ajou University School of Medicine, Suwon 443380, Korea.
[c]Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan.

*Corresponding author
kurata@bio.kyutech.ac.jp

Molecular OMICS
https://doi.org/10.1039/C9MO00098D

Abstract

Cysteine S-nitrosylation is a type of reversible post-translational modification of proteins, which controls diverse biological processes. It is associated with redox-based cellular signaling to protect against oxidative stress. The identification of S-nitrosylation sites is an important step to reveal the function of proteins; however, experimental identification of S-nitrosylation is expensive and time-consuming work. Hence, sequence-based computational prediction of potential S-nitrosylation sites is highly sought before experimentation. Herein, a novel predictor PreSNO has been developed that integrates multiple encoding schemes by the support vector machine and random forest algorithms. The PreSNO achieved an accuracy and Matthews correlation coefficient value of 0.752 and 0.252 respectively in classifying between SNO and non-SNO sites when evaluated on the independent dataset, outperforming the existing methods. The web application of the PreSNO and its associated datasets are freely available at http://kurata14.bio.kyutech.ac.jp/PreSNO/.

Keywords: Cysteine S-nitrosylation, Support vector machine, Random forest, Feature encoding, Feature selection

Introduction

S-nitrosylation (SNO) is a type of reversible post-translational modification (PTM) of proteins that play a key role in regulating many cellular functions[1, 2]. In the SNO process, a thiol group of cysteine residues is covalently attached by recycling nitric oxide [3-8]. Different studies suggest that SNO on cysteine is critically responsible for redox pathways, cardiovascular, immune, and neuronal systems [9-12] and affects various pathophysiological events such as cancers and diabetes[13-19]. Detailed mechanisms of SNO remain to be elucidated, due to the low abundance and labile nature of SNO. Therefore, identification of SNO sites is essential for an understanding of both the pathological and physiological mechanisms as well as the basic design of drugs.

To identify the SNO sites of proteins by using the molecular signature, large-scale proteomic experimental works have been accomplished [2, 20-22]. Notwithstanding the increasing number of experimentally determined SNO proteins, the explicit identification of SNO sites remains challenging. In particular, large-scale experimental screenings of SNO sites are time-consuming and laborious works. As an alternative to experimental efforts, the computational methodology can serve to provide a potential proteome-wide identification of SNO sites.

To date, a few computational models, e.g., GPS-SNO [23], SNOSite [24], and iSNOPseAAC [25], have been developed to predict the SNO sites. The GPS-SNO used their Group-based Prediction System (GPS) algorithm with the encoding schemes including matrix transformation, weight training, and motif selection, and it was trained by using 504 SNO sites of 327 proteins. The SNOSite used the maximal dependence decomposition via support vector machine (SVM), trained by 586 SNO sites of 384 proteins. The iSNOPseAAC implemented a Conditional Random Field (CRF) algorithm with the encoding scheme of the pseudo amino acid composition, trained by using 731 SNO sites of 438 proteins. Recently, DeepNitro has been developed that employed a deep learning algorithm with the encoding schemes of the composition of amino acid pairs and position-specific scoring matrix (PSSM) [26]. Existing predictors still remain to be improved. First, since the existing predictors of GPS-SNO [23], SNOSite [24], and iSNOPseAAC [25] used a small training dataset, they provided poor predictions when evaluated with the independent dataset. Second, although feature extraction and selection are critically important for machine learning (ML)-based algorithms, the existing algorithms used only position-wise encoding methods, which were unable to fully characterize the potential SNO sites. Third, since the SNOSite and iSNOPseAAC predictors do not provide the probability scores of SNO sites, users cannot understand the stringency of prediction. Finally, most of the methods used the old versions of datasets, which include many false-negative samples that are now verified experimentally as the positive samples.

In this work, to overcome those problems, we have developed a novel predictor PreSNO (Prediction of S-nitrosylation sites) using the latest, large-scale dataset by integrating two different classifiers of the support vector machine (SVM) and random forest (RF), as shown in **Figure 1**. In particular, we combined different established encoding schemes, namely, the composition of profile-based amino acid pair (CPA)), the *k*-space spectral amino acid composition (SAC), tripeptide composition from the PSSM (TCP), and physicochemical properties of amino acids (PPA). These four encodings were inputted separately into SVM and RF. Finally, all these models were integrated via a linear regression (LR) model to calculate the probability score of S-nitrosylation at each cysteine residue. To construct the PreSNO and assess its prediction performance, 5-fold cross-validation (CV) was carried on the training dataset, and the prediction was executed on the independent data. The PreSNO outperformed other existing prediction models. Additionally, we employed two other combination methods of the sequential combinational model and meta-classifier, to demonstrate the advantage of the LR-based combination employed by the proposed PreSNO.

Materials and methods

Dataset

Recently, Xie et al. have constructed a high-quality dataset based on extensive literature search and previously reported datasets[26], where the positive samples are experimentally confirmed as S-nitrosylation sites, to develop the DeepNitro predictor. Any SNO sites other than experimentally confirmed SNO sites were defined as the negative samples. This procedure is commonly employed to generate negative samples [27-29], although erroneous data may deteriorate the prediction performance.

In this study, we utilized the DeepNitro dataset that encompassed 3,113 unique proteins with 4,762 SNO sites. To avoid the overestimation of the prediction model, we filtered the protein sequences with an identity cut-off of 30% by using CD-HIT [30], signifying that the sequence identity was >30% in these cases. In general, a decrease in the sequence identity cut-off is able to avoid overfitting risks caused by redundant samples containing many homologous sites[26], while decreasing the number of available samples. Since our dataset was large, we used a low sequence identity cut-off (30%). After CD-HIT elimination, we obtained 3,734 positive and 20, 548 negative samples. The experimentally verified SNO sites were considered as the positive samples (SNO sites), whereas the remaining cysteine residues, which had not been experimentally verified as SNO sites, were considered as the negative samples (non-SNO sites). Subsequently, each sequence window with length 2*w*+1, having cysteine residue (C) at the center, was characterized, where *w* is the number of residues. We eliminated the identical window sequence (i.e., if the given SNO or non-SNO sites share an identical flanking sequence, the negative one is deleted) [26]. Finally, we obtained 3,734 positive and 20,333 negative samples. From these samples, we randomly selected 20% as the independent dataset (351 SNO sites with 3,168 non-SNO sites), while the remaining samples of 3,383 SNO sites and 17,165 non-SNO sites were considered as the training dataset. Generally, the prediction accuracy is often impaired by an unbalanced ratio of positive to negative samples in the training data[31-33]. To solve the potentially biased prediction, the non-SNO fragment sequences were randomly pooled from the entire non-SNO samples to keep a ratio of SNO to non-SNO sites at 1:1. All of the curated training and independent datasets are available in our web server.

Feature vectors

To encode the SNO and non-SNO sequences, four encoding schemes of the CPA, SAC, TCP, and PPA were used. Each of the encoding schemes is summarized as follows.

CPA encoding

The CPA encoding was developed from the PSSM profile [31, 32, 34]. In brief, the PSSM was generated from the Swiss-Prot (December 2010) database by using PSI-BLAST (version 2.2.26+) with two constraints: iteration times and e-value of 3 and 1.0×10^{-3}, respectively. Then, we generated potential *k*-space composition of the profile-based amino acids, i.e., CPA, in the same way as the previous study on pupylation site prediction [32]. For a window sequence, a 2,205-dimensional feature vector was generated by the CPA encoding. The limitation of PSSM is that it requires a long computational time to generate profile information for a given sequence.

SAC encoding

We calculated the *k*-space spectrum composition of amino acids, SAC, to measure the sequence context of SNO or non-SNO sites. We scanned the whole curated sliding window with length 2w+1 and counted all the potential numbers of amino acid pairs. 441 (=21×21) pairs of amino acids (including the null residue (-)) are generated for a single *k*-space (i.e., AA, AC, AD, …, --). 21×(k_{max}+1)×21 residue pairs are generated when *k* signifies the space between two residues. In this study, we set k_{max} to 4. The SAC encoding is widely used in computational biology research[32, 35, 36].

TCP encoding

The TCP is a novel encoding scheme generated from the PSSM profile. After generating the PSSM by using PSI-BLAST, we calculated a score with respect to each component of three residues from the PSSM. In brief, for a positive or negative sequence with 21 amino acid residues (including null residues), the TCP scheme provided a 9,261 (=21×21×21)-dimensional feature vector for an SNO or non-SNO site. The score value of each tripeptide (q_i, q_j, q_k, where i, j, k = 1, 2, …, 21) were calculated and normalized as follows:

$$V_{i,j,k}(N) = \frac{\max\{min\{\mathrm{PSMM}(t, q_i), \mathrm{PSSM}(t+1, q_j), \mathrm{PSSM}(t+2, q_k)\}, 0\}}{2w} \qquad (1)$$

where *N* is the index of the curated tripeptide (*N*=1, 2, …, 9,261) and *t* is the row position of the first residue of each curated tripeptide in the PSSM. The PSSM (t, q_i) is the score of amino acid residue q_i at the position of t^{th} row. The PSSM (t+1, q_j) and PSSM (t+2, q_k) stand for the scores of residue (t+1)th and (t+2)th row positions, respectively.

PPA encoding

The PPA database (version 9.1) includes the various mathematical indices of physicochemical properties of amino acids [37] and is widely used for protein and peptide prediction [38-43]. We used 15 types of informative amino acid indices to encode SNO and non-SNO samples (**Table S1**). At w=20, a 615 (=(2×20+1)×15)-dimensional feature vector was obtained for an SNO or non-SNO site through the PPA encoding scheme.

Machine learning model

The SVM and RF algorithms were employed to classify the SNO and non-SNO sites. Both algorithms have been extensively used to predict binary class samples [24, 32, 44-46]. To minimize the classification error, the SVM aims to find the optimal hyperplane to accurately classify samples based on the consecutive features of the training dataset. For numerical calculations, the provided sequences were converted into the representative feature vectors with fixed length and the class labels of the SNO site and non-SNO site are set to 1 and 0, respectively. We used an SVM^{light} package function with default parameters at http://svmlight.joachims.org/ [36, 47, 48].

The RF is an ensemble learning of ML algorithms [49]. In brief, the RF consists of N individual decision trees, $T = \{T_1, T_2, \ldots, T_N\}$. The RF generates new training datasets for N trees by utilizing the bootstrap sampling and then assigns M features to each node of the trees to give the best split according to the Gini impurity. To improve the prediction performance, the RF scores were combined as a weighted sum. An R package was employed (https://cran.r-project.org/web/packages/randomForest/) with the default of 1,000 trees to estimate the performance.

Feature optimization

There are several feature ranking procedures[32, 39, 40, 48], including mRMR, Chi-square, and Wilcoxon rank-sum (WR) test. In this study, the WR test was employed. According to the relevance to the redundancy between the features, the WR test can rank all the features themselves [50, 51].

Performance evaluation

To evaluate the prediction performance of the PreSNO, commonly used four threshold-dependent yardstick measures were applied[52, 53]: accuracy (AC), sensitivity (SN), specificity (SP), and Matthews' correlation coefficient (MCC) defined by:

$$\mathrm{AC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{2}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})}} \quad (5)$$

where TP, TN, FP, and FN illustrate the numbers of true positive (i.e., accurately predicted as SNO), true negative (i.e., accurately predicted as non-SNO), false positive (i.e., wrongly predicted as SNO), and false negative (wrongly predicted as non-SNO), respectively. As the threshold-independent measure, the area under the ROC curves (AUC) was used.

Combined model

LR-based combination model

To enhance the performance of the PreSNO, we combined the SVM and RF probability scores via the LR model[54]. The SVM and RF scores provided by each encoding scheme of CPA, SAC, TCP, and PPA were linearly combined as follows,

$$SVM_{com} = \mathrm{CPA} \times w_1 + \mathrm{SAC} \times w_2 + \mathrm{TCP} \times w_3 + \mathrm{PPA} \times w_4 \quad (6)$$

$$\mathrm{RF}_{com} = \mathrm{CPA} \times w_5 + \mathrm{SAC} \times w_6 + \mathrm{TCP} \times w_7 + \mathrm{PPA} \times w_8 \quad (7)$$

Furthermore, both the scores of the $\mathrm{SVM}_{\mathrm{com}}$ and $\mathrm{RF}_{\mathrm{com}}$ models were linearly combined as follows.

$$\mathrm{PreSNO} = \mathrm{SVM}_{com} \times w_9 + \mathrm{RF}_{com} \times w_{10} \quad (8)$$

where w_1, w_2, w_3, w_4, w_5, w_7, w_8, w_9, and w_{10} are the weight coefficients. The sum of the weight coefficients for each combined model is 1. Each weight coefficient was adjusted between 0 and 1 with an interval of 0.05.

Sequential combination model

To construct a sequential combination model, we combined the four encoding feature vectors of the CPA, SAC, TCP, and PPA in a row, as follows:

$$F_{com} = F(\mathrm{CPA}).F(\mathrm{SAC}).F(\mathrm{TCP}).F(\mathrm{PPA}) \quad (9)$$

where F_{com} is the combined feature vector and F(.) represents each encoding feature vector. The total dimension of F_{com} was 14,286.

Meta-classifier

To construct the meta-classifier for sequence *S*, many probability scores were estimated by implementing different encoding schemes and then they were combined as the new feature vector defined by:

$$P_{com} = P(C(1), \mathrm{En}(1)), \ldots P(C(i), \mathrm{En}(j)), \ldots, P(C(n), \mathrm{En}(m)) \qquad (10)$$

where P_{com} is the new feature vector, $P(C(i), \mathrm{En}(j))$ the prediction probability by each classifier $C(i)$ with encoding scheme En(j), i the index of the classifier, j the index of the encoding scheme, n the number of classifiers and m the number of encoding methods. Finally, S was classified by MLs based on the new feature vectors. In this study, we used 4 encoding schemes for S and two ML algorithms, which generated 8 types of probability scores. The feature vectors consisting of 8 probability scores were used to train the SVM and RF models.

PreSNO web server

The web application programs of the PreSNO were written in PERL, R, HTML, PHP, and CGI scripts. After submitting a query protein, the web application returns the prediction result within several minutes. The final output webpage provides the query sequence name, all predicted cysteine site positions, and probability scores of the predicted SNO sites, together with a job ID, like "2019012100011". Users can save this ID on behalf of the future query for a month.

Results and Discussion

Analysis of SNO and non-SNO sites sequence

We scrutinized the amino acid residue preference of the window sequences of the SNO and non-SNO samples by a two-sample pLogo [55]. As shown in **Figure 2**, over- and under-represented residues (SNO and non-SNO samples) for a given window sequence (p <0.05) are displayed at each position above and below the X-axis, respectively. The height of the logos is in proportion to their corresponding amino acid occurrence frequency of SNO or non-SNO samples. The cumulative percentages of over- or under-represented amino acids are displayed on the Y-axis. A substantial dissimilarity in the window sequences was found between SNO and non-SNO samples. Particularly, in the sequences having the SNO site, the charged residues of the aspartic acid (D), glutamic acid (E), lysine (K), and arginine (R) were enriched. For the sequences having a non-SNO site, neutral amino acids of cysteine (C), and tryptophan (W) were seen. These results demonstrated distinct position-specific sequence preferences between the SNO and non-SNO sites, suggesting that position-specific amino acids are effective in identifying the SNO sites.

Interestingly, in the sequences with a non-SNO site neutral amino acids including "C" and "W" were frequently observed. On the other hand, charges residues of E", "D", "K", and "R were often found around the SNO site. These charged residues would be more exposed to solvent rather than neutral residues, increasing the accessibility to the SNO sites. They are also responsible for salt bridges to facilitate two non-covalent interactions: hydrogen bonding and ionic bonding, which may promote nitrosylation.

Optimization of dataset ratio and window size

Generally, the use of an unbalanced ratio of positive to negative samples, employed for training the ML model, deteriorates the prediction performance [33]. To solve this issue, a well-adjusted ratio of the positive to negative samples have been considered by many PTM site prediction studies [31, 34, 51]. In this study, we used different ratios of SNO to non-SNO samples to train the SVM and RF classifiers implementing each scheme of CPA, SAC, TCP, and PPA (**Figure S1**). The performance was evaluated by a 5-fold CV test on the training dataset. In both the SVM and RF algorithms, a ratio of 1:1 achieved higher performance than any other ratios (**Figure S1**).

To distinguish the SNO from non-SNO sites, the window size is an essential factor that affects the prediction performance. Hence, we optimized the window size in a range from 13 to 45 using four different encodings (CPA, SAC, TCP, and PPA) and two different classifiers (SVM and RF) by a 5-fold CV test on the training dataset. Figure S2 shows that the AUC of RF and SVM peaked at size 41. Therefore, we used the optimal sequence length of 41 for the subsequent analysis (model construction).

Construction and evaluation of PreSNO

The training dataset is transformed into feature vectors by using four encoding schemes (CPA, SAC, TCP, and PPA) and individually inputted to SVM and RF. Particularly, we selected 350 and 330 critical TCP-encoding features (identified by WR test) and inputted them to SVM and RF, respectively. On the other hand, the CPA, SAC, and PPA encoding features were used as such (without applying feature selection). The resultant prediction models were verified using 5-fold CV.

Table 1 shows that the CPA encoding achieved the best performance with AUC values of 0.768 and 0.819 for SVM and RF, respectively, followed by the SAC scheme. Next, we evaluated the SVM and RF models implementing one of the four encoding schemes on the independent dataset. The CPA encoding achieved the best performance for the four different encodings, with AUC values of 0.714, and 0.694 for SVM and RF, respectively. The TCP and PPA showed a reasonable performance regardless of classifiers on both the training and independent datasets. Since the TCP and PPA encodings represent different features from the CPA and SAC, the four features are integrated to expect high performance. Hence, we employed all the four encodings for the subsequent analysis.

Finally, we combined the above models implementing one of the four encoding schemes to construct three predictors, SVM_{com}, RF_{com}, and PreSNO. The weight coefficients of them were optimized to maximize the AUC. For the SVM_{com}, the weight coefficients for the CPA, SAC, TCP, and PPA schemes were 0.15, 0.3, 0.1, and 0.45, respectively (**Table 2**). In the RF_{com}, the weight coefficients for the CPA, SAC, TCP, and PPA schemes were 0.25, 0.15, 0.5, and 0.1, respectively (**Table 2**). In the SVM_{com}, the PPA and SAC-based models significantly contributed to the prediction, compared to the other encoding models. In the RF_{com}, the TCP greatly contributed to the prediction and the CPA moderately did. The contribution of each encoding scheme depended on the classifier algorithms. In the PreSNO, the weight coefficients for the scores of the $\text{SVM}_{\text{com}}$and RF_{com} were 0.35, and 0.65, respectively. Performance comparison of these three predictors showed that the PreSNO achieved the highest AUC value of 0.837 on the training dataset (**Figure 3A)**, where SP, SN, AC, and MCC were 0.863, 0.536, 0.700, and 0.422, respectively (**Table 3**). Particularly, the AC of the PreSNO was ~2-4% higher than RF_{com} and SVM_{com}, showing the advantage of integrating multiple classifiers. Furthermore, we compared the performances of the three predictors on the independent dataset. As shown in **Figure 3B**, the PreSNO achieved higher performance than any other individual classifiers. The PreSNO provided the best performance not only on the training dataset but also on the independent dataset.

Sequential combination model and meta-classifier

To demonstrate the strength of the combination method employed by the PreSNO, we built two competitive combination models. First, we assembled the four feature encoding vectors of CPA, SAC, TCP, and PPA in a row. It was named as the sequential combination model. The total dimension of the sequential combination model was 14,286. Based on the WR test, we selected the top 1,250 and 1,500 features and inputted them to RF and SVM, respectively, and evaluated the resultant prediction models using the 5-fold CV test. The two sequential combination models of the SVM and RF yielded AUC values of 0.811 and 0.829 on the training dataset, respectively (**Figure S3A**), and achieved 0.746 and 0.737 on the independent dataset, respectively **(Figure S3B**).

Second, we built a meta-classifier that integrated the output scores of different algorithms[46, 56, 57]. To construct the meta-predictor, the eight models utilized by the PreSNO (prior to weight optimization) were employed. The predicted probability scores from the eight models were inputted to SVM and RF individually. We characterized the above two approaches as shown in **Figure S3A**. The PreSNO showed an AUC of 0.837 on the training dataset, which was ~0.7- 4.0% higher than the sequential combination model and meta-classifier. It presented an AUC of 0.756 on the independent dataset, which was ~1.0-2.0% higher than the two models. Differing from the LR-based model and sequential combination model, in the meta-classifier model, the RF performed better than SVM on the training datasets; the SVM was slightly superior to the RF algorithm on the independent dataset **(Figures 3, S3, and S4)**. Since the PreSNO outperformed the two combination models, we selected the PreSNO as the final predictor.

Analysis of feature importance

By using the prediction models, the critical features were analyzed. We collected the average scores of the top 20 amino acid features and ranked them for the CPA, TCP, and SAC schemes via the WR test. The top 20 amino acid patterns were identified as critically important residues of adjacent SNO and non-SNO sites. The p-value of the residue pattern scores, selected for the CPA, SAC and TCP schemes, are shown in **Table S2**. The important features are depicted using a radar diagram (**Figure 4ABC**). In the CPA scheme the pattern of "AL" was top-ranked **(Table S2)**, but its significance was low. It is because the WR test selected the SNO and non-SNO site-specific patterns based on the sum of the ranks[28]. The "A×××L" pattern was enriched in the SNO sites. The patterns of "LL", "K××A", "Y×E", "L×××A", "L×××V", "KK", and "LS" were enriched around the non-SNO sites (Figure 4A). In the SAC scheme, the top-ranking patterns of "LL" and the second one of "L×E" were enriched around non-SNO sites (**Figure 4B**). In both the CPA and SAC schemes *k*-spaces residue pairs (i.e., " ", "×", "××", "×××", "××××") were observed **(Figure 4)**, indicating that 0,1,2,3 and 4 spaces of residue pairs are useful for the CPA and SAC schemes. In the TCP scheme, the top-ranking pattern "LKK" was enriched around non-SNO sites. The patterns of "NLE", "DKL", "GLK", "AAL", "ALL", and "DAK" were enriched around SNO sites (**Figure**

4C). Those analyses suggested that statistically, different sequence patterns exist between the SNO and non-SNO samples.

Furthermore, we investigated the significant residue patterns for each encoding from **Table S2** and compared the amino acid preferences among them. As shown in **Table S3**, in the CPA, SAC, TCP, and PPA encodings, charged, hydrophobic, and polar amino acids containing patterns were distributed on both the SNO and non-SNO sequences. It was hard to find specific patterns to SNO, while different amino acid patterns significantly existed between the SNO and non-SNO sequences. In the SAC, TCP, and PPA, charged amino acid containing patterns would be preferred by the SNO sites rather than by the non-SNO ones, which may suggest that the charged residues play a role in SNO.

Comparison with other existing tools

To date, four predictors (GPS-SNO [23], SNOSite [24], iSNOPseAAC [25], and DeepNitro [26]) are publicly available to predict SNO sites. To make a fair comparison among the four available predictors, we submitted the independent samples (351 SNO sites with 3,168 non-SNO sites) to them and measured the numbers of TP, FP, TN, and FN and calculated SP, SN, AC, and MCC. As shown in **Table 4**, the PreSNO (SP=0.769, AC=0.752, and MCC=0.252) greatly outperformed the SNOsite, iPseAAC, GPS-SNO, and DeepNitro. In terms of balanced performance, the PreSNO and DeepNitro were the two top methods showing the lowest difference between SN and SP. Other prediction models were biased towards either SN or SP, where the difference (|SN-SP|) was large, due to their imbalanced ratios of the training to independent samples. Overall, the proposed PreSNO provided a more reliable prediction than the existing tools.

Advantages of PreSNO

The advantages of the PreSNO over existing predictors are summarized: (1) The PreSNO integrated four types of complementary encoding schemes to train the SVM and RF models, while the existing predictors used only position-wise encoding methods that were unable to fully characterize the potential SNO sites. (2) The PreSNO employed the most updated version of the dataset as well as the DeepNitro predictor, while the existing GPS-SNO [23], SNOSite [24], and iSNOPseAAC [25] predictors used the small and old version of SNO datasets. (3) The PreSNO server provided the probability scores of the SNO sites so that users can understand the actual prediction results, while the existing SNOSite and iSNOPseAAC predictors did not.

Conclusions

We have established a computational tool PreSNO to predict SNO sites by integrating the four encoding schemes with SVM and RF algorithms through an LR model. The PreSNO is a promising predictor that outperforms the existing prediction models. The LR-based combination of the PreSNO was demonstrated to outperform two typically used combination methods (sequential combination method and meta classifier). Furthermore, a feature selection analysis characterized significant sequence patterns to facilitate an understanding of the prediction model. Finally, a web application of our tool is provided for the public.

Authors' contributions

MMH and HK conceived the project. MMH collected and analyzed the datasets. MMH drafted the manuscript. HK, MMH, MSK, and MB thoroughly revised the manuscript. All authors approved and read the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

MMH is a JSPS international research fellow http://www.jsps.go.jp/english. This work was supported by the Grant-in-Aid for Scientific Research (B) (19H04208) and partially supported by the developing key technologies for discovering and manufacturing pharmaceuticals used for next-generation treatments and diagnoses both from the Ministry of Economy, Trade and Industry, Japan (METI) and from Japan Agency for Medical Research and Development (AMED).

References

1. I. Gusarov and E. Nudler, *Molecular cell*, 2018, **69**, 351-353.
2. M. Lenarcic Zivkovic, M. Zareba-Koziol, L. Zhukova, J. Poznanski, I. Zhukov and A. Wyslouch-Cieszynska, *The Journal of biological chemistry*, 2012, **287**, 40457-40470.
3. H. P. Monteiro, P. E. Costa, A. K. Reis and A. Stern, *Biomedical journal*, 2015, **38**, 380-388.

4. M. W. Foster, D. T. Hess and J. S. Stamler, *Trends in molecular medicine*, 2009, **15**, 391-404.
5. B. Derakhshan, G. Hao and S. S. Gross, *Cardiovascular research*, 2007, **75**, 210-219.
6. D. T. Hess, A. Matsumoto, S. O. Kim, H. E. Marshall and J. S. Stamler, *Nature reviews. Molecular cell biology*, 2005, **6**, 150-166.
7. S. R. Jaffrey, H. Erdjument-Bromage, C. D. Ferris, P. Tempst and S. H. Snyder, *Nature cell biology*, 2001, **3**, 193-197.
8. J. S. Stamler, S. Lamas and F. C. Fang, *Cell*, 2001, **106**, 675-683.
9. C. T. Stomberski, D. T. Hess and J. S. Stamler, *Antioxidants & redox signaling*, 2019, **30**, 1331-1351.
10. J. Feng, L. Chen and J. Zuo, *Journal of integrative plant biology*, 2019, DOI: 10.1111/jipb.12780.
11. S. B. Wang, V. Venkatraman, E. L. Crowgey, T. Liu, Z. Fu, R. Holewinski, M. Ranek, D. A. Kass, B. O'Rourke and J. E. Van Eyk, *Circulation research*, 2018, **122**, 1517-1531.
12. E. Vanzo, J. Merl-Pham, V. Velikova, A. Ghirardo, C. Lindermayr, S. M. Hauck, J. Bernhardt, K. Riedel, J. Durner and J. P. Schnitzler, *Plant physiology*, 2016, **170**, 1945-1961.
13. V. Mahishale, B. Patil, M. Lolly, A. Eti and S. Khan, *Chonnam medical journal*, 2015, **51**, 86-90.
14. J. Romero-Aguirregomezcorta, A. P. Santa, F. A. Garcia-Vazquez, P. Coy and C. Matas, *PloS one*, 2014, **9**, e115044.
15. L. F. Anderson, S. Tamne, J. P. Watson, T. Cohen, C. Mitnick, T. Brown, F. Drobniewski and I. Abubakar, *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 2013, **18**.
16. Y. Koriyama, *Yakugaku zasshi : Journal of the Pharmaceutical Society of Japan*, 2013, **133**, 843-848.
17. Z. Wang, *Cancer letters*, 2012, **320**, 123-129.
18. T. Nakamura, O. A. Prikhodko, E. Pirie, S. Nagar, M. W. Akhtar, C. K. Oh, S. R. McKercher, R. Ambasudhan, S. Okamoto and S. A. Lipton, *Neurobiology of disease*, 2015, **84**, 99-108.
19. T. Yasukawa, E. Tokunaga, H. Ota, H. Sugita, J. A. Martyn and M. Kaneki, *The Journal of biological chemistry*, 2005, **280**, 7511-7518.
20. T. Nakamura, S. Tu, M. W. Akhtar, C. R. Sunico, S. Okamoto and S. A. Lipton, *Neuron*, 2013, **78**, 596-614.

21. P. Anand and J. S. Stamler, *Journal of molecular medicine*, 2012, **90**, 233-244.
22. S. L. Cook and G. P. Jackson, *Journal of the American Society for Mass Spectrometry*, 2011, **22**, 221-232.
23. Y. Xue, Z. Liu, X. Gao, C. Jin, L. Wen, X. Yao and J. Ren, *PloS one*, 2010, **5**, e11290.
24. T. Y. Lee, Y. J. Chen, T. C. Lu, H. D. Huang and Y. J. Chen, *PloS one*, 2011, **6**, e21849.
25. Y. Xu, J. Ding, L. Y. Wu and K. C. Chou, *PloS one*, 2013, **8**, e55844.
26. Y. Xie, X. Luo, Y. Li, L. Chen, W. Ma, J. Huang, J. Cui, Y. Zhao, Y. Xue, Z. Zuo and J. Ren, *Genomics, proteomics & bioinformatics*, 2018, **16**, 294-306.
27. F. Luo, M. Wang, Y. Liu, X. M. Zhao and A. Li, *Bioinformatics*, 2019, **35**, 2766-2773.
28. Z. Chen, X. Liu, F. Li, C. Li, T. Marquez-Lago, A. Leier, T. Akutsu, G. I. Webb, D. Xu, A. I. Smith, L. Li, K. C. Chou and J. Song, *Briefings in bioinformatics*, 2018, DOI: 10.1093/bib/bby089.
29. X. Wang, R. Yan, J. Li and J. Song, *Molecular bioSystems*, 2016, **12**, 2849-2858.
30. L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, *Bioinformatics*, 2012, **28**, 3150-3152.
31. M. M. Hasan, M. S. Khatun, M. N. H. Mollah, C. Yong and G. Dianjing, *Molecules*, 2018, 23(7), 1667.
32. M. M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song and Z. Zhang, *PloS one*, 2015, **10**, e0129635.
33. F. Provost, *AAAI Technical Report,* 2000, 1-3.
34. M. M. Hasan and H. Kurata, *IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung*, 2018, 356-359.
35. Y. Zhou, P. Zeng, Y. H. Li, Z. Zhang and Q. Cui, *Nucleic acids research*, 2016, **44**, e91.
36. Z. Chen, Y. Zhou, Z. Zhang and J. Song, *Briefings in bioinformatics*, 2015, **16**, 640-657.
37. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, *Nucleic acids research*, 2008, **36**, D202-205.
38. B. Manavalan, T. H. Shin, M. O. Kim and G. Lee, *Frontiers in Immunology*, 2018, **9**, 1695.
39. M. M. Hasan, S. Yang, Y. Zhou and M. N. Mollah, *Molecular bioSystems*, 2016, **12**, 786-795.

40. M. M. Hasan, M. S. Khatun, M. N. H. Mollah, C. Yong and D. Guo, *International journal of nanomedicine*, 2017, **12**, 6303-6315.

41. M. M. Hasan, M. S. Khatun and H. Kurata, *Protein and peptide letters*, 2018, **25**, 815-821.

42. M. M. Hasan, D. Guo and H. Kurata, *Molecular bioSystems*, 2017, **13**, 2545-2550.

43. B. Manavalan, T. H. Shin, M. O. Kim and G. Lee, *Frontiers in pharmacology*, 2018, **9**, 276.

44. M. M. Hasan, M. S. Khatun and H. Kurata, *Cells*, 2019, 8(2), 95.

45. V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan and D. C. Yang, *Int J Mol Sci*, 2019, 20 (8), 1964.

46. B. Manavalan, S. Basith, T. H. Shin, L. Wei and G. Lee, *Molecular Therapy-Nucleic Acids*, 2019, 16, 733-74.

47. S. Khatun, M. Hasan and H. Kurata, *FEBS Lett*, 2019, DOI: 10.1002/1873-3468.13536.

48. Z. Chen, Y. Zhou, J. Song and Z. Zhang, *Biochimica et biophysica acta*, 2013, **1834**, 1461-1467.

49. L. Breiman, *Machine learning*, 2001, **45**, 5-32.

50. M. M. Hasan, M. M. Rashid, M. S. Khatun, and H. Kurata, *Sci Rep*, 2019, 9, 8258.

51. M. M. Hasan and H. Kurata, *PloS one*, 2018, **13**, e0200283.

52. B. Manavalan, R. G. Govindaraj, T. H. Shin, M. O. Kim and G. Lee, *Front Immunol*, 2018, **9**, 1695.

53. B. Manavalan, T. H. Shin, M. O. Kim and G. Lee, *Front Immunol*, 2018, **9**, 1783.

54. M. S. Khatun, M. M. Hasan and H. Kurata, *Frontiers in Genetics*, 2019, 10:129.

55. J. P. O'Shea, M. F. Chou, S. A. Quader, J. K. Ryan, G. M. Church and D. Schwartz, *Nature methods*, 2013, **10**, 1211-1212.

56. L. Wei, R. Su, S. Luan, Z. Liao, B. Manavalan, Q. Zou and X. Shi, *Bioinformatics*, 2019, DOI: 10.1093/bioinformatics/btz408.

57. B. Manavalan, S. Basith, T. H. Shin, L. Wei and G. Lee, *Bioinformatics*, 2019, 35, 2757-2765.

Figure Legends

Figure 1. The overall framework of the PreSNO.

Figure 2. Amino acid residue preference around the SNO and non-SNO sites. The residues flanking the SNO sites that were significantly enriched or depleted (p<0.05) are shown. The pLogo of the two-sample sequence was prepared using the webserver http://www.twosamplelogo.org/.

Figure 3. ROC curves of the SVM_{com}, RF_{com}, and PreSNO.
(A) Training data. (B) Independent data.

Figure 4. Average scores of top 20 amino acid patterns selected by the WR test. Green color denotes the SNO sites, while blue color denotes the non-SNO sites.
(A) CPA, (B) SAC, and (C) TCP scheme.

Supplementary Information

Figure S1. Effect of the ratio of positive to negative training datasets on prediction performances by the ML models with a single encoding scheme of the CPA, SAC, TCP, or PPA.

(A) AUC values provided by the SVM. (B) AUC values provided by the RF.

Figure S2. Effect of different window sizes on the AUC values with a single encoding scheme of the CPA, SAC, TCP, or PPA on the training datasets. (A) SVM and (B) RF algorithms.

Figure S3. Prediction performance provided by the RF and SVM with the sequential combination of the CPA, SAC, TCP, and PPA.

(A) Training data. (B) Independent data.

Figure S4. Prediction performance by the meta-classifier algorithms of the SVM and RF.

(A) Training data. (B) Independent data.

Table S1. Selected amino acid index properties for the PPA encoding scheme.

Table S2. Top 20 selected features based on the CPA, SAC, TCP, and PPA encoding schemes by the WR test.

Table S3. Comparison of unique amino acids between different encodings.

Tables

Table 1. Effect of the four types of encoding schemes on the AUCs by the SVM and RF on the training and independent datasets.

Methods	SVM		RF	
	Training	Independent	Training	Independent
CPA	0.768	0.714	0.819	0.694
SAC	0.764	0.709	0.788	0.682
TCP	0.738	0.682	0.763	0.672
PPA	0.731	0.703	0.759	0.680

Table 2. Weight coefficients of each encoding scheme for two combined models

Combined model	**CPA**	**SAC**	**TCP**	**PPA**
SVM$_{com}$- weight coefficient	0.15	0.30	0.10	0.45
RF$_{com}$-weight coefficient	0.25	0.15	0.50	0.10

Table 3. Prediction performance by the combined models on the training dataset

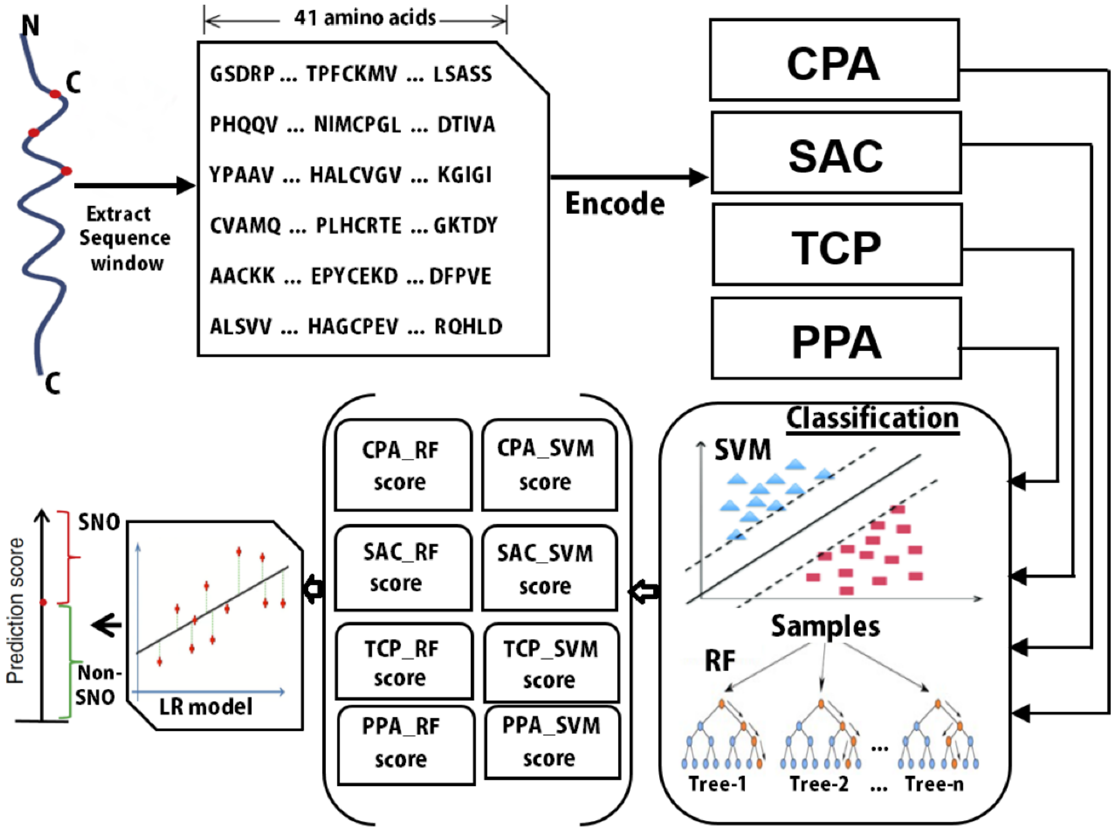
Methods	TP	FP	TN	FN	SP	SN	AC	MCC
SVM$_{com}$	1586	500	2883	1797	0.852	0.469	0.661	0.348
RF$_{com}$	1709	465	2913	1674	0.862	0.505	0.684	0.393
PreSNO	1812	462	2921	1571	0.863	0.536	0.700	0.422

In the PreSNO, the weight coefficients of the SVMcom and RFcom scores were 0.35 and 0.65, respectively.

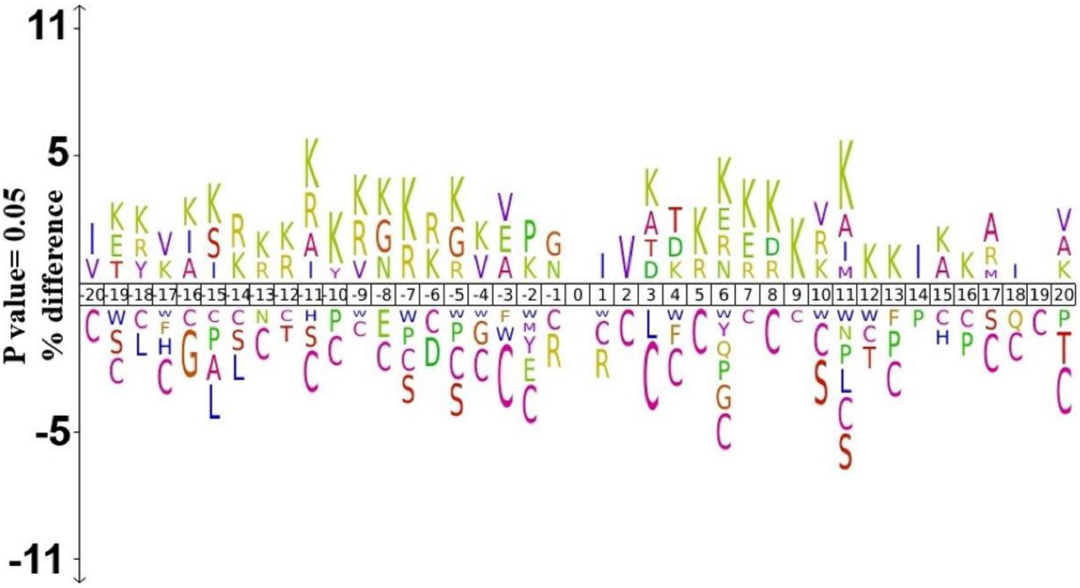
Table 4. Comparison of the PreSNO with existing predictors

Model	TP	FP	TN	FN	SP	SN	AC	MCC	**AUC**
GPS-SNO	99	825	2337	253	0.739	0.281	0.693	0.014	0.523
iSNOPseAAC	101	768	2394	251	0.757	0.287	0.710	0.031	-
SNOSite	235	1749	1413	117	0.447	0.668	0.469	0.069	-
DeepNitro	202	776	2386	148	0.755	0.578	0.737	0.222	0.731
PreSNO	211	733	2431	141	0.769	0.604	0.752	0.252	0.756

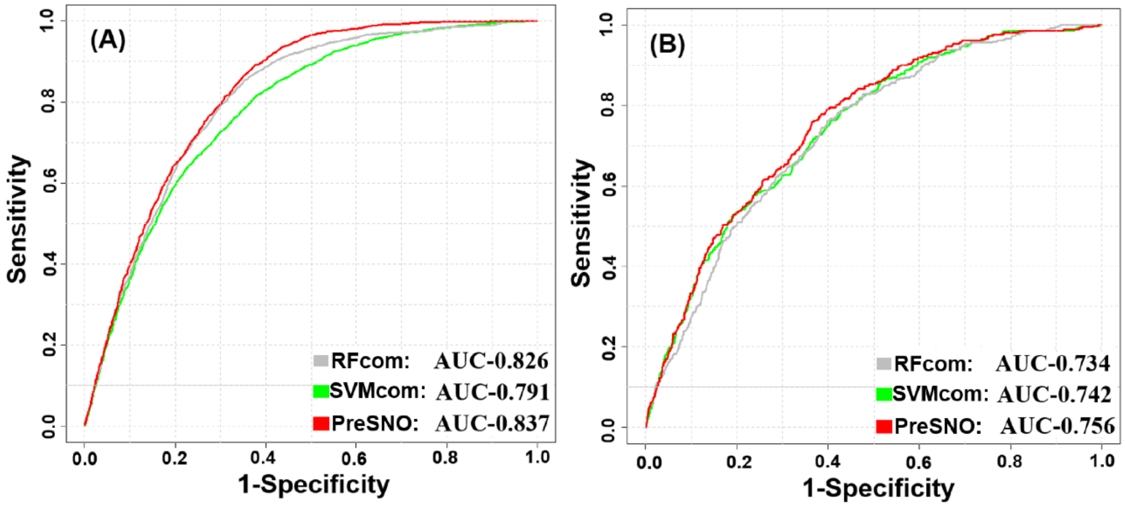
All threshold values of the GPS-SNO and DeepNitro were considered. In the iSNOPseAAC and SNOSite predictors, the medium threshold was used in their corresponding online servers.



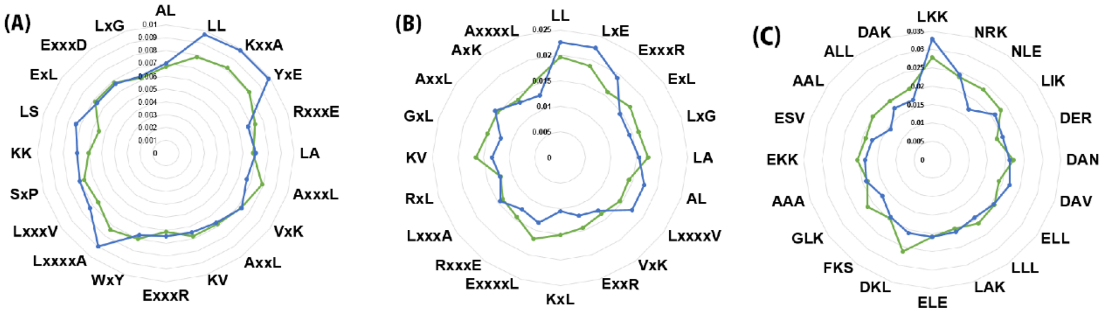
(Hasan et al. **Figure 1**)



(Hasan et al. **Figure 2**)



(Hasan et al. **Figure 3**)



(Hasan et al. **Figure 4**)

Supplementary Information

Prediction of S-nitrosylation Sites by Integrating Support Vector Machine and Random Forest

Md. Mehedi Hasan[a], Balachandran Manavalan[b] , Mst. Shamima Khatun[a], and Hiroyuki Kurata[a,c*]

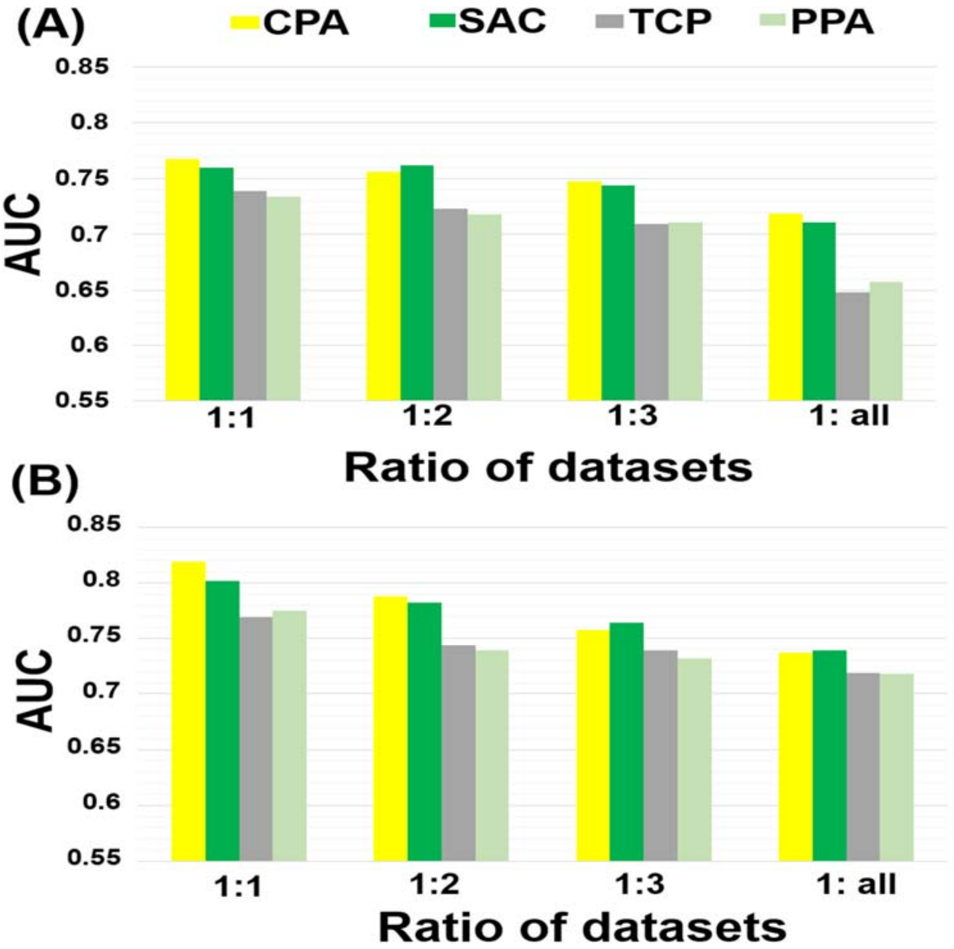


Figure S1. Effect of the ratio of positive vs negative training datasets on prediction performances by the ML models with a single encoding scheme of the CPA, SAC, TCP, or PPA.

(A) AUC values provided by the SVM. (B) AUC values provided by the RF

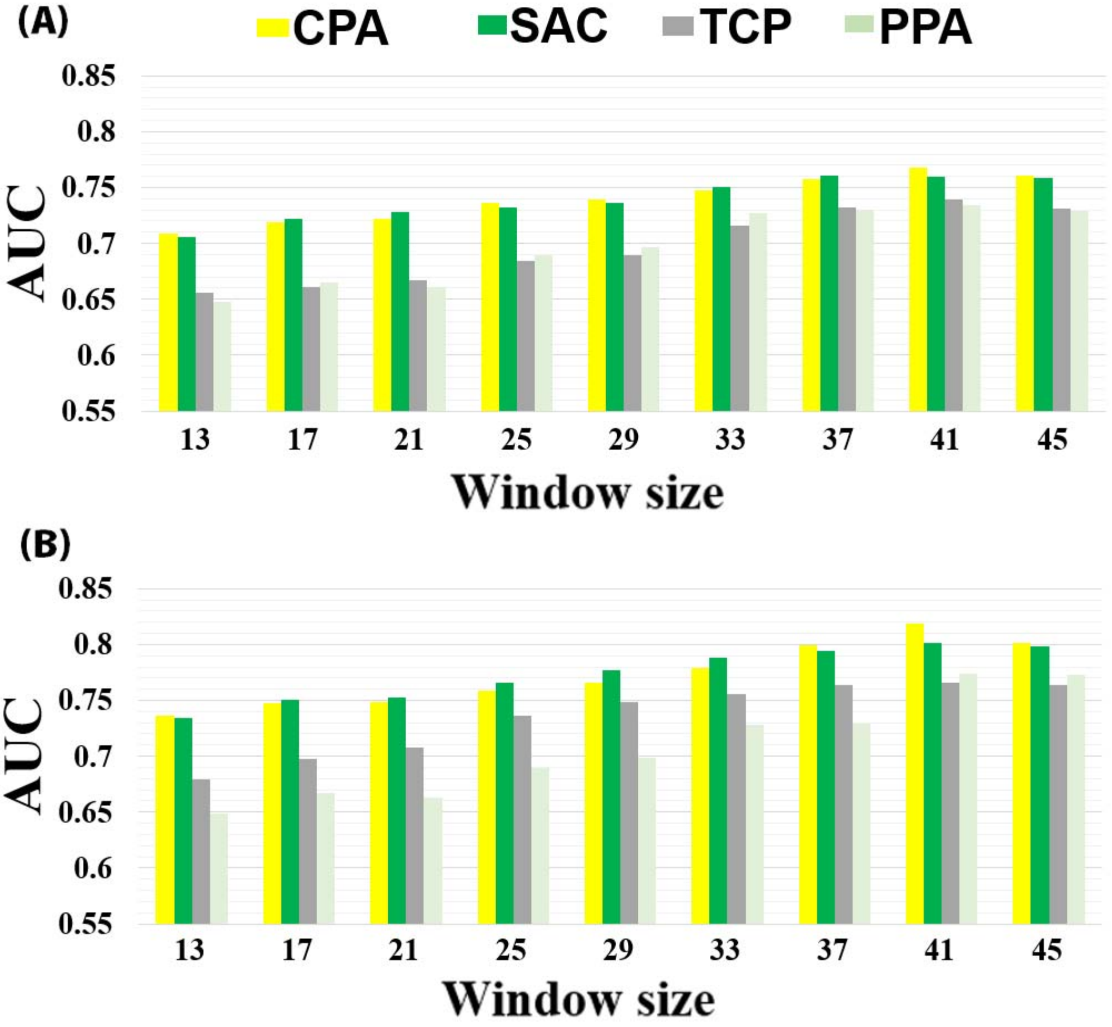


Figure S2. Effect of window sizes on the AUC values by a single encoding scheme of the CPA, SAC, TCP, or PPA on the training datasets by 5-fold CV test.

(A) SVM and (B) RF algorithms.

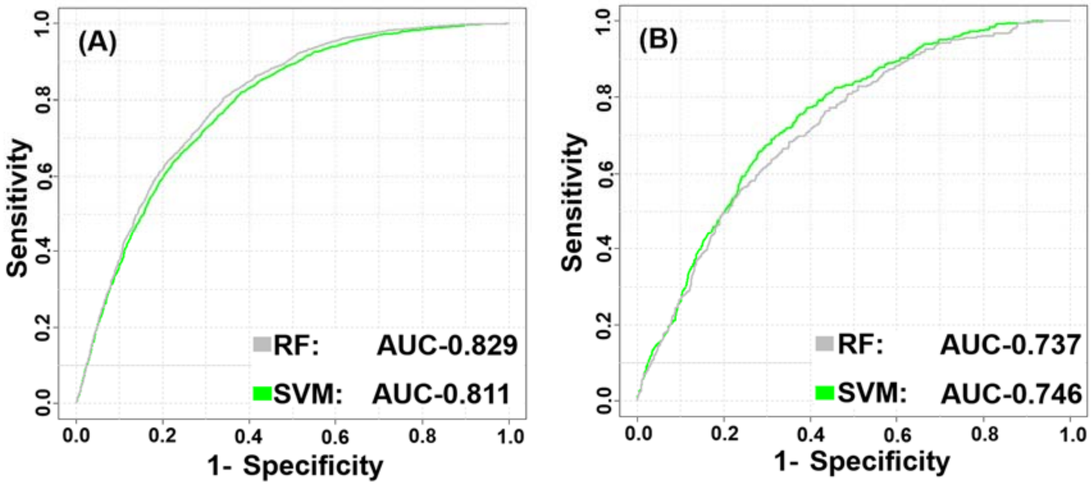


Figure S3. Prediction performance provided by the RF and SVM with the sequential combination of the CPA, SAC, TCP, and PPA.

(A) Training data. (B) Independent data.

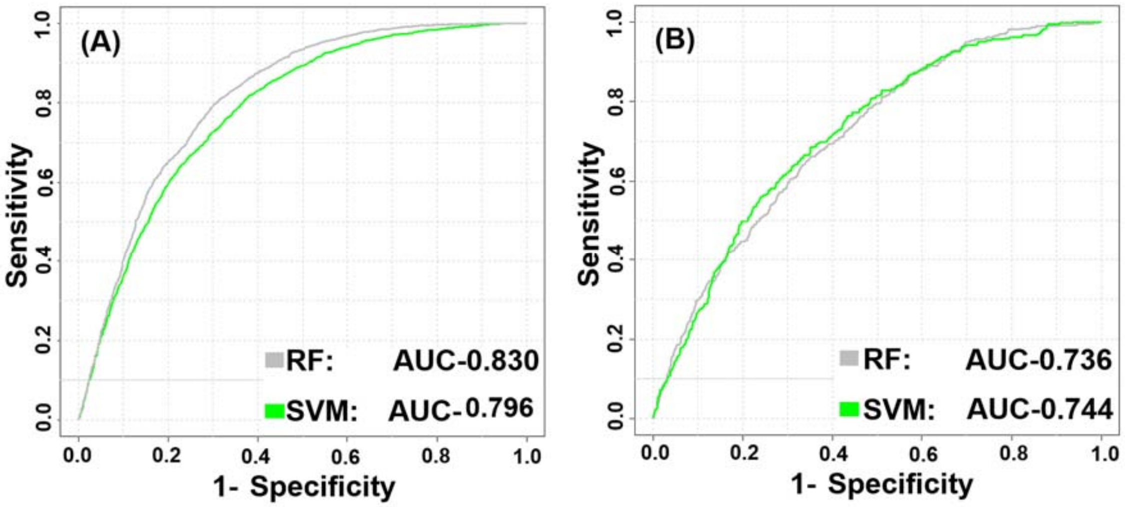


Figure S4. Prediction performance by the meta-classifiers of the SVM and RF.

(A) Training data. (B) Independent data.

Table S1. Selected amino acid index properties for the PPA encoding scheme.

PPA property	Amino acid index properties						
BLAM930101	0.96	0.77	0.39	0.42	0.42	0.80	0.53
	0.00	0.57	0.84	0.92	0.73	0.86	0.59
	-2.50	0.53	0.54	0.58	0.72	0.63	
MAXF760101	1.43	1.18	0.64	0.92	0.94	1.22	1.67
	0.46	0.98	1.04	1.36	1.27	1.53	1.19
	0.49	0.70	0.78	1.01	0.69	0.98	
TSAJ990101	89.3	190.3	122.4	114.4	102.5	146.9	138.8
	63.8	157.5	163.0	163.1	165.1	165.8	190.8
	121.6	94.2	119.6	226.4	194.6	138.2	
NAKH920108	9.36	0.27	2.31	0.94	2.56	1.14	0.94
	6.17	0.47	13.73	16.64	0.58	3.93	10.99
	1.96	5.58	4.68	2.20	3.13	12.43	
CEDJ970104	7.9	4.9	4.0	5.5	1.9	4.4	7.1
	7.1	2.1	5.2	8.6	6.7	2.4	3.9
	5.3	6.6	5.3	1.2	3.1	6.8	
LIFS790101	0.92	0.93	0.60	0.48	1.16	0.95	0.61
	0.61	0.93	1.81	1.30	0.70	1.19	1.25
	0.40	0.82	1.12	1.54	1.53	1.81	
NOZY710101	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.5	1.8	1.8	0.0	1.3	2.5
	0.0	0.0	0.4	3.4	2.3	1.5	
HUTJ700103	154.33	341.01	207.90	194.91	219.79	235.51	223.16
	127.90	242.54	233.21	232.30	300.46	202.65	204.74
	179.93	174.06	205.80	237.01	229.15	207.60	
NAKH900109	9.25	3.96	3.71	3.89	1.07	3.17	4.80
	8.51	1.88	6.47	10.94	3.50	3.14	6.36
	4.36	6.26	5.66	2.22	3.28	7.55	
BIOV880101	16.	-70.	-74.	-78.	168.	-73.	-106.
	-13.	50.	151.	145.	-141.	124.	189.
	-20.	-70.	-38.	145.	53.	123.	
MIYS990104	-0.04	0.07	0.13	0.19	-0.38	0.14	0.23
	0.09	-0.04	-0.34	-0.37	0.33	-0.30	-0.38
	0.19	0.12	0.03	-0.33	-0.29	-0.29	

PUNT030101	-0.17	0.37	0.18	0.37	-0.06	0.26	0.15
	0.01	-0.02	-0.28	-0.28	0.32	-0.26	-0.41
	0.13	0.05	0.02	-0.15	-0.09	-0.17	
WOEC730101	7.0	9.1	10.0	13.0	5.5	8.6	12.5
	7.9	8.4	4.9	4.9	10.1	5.3	5.0
	6.6	7.5	6.6	5.3	5.7	5.6	
BASU050102	0.0728	0.0394	-0.0390	-0.0552	0.3557	0.0126	-0.0295
	-0.0589	0.0874	0.3805	0.3819	-0.0053	0.1613	0.4201
	-0.0492	-0.0282	0.0239	0.4114	0.3113	0.2947	
SUYM030101	-0.058	0.000	0.027	0.016	0.447	-0.073	-0.128
	0.331	0.195	0.060	0.138	-0.112	0.275	0.240
	-0.478	-0.177	-0.163	0.564	0.322	-0.052	

Table S2. Top 20 selected features based on the CPA, SAC, TCP, and PPA schemes by the WR test.

Sequential order	CPA		SAC		TCP		PPA	
	p-value	Selected pattern	*p*-value	Selected pattern	*p*-value	Selected pattern	*p*-value	Window position
1	6.03E-01	AL	3.06E-03	LL	2.54E-03	LKK	1.35E-02	(L, -9)
2	7.37E-06	LL	1.63E-04	L×E	6.52E-01	NRK	6.33E-04	(E, +11)
3	1.42E-04	K××A	1.70E-04	E×××R	2.49E-02	NLE	4.47E-02	(K, -2)
4	1.63E-05	Y×E	6.52E-02	E×L	9.95E-01	LIK	3.33E-06	(H, +6)
5	8.61E-01	R×××E	3.47E-02	L×G	2.56E-01	DER	5.44E-01	(D, -3)
6	6.90E-01	LA	8.41E-02	LA	9.37E-01	DAN	3.67E-02	(E, +6)
7	2.54E-03	A×××L	2.28E-03	AL	1.31E-02	DAV	7.57E-02	(V, -11)
8	9.17E-02	V×K	5.10E-03	L××××V	8.58E-01	ELL	1.14E-03	(S, -2)
9	6.16E-01	A××L	2.82E-01	V×K	9.38E-01	LLL	3.11E-03	(T, -12)
10	3.67E-01	KV	3.22E-02	E××R	4.25E-01	LAK	7.88E-02	(R, +6)
11	2.58E-01	E×××R	2.24E-07	K×L	9.84E-01	ELE	3.15E-02	(P, +8)
12	3.19E-01	W×Y	5.96E-04	E××××L	4.081E-04	DKL	1.02E-03	(V, -15)
13	8.25E-05	L××××A	2.68E-02	R×××E	7.25E-01	FKS	4.11E-02	(P, +2)
14	1.39E-03	L×××V	5.20E-01	L×××A	7.95E-05	GLK	3.49E-02	(K, +10)
15	2.61E-01	S×P	9.10E-02	R×L	8.29E-01	AAA	7.46E-02	(E, +7)
16	3.16E-02	KK	7.09E-04	KV	9.8E-03	EKK	5.44E-03	(N, +18)
17	1.15E-07	LS	1.34E-03	G×L	1.17E-01	ESV	3.11E-02	(Y, -13)
18	6.65E-01	E×L	5.88E-01	A××L	5.89E-08	AAL	3.45E-03	(R, +11)
19	7.42E-01	E×××D	3.86E-01	A×K	4.98E-02	ALL	2.22E-03	(G, +16)
20	5.57E-01	L×G	1.08E-03	A××××L	1.49E-02	DAK	2.54E-02	(L, +9)

A *p*-value is calculated by a two-sample *t*-test.

Table S3. Comparison of the amino acid residue patterns selected by different encodings.

Encoding	CPA		SAC		TCP		PPA	
Sample	SNO	non-SNO	SNO	non-SNO	SNO	non-SNO	SNO	non-SNO
Pattern	AxxxL	LL KxxA YxE LxxxxA LxxxV KK LS	LxG ExxxR KxL ExxxL RxxxE KV GxL AxxxL	LL LxE ExxR AL LxxxV	NLE DKL GLK EKK AAL ALL DAK	LKK DAV	(L, -9) (E, +6) (V, -15) (R, +11) (K, +10) (L, +9) (Y, -13)	(E, +11) (K, -2) (H, +6) (S, -2) (T, -12) (P, +8) (G, +16) (P, +2) (N, 18)
#charged	0	3	3	2	4	2	3	2
#hydrophobic	2	3	3	3	3	3	2	2
#polar	0	1	0	0	1	0	1	4

Blue indicates charged amino acids; red hydrophobic amino acids; black polar amino acids.

#charged, the unique number of charged amino acid containing patterns;

#hydrophobic, the unique number of hydrophobic amino acid containing patterns;

#polar, the unique number of polar amino acid containing patterns.