Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Accurately predicting nitrosylated tyrosine sites using probabilistic sequence information

Afrida Rahman[a,1], Sabit Ahmed[a,1], Md. Al Mehedi Hasan[a], Shamim Ahmad[b], Iman Dehzangi[c,d,*]

[a] *Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh*
[b] *Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh*
[c] *Department of Computer Science, Rutgers University, Camden, NJ 08102, USA*
[d] *Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA*

ARTICLE INFO

Edited by John Doe

Keywords:
Post-translational modification
Nitrotyrosine Sites Prediction
Sequence-coupling Model
General PseAAC
Data Imbalance Issue
Support Vector Machine

ABSTRACT

Post-translational modification (PTM) is defined as the enzymatic changes of proteins after the translation process in protein biosynthesis. Nitrotyrosine, which is one of the most important modifications of proteins, is interceded by the active nitrogen molecule. It is known to be associated with different diseases including autoimmune diseases characterized by chronic inflammation and cell damage. Currently, nitrotyrosine sites are identified using experimental approaches which are laborious and costly. In this study, we propose a new machine learning method called PredNitro to accurately predict nitrotyrosine sites. To build PredNitro, we use sequence coupling information from the neighboring amino acids of tyrosine residues along with a support vector machine as our classification technique.Our results demonstrates that PredNitro achieves 98.0% accuracy with more than 0.96 MCC and 0.99 AUC in both 5-fold cross-validation and jackknife cross-validation tests which are significantly better than those reported in previous studies. PredNitro is publicly available as an online predictor at: http://103.99.176.239/PredNitro.

1. Introduction

Post-translational modifications (PTMs) introduce a series of crucial protein modifications after the translation phase (Saraswathy and Ramalingam, 2011). PTMs alter and derivate intra-molecular bonds of amino acids with drastic impacts on proteomic analysis and biological processes, such as cellular signal transduction, metabolism, subcellular localization, protein folding, and protein degradation (McDowell and Philpott, 2016; Weissman et al., 2003; Ghauri et al., 2018; Blantz and Munger, 2002). Hence, efficient identification and appropriate understanding of PTM sites are essential for basic research in the fields of disease detection and prevention, and drug development (Chou, 2015; Qiu et al., 2016). Among 20 fundamental amino acid residues that build proteins, modifications at tyrosine residue (Y) are usually referred to as tyrosine PTM or Y-PTM. There are several tyrosine PTMs such as amidation, phosphorylation, nitration, hydroxylation, sulfation, and ubiquitination (Lee et al., 2006).

Among several forms of PTMs, protein nitrotyrosine is of critical importance. It is generated by the interaction of tyrosine along with nitrate molecules in peroxynitrite (ONOO-), which is reactive and often derived from an aggregation of superoxide radical anion (O2-) and nitric oxide (NO) (Abello et al., 2010) shown in Fig. 1. Nitrotyrosine is regarded as an indicator of inflammation and cell injury. It is also shown to be involved in diseases such as septic shock, Alzheimer, lung cancer, rheumatoid arthritis, celiac disease, cardiovascular disease and asthma (Giasson et al., 2000; Donnini et al., 2008; Brindicci et al., 2010).The experimental method for precisely identifying nitrotyrosine sites is expensive and time-consuming. It is even more sensitive toward proteins that are plentiful. (Hasan et al., 2018). Therefore, there is a demand to develop fast computational approaches to accurately predict nitrotyrosine sites. (Qiu et al., 2017; Rahman et al., 2020).

During the past few years, a wide range of computational methods have been proposed to predict nitrotyrosine sites. The first predictor of nitrotyrosine sites named 'GPS-YNO2' was proposed by Liu et al. (2011). It was developed using four statistical analyses including matrix mutation, weight training, k-means clustering, and motif length selection.

[*] Corresponding author at: Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh.
E-mail address: i.dehzangi@rutgers.edu (I. Dehzangi).

[1] Contributed equally to this work.

https://doi.org/10.1016/j.gene.2022.146445
Received 8 September 2021; Received in revised form 16 February 2022; Accepted 18 March 2022
Available online 28 March 2022
0378-1119/© 2022 Elsevier B.V. All rights reserved.

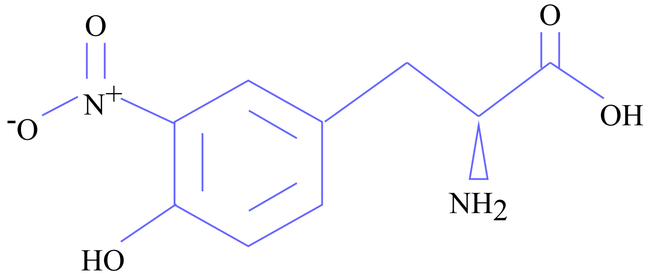

Fig. 1. Chemical structure of 3-Nitrotyrosine.

Table 1
Summary Of dataset.

Dataset	Positive sites	Negative sites
Training dataset	1191	1191
Independent dataset	203	1022

Later on, Xu et al. (2014) constructed ’iNitro-Tyr’ using the composition of pseudo amino acid encoding. After that, Ghauri et al. developed ’pNitro-Tyr-PseAAC’ (Ghauri et al., 2018) using a backpropagation neural network. In a different study, Xie et al. (2018) constructed ’DeepNitro’ using deep learning with four encoding schemes (i.e., positional amino acid distributions, sequence contextual dependencies, physicochemical properties, and position-specific scoring features) to predict nitrotyrosine sites. At the same time, Hasan et al. (2018) developed ’NTyroSite’ in which the Wilcoxon-rank sum test was applied as a feature selection technique and Random Forest was applied as a classifier. Most recently, a new machine learning method named ’PredNTS’ (Nilamyani et al., 2021) was developed by integrating multiple feature encoding schemes such as K-mer, composition of k-spaced amino acid pairs (CKSAAP), AAindex, and binary as well as using Random Forest as a classifier. It achieved better performance compared to the other previous studies. PredNTS achieved 91% AUC in 5-fold cross-validation test. However, PredNTS does not achieve similar results in terms of sensitivity and MCC. It means that it is better in predicting negative nitrotyrosine samples rather than positive nitrotyrosine samples.

Recognizing the aforementioned scenario, there is a demand for developing novel system for identifying nitrotyrosine sites with higher efficacy. In order to develop an efficient predictor, appropriate sequence patterns associated with tyrosine nitration need to be retrieved. In this study, we propose a new machine learning method called PredNitro to accurately predict nitrotyrosine sites. To build this model, we utilize the vectorized sequence-coupled features to capture the useful information out of the protein sequences and a support vector machine (SVM) as our classification technique (Dehzangi et al., 2015; Vapnik, 2013; Ahmed et al., 2021; Ahmed et al., 2021; Rahman et al., 2020; Ahmed et al., 2021). PredNitro achieves more than 99% AUC in jackkinfe test, k-fold cross-validation test, and independent test. The exploratory results of other crucial metrics demonstrate the superior performance of PredNitro over the other existing approaches. PredNitro is publicly available as an online predictor at: http://103.99.176.239/PredNitro.

2. Materials and methods

2.1. Dataset

The nitrotyrosine dataset for this study was collected from multiple databases (dbPTM, SysPTM2.0, GPS-YNO2) including DeepNitro and iNitro-Tyr, as stated by Nilamyani et al. (2021). It contains 796 proteins with 1406 experimentally validated nitrotyrosine sites. Subsequently, redundant sequences were discarded using CD-HIT with a similarity cutoff of 40% to prevent overfitting issues in performance measurement, since this degree of redundancy reduction is widely acknowledged. For the independent test dataset, 20% of the samples were picked randomly to assure the feasibility of the proposed predictor for new and unseen proteins. Thereafter, a 1:1 ratio of positive to negative samples was chosen from the entire remaining dataset to construct the training set which is identical to the ’PredNTS’ (Nilamyani et al., 2021). As a result, the non-redundant training dataset was attained composing 1191 experimentally positive nitrotyrosine samples and 1191 negative nitrotyrosine samples. On the other hand, the independent test dataset was attained composing 203 experimentally positive nitrotyrosine samples and 1022 negative nitrotyrosine samples (See Table 1). The training and independent test dataset with corresponding peptide sequences and site positions are hosted in a GitHub repository at https://github.com/Sabit-Ahmed/PredNitro. An overview of the dataset preparation as well as the general architecture of PredNitro is shown in Fig. 2. For analyzing the statistically verified disparity among the positive and negative nitrotyrosine samples in our dataset, the distribution of amino acid residues in the positive samples and negative samples are visually explored by the guidance of WebLogo in Fig. 3 and Fig. 4 (Crooks et al., 2004).

2.2. Feature construction

In this study, Chou’s scheme (Chou, 1993; Ahmed et al., 2021) was implemented to encode more scrupulously and efficiently the sequences of the nitrotyrosine sites and extract features from the peptide segment. Based on the Chou’s conception, a tyrosine residue centered peptide can be represented by:

$$\Theta_{\zeta}(Y) = R_{-\zeta}R_{-(\zeta-1)}\ldots R_{-2}R_{-1}YR_{1}R_{2}\ldots R_{+(\zeta-1)}R_{+\zeta} \tag{1}$$

In this equation, $R_{-\zeta}$ and $R_{+\zeta}$ denote the ζ-th leftward and rightward amino acid residues, respectively, while ζ being an integer and ‘Y’(center) indicating “Tyrosine” (Ahmed et al., 2021).Again, the peptide sequence $\Theta_{\zeta}(Y)$ is categorized into two types: $\Theta_{\zeta}^{+}(Y), \Theta_{\zeta}^{-}(Y)$ are true nitrated peptide and false nitrated peptide with a tyrosine residue at its center (Rahman et al., 2020; Ahmed et al., 2021). To segment the nitrotyrosine protein sequences, the sliding window method was adopted. According (Ghauri et al., 2018) which introduced pNitro-Tyr-PseAAC, using ζ=20 as the window size obtained the best results. Therefore, we use the same window size in this study meaning that the corresponding peptide segment contained $(2\zeta+1) = 41$ amino acid residues. With a sequence fragment of window size 41, Eq. 1 can be presented as:

$$\Theta_{20}(Y) = Q_{1}Q_{2}\ldots Q_{19}Q_{20}YQ_{21}Q_{22}\ldots Q_{39}Q_{40} \tag{2}$$

In the segmentation process, the absent amino acids are filled with dummy residues denoted by ’X’ as it was discussed in Xu et al. (2015), Ahmed et al. (2021) for the processing of site sequences of identical length. Therefore, the nitrotyrosine dataset is taken the following pattern:

$$S_{\zeta}(\mathrm{Y}) = S_{\zeta}^{+}(\mathrm{Y}) \cup S_{\zeta}^{-}(\mathrm{Y}) \tag{3}$$

where the positive subset $S_{\zeta}^{+}(\mathrm{Y})$ could contain only $\Theta_{\zeta}^{+}(Y)$ samples, while the negative subset $S_{\zeta}^{-}(\mathrm{Y})$ could contain only $\Theta_{\zeta}^{-}(Y)$ samples with their center residue Y. In this study, the vectorized sequence-coupled model has been adopted to extract features from the tyrosine nitrated sites eliciting the sequence pattern details (Chou, 1993; Ahmed et al., 2021). According to Chou’s general PseAAC (Chou, 2011), the peptide sample in (2), can be presented as:

$$\Theta(Y) = \Theta^{+}(Y) - \Theta^{-}(Y) \tag{4}$$

where

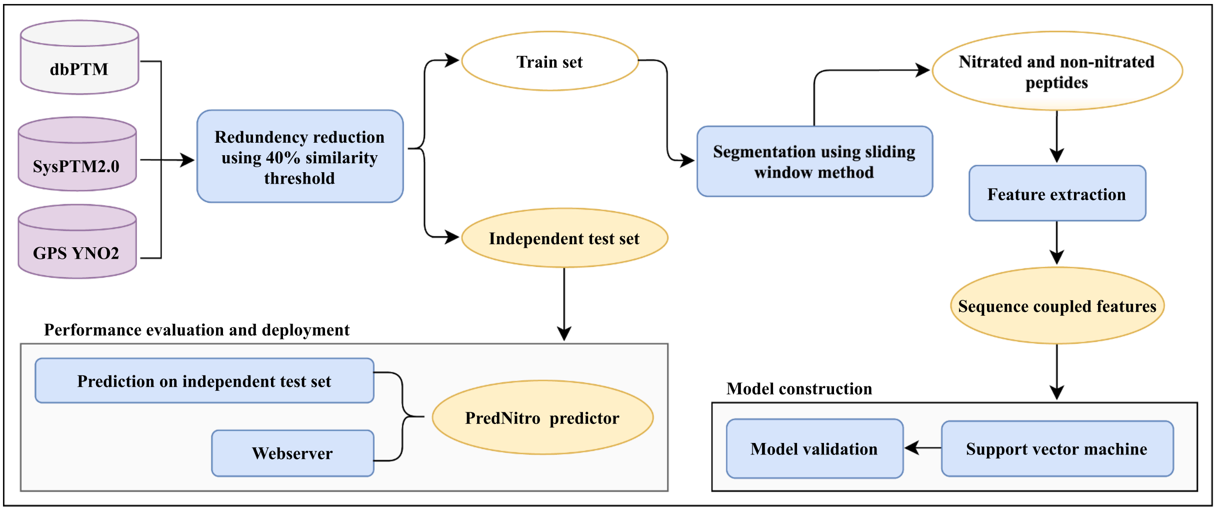
dbPTM
SysPTM2.0
GPS YNO2
Redundency reduction using 40% similarity threshold
Train set
Independent test set
Segmentation using sliding window method
Nitrated and non-nitrated peptides
Feature extraction
Sequence coupled features
Model construction
Model validation
Support vector machine
Performance evaluation and deployment
Prediction on independent test set
Webserver
PredNitro predictor

Fig. 2. The general architecture of PredNitro. The nitrotyrosine dataset has been collected from three different databases and splitted it into train set and independent test set. Finally, a new machine learning tool has been developed utilizing the sequence-coupled information and support vector machine called PredNitro to predict nitrotyrosine sites in proteins.

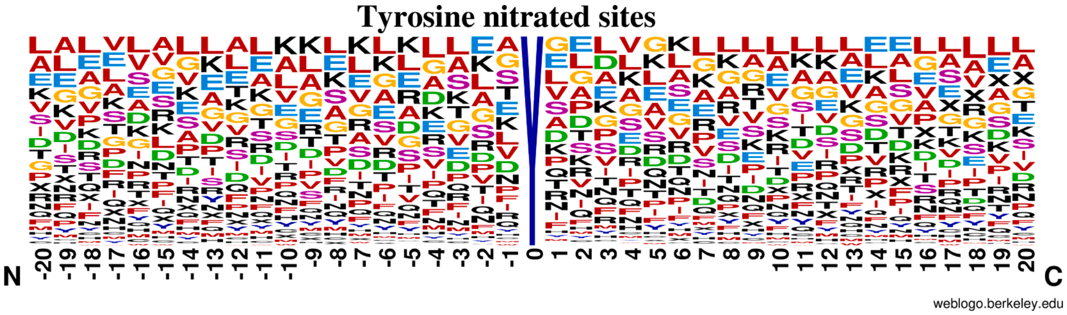

Fig. 3. Frequency plot of positive nitrotyrosine samples.

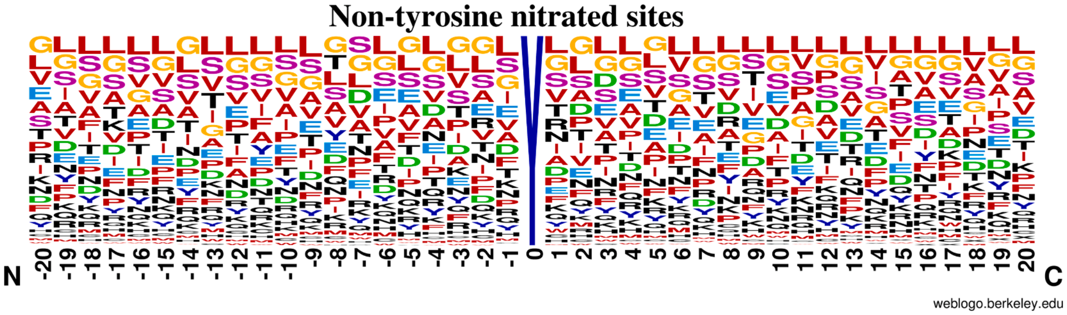

Fig. 4. Frequency plot of negative nitrotyrosine samples.

Table 2
Performances for different classification techniques on the benchmark dataset.

Predictor	Cross-validation Technique	Sn (%)	Sp (%)	ACC (%)	MCC (%)	AUC (%)
KNN	Jackknife	98.07	95.55	96.81	93.65	98.92
	10-Fold	**98.10**	95.56	96.83	93.69	98.96
	5-Fold	97.95	95.53	96.74	93.51	99.00
RF	Jackknife	95.97	97.15	96.56	93.12	98.81
	10-Fold	96.13	97.04	96.58	93.17	98.80
	5-Fold	95.8	96.84	96.32	92.65	98.83
SVM	Jackknife	97.73	98.32	98.03	96.06	**99.48**
	10-Fold	97.74	**98.40**	**98.07**	**96.15**	**99.48**
	5-Fold	97.67	98.34	98.00	96.01	**99.48**

Table 3
Jackknife cross-validation performance comparison on different datasets.

Predictor	Threshold	Sn (%)	Sp (%)	ACC (%)	MCC (%)	AUC (%)
GPS-YNO2*	High	28.89	90.02	82.57	18.84	-
	Medium	40.53	85.02	79.60	21.71	-
	Low	50.09	80.18	76.51	23.35	-
iNitro-Tyr*	-	81.76	85.89	84.52	49.05	-
pNitro-Tyr-PseAAC*	-	85.64	89.40	88.09	62.70	-
Proposed Method[a]	-	93.05	93.86	93.75	77.07	98.20
Proposed Method[b]	-	97.73	98.32	98.03	96.06	99.48

[a] corresponds to the performance derived from the benchmark dataset used in Ghauri et al. (2018).
[b] corresponds to the performance on the benchmark dataset used in Nilamyani et al. (2021).
* corresponds to the performance reported by Ghauri et al. (2018).

Table 4
K-fold cross-validation performance comparison on different datasets.

Predictor	Threshold	Sn(%)	Sp(%)	ACC (%)	MCC (%)	AUC (%)
pNitro-Tyr-PseAAC	-	84.00	93.02	89.10	-	-
Proposed Method[a]	-	93.26 ± 0.12	93.77 ± 0.06	93.70 ± 0.06	77.00 ± 0.20	98.20 ± 0.00
DeepNitro	High	17.70	95.00	84.90	17.20	-
	Medium	29.10	90.00	82.00	19.50	-
	Low	38.50	85.00	78.90	20.60	-
Proposed Method[b]	-	93.46 ± 0.10	93.86 ± 0.02	93.81 ± 0.02	77.52 ± 0.07	98.29 ± 0.01
PredNTS	-	-	-	-	-	91.00
Proposed Method[c]	-	97.67 ± 0.11	98.34 ± 0.11	98.00 ± 0.06	96.01 ± 0.13	99.48 ± 0.00

[a,b] correspond to the 10-fold cross-validation performance derived from the respective dataset used in Ghauri et al. (2018),Xie et al. (2018).
[c] corresponds to the 5-fold cross-validation performance on the dataset used in Nilamyani et al. (2021).

Table 5
Independent test performance comparison with the existing predictors.

Predictor	Sn(%)	Sp(%)	ACC(%)	MCC(%)	AUC(%)
GPS-YNO2	33.40	80.10	72.40	12.20	-
DeepNitro	33.90	80.30	72.60	12.80	-
NTyroSite	44.00	79.30	74.40	19.60	-
PredNTS	52.20	80.90	76.10	28.60	86.00
PredNitro	**100.00**	**88.16**	**90.12**	**74.32**	**99.59**

$$\Theta^{+}(Y)=\begin{bmatrix}\Theta_1^{+}(Q_1|Q_2)\\ \Theta_2^{+}(Q_2|Q_3)\\ \vdots\\ \Theta_{19}^{+}(Q_{19}|Q_{20})\\ \Theta_{20}^{+}(Q_{20})\\ \Theta_{21}^{+}(Q_{21})\\ \Theta_{22}^{+}(Q_{22}|Q_{21})\\ \vdots\\ \Theta_{39}^{+}(Q_{39}|Q_{38})\\ \Theta_{40}^{+}(Q_{40}|Q_{39})\end{bmatrix} \tag{5}$$

$$\Theta^{-}(Y)=\begin{bmatrix}\Theta_1^{-}(Q_1|Q_2)\\ \Theta_2^{-}(Q_2|Q_3)\\ \vdots\\ \Theta_{19}^{-}(Q_{19}|Q_{20})\\ \Theta_{20}^{-}(Q_{20})\\ \Theta_{21}^{-}(Q_{21})\\ \Theta_{22}^{-}(Q_{22}|Q_{21})\\ \vdots\\ \Theta_{39}^{-}(Q_{39}|Q_{38})\\ \Theta_{40}^{-}(Q_{40}|Q_{39})\end{bmatrix} \tag{6}$$

where $\Theta_1^{+}(Q_1|Q_2)$ is the conditional probability of amino acid Q_1 at the leftmost position given that its adjacent right member is Q_2 and so forth (Ahmed et al., 2021). Similarly, $\Theta_{40}^{+}(Q_{40}|Q_{39})$ denotes the conditional probability of amino acid Q_{40} at the rightmost position given that its adjacent left member is Q_{39} and so on. In contrast, only $\Theta_{20}^{+}(Q_{20})$ and $\Theta_{21}^{+}(Q_{21})$ are of non-conditional probability as Y is the adjoining member of both amino acids at position Q_{20} and Q_{21} (Chou, 1993; Ahmed et al., 2021; Rahman et al., 2020; Ahmed et al., 2021; Ahmed et al., 2021). The probability values can be extracted from the set of nitrated peptides using the frequency of a given acid corresponding to their positions. Accordingly, $\Theta^{-}(Y)$ in (4), and its probability components can be deduced from the non-nitrated peptide set in the same way as shown in (6). Finally, a 40-dimensional feature vector was obtained by using Eqs. ()()()(4)–(6) for each potential nitrated and non-nitrated sample.

In order to facilitate visualization and insight into the sequence-coupling effects at various places in each sample, we have stored all conceivable combinations of conditional probability values extracted from the positive training subset i.e. $\Theta^{+}(Q_1|Q_2)$ to $\Theta^{+}(Q_{19}|Q_{20})$ and $\Theta^{+}(Q_{22}|Q_{21})$ to $\Theta^{+}(Q_{40}|Q_{39})$ in one data frame and non-conditional probability values for each amino acid residue retrieved from the positive training subset i.e. $\Theta^{+}(Q_{20})$ and $\Theta^{+}(Q_{21})$ in another data frame, where the columns represent the formulated sample positions and the rows represent the amino acid residues. It should be noted that there are $21 \times 21 = 441$ different combinations of conditional probability values and 21 non-conditional probability values for each position in any formulated sample (including the dummy amino acid residue $'X'$) (Chou,

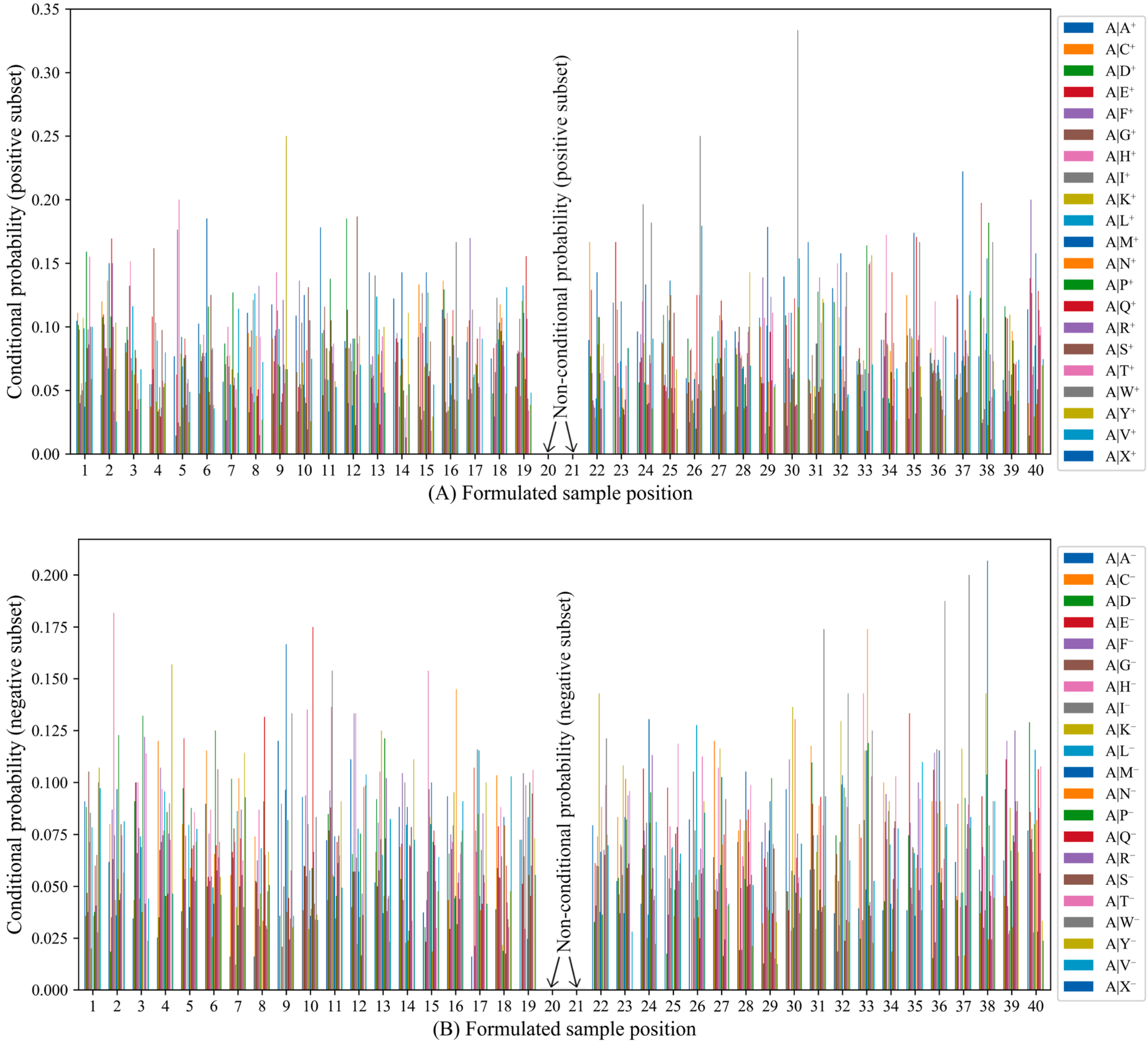

Fig. 5. (A) The conditional probability values of amino acid residue $'A'$ which have been calculated from the positive subset, given that its right member is any of the 21 amino acid residues at sample positions 1 to 19 and the conditional probability values of any of the 21 amino acid residue given that the left member is $'A'$ at sample positions 22 to 40. (B) The conditional probability values of amino acid residue $'A'$ which have been calculated from the negative subset, given that its right member is any of the 21 amino acid residues at sample positions 1 to 19 and the conditional probability values of any of the 21 amino acid residue given that the left member is $'A'$ at sample positions 22 to 40.

1993). Likewise, the conditional and non-conditional probability values derived from the negative subset are stored in two distinct data frames. All the corresponding conditional and non-conditional probability values retrieved from the positive and negative subsets are provided in the supplementary material S1.

2.3. Classification algorithm

During the last decade, different types of machine learning models such as random forest (Lv et al., 2020; Shi et al., 2019), logistic regression (Dai et al., 2021), stacking method (Bin et al., 2020), k-nearest neighbor (Wang et al., 2020), and neural network (Xie et al., 2018) had been used to predict the nitration sites (Ghauri et al., 2018, 2011, 2018, 2014, 2018). However, support vector machine (SVM) which is considered as one of the state-of-the-at classification techniques have never been used for this task. SVM aim at enhancing the classification performance by finding the maximal marginal hyperplane (MMH) to separate different classes (Cortes and Vapnik, 1995; Zhang et al., 2019). SVM is widely used in the literature to predict other PTMs and obtained promising results (Ahmed et al., 2021; Ahmed et al., 2021; Ahmed et al., 2021; Rahman et al., 2020; Chandra et al., 2020; Singh et al., 2020; Chandra et al., 2019; Reddy et al., 2019; Chandra et al., 2019). An SVM is designed to solve the following constrain minimization problem:

$$min^{\frac{1}{2}}_{w,\xi}\|w\|^2 + C^+\sum_{k=1}^{q}\xi_k + C^-\sum_{k=q+1}^{n}\xi_k \tag{7}$$

(Subject to: $Y_k(w.\varphi(X_k)+a)\geqslant 1-\xi_k$ for all, $k=1,2,..,n$)where the training set is denoted by $\{(X_k,Y_k),k=1,2,\ldots,n\}$ and first q examples (i.

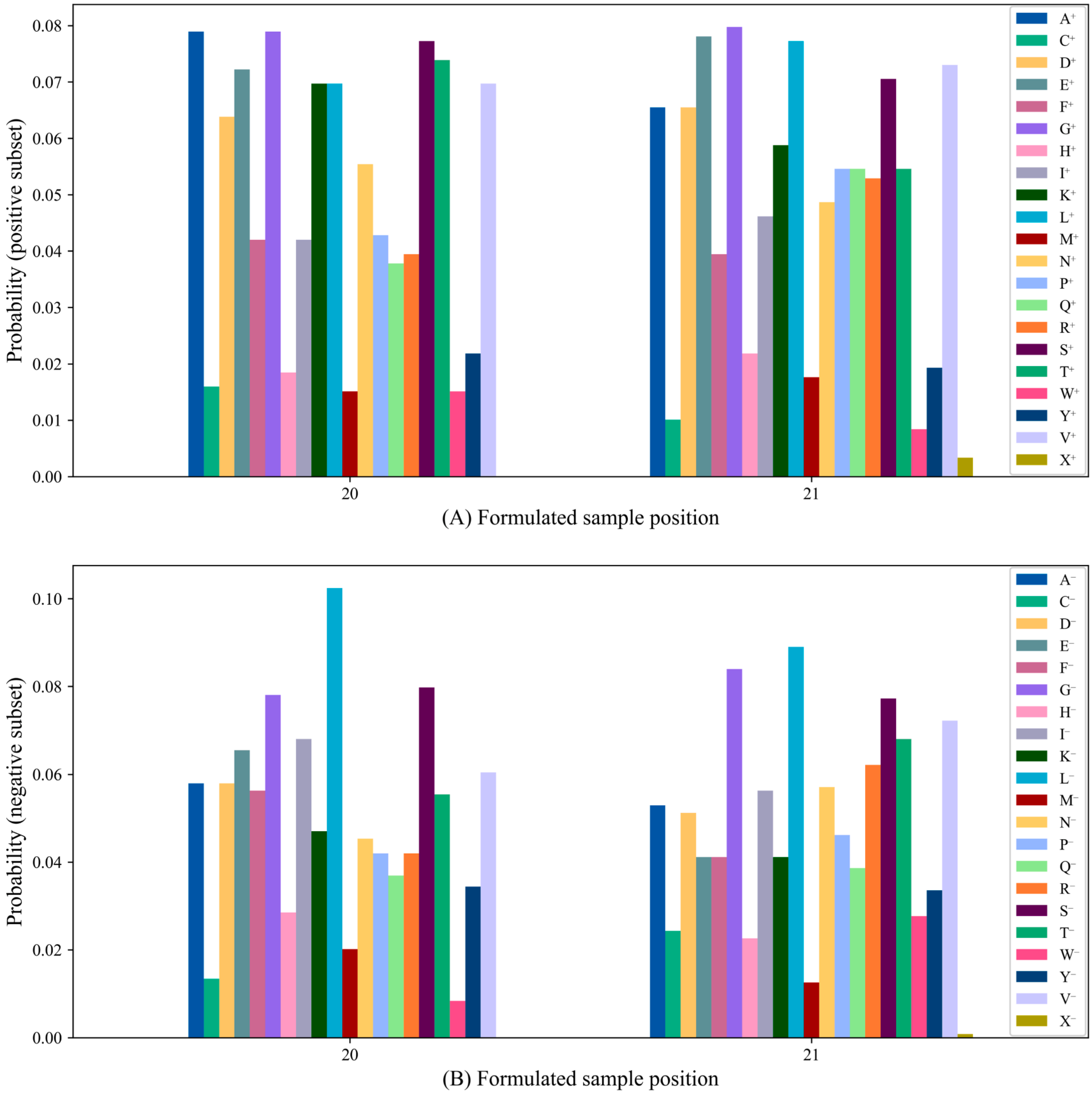

Fig. 6. (A) The non-conditional probability values of 21 amino acid residues derived from the positive subset at sample positions 20 and 21. (B) The non-conditional probability values of 21 amino acid residues derived from the negative subset at sample position 20 and 21.

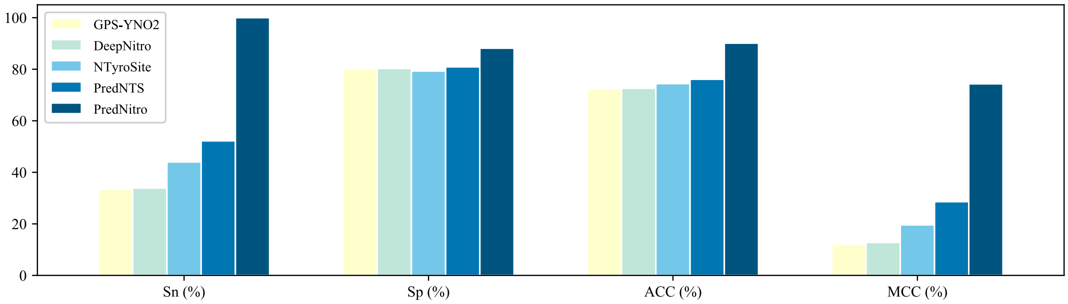

Fig. 7. Performance comparison between PredNitro and the existing methods based on the independent test.

e. $Y_k = 1, k = 1,2,\ldots,q$) are assumed as the positive examples while the rest are assumed as the negative examples (i.e. $Y_k = -1, k = q+1, q+2,\ldots,n$). The non-linear feature mapping and slack variables are denoted by $\varphi(X)$ and $\xi_k (k = 1,2,\ldots,n)$ respectively (Ju and Wang, 2020). In our experiments, the Gaussian kernel function is described as: $\Upsilon(X_k, X_j) = \varphi(X_k)^T\varphi(X_j) = exp(-\frac{\|X_k - X_j\|^2}{\partial})$, where ∂ is the width of the function.

2.4. Prediction metrics

To investigate the prediction quality of PredNitro, we have utilized four intuitive evaluation metrics, such as accuracy (ACC), sensitivity (Sn), specificity (Sp), and Matthew's Correlation Coefficient (MCC) which have been widely used in the literature for this task (Dehzangi et al., 2018; Dehzangi et al., 2015; Ahmed et al., 2021; Rahman et al., 2020; Chandra et al., 2019). These performance metrics can be calculated as follows:

$$Sn = \frac{TP}{TP + FN} \tag{8}$$

$$Sp = \frac{TN}{TN + FP} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

Where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. Here, we also calculate the area under the ROC curve (AUC) and MCC to check the stability and robustness of our system (Ahmed et al., 2021; Rahman et al., 2020).

2.5. Jackknife cross-validation test

Jackknife cross-validation is widely used to evaluate the performance of different models to tackle this problem test (Liu et al., 2011; Ghauri et al., 2018). The jackknife test can provide a unique outcome for a given data set, and it is highly beneficial to reduce the computational sophistication of model creation (Li et al., 2016). Also, the arbitrariness problem and the 'memory' influence can be addressed when using this test (Chou, 2011; Ghauri et al., 2018). In this mode, if the total sample number of the cross-validation is *n*, each sample will be used as a test set once, while the remaining *n-1* samples will be used as training set (Li et al., 2016). In this way, the overall prediction metrics are calculated for each population mixture of size *n-1* using iterative process (Qiu et al., 2014; Liu et al., 2015; Khan et al., 2014). After the completion of N estimation turns, the averaged performance of test sets is calculated as the ultimate outcome for the jackknife cross-validation.

2.6. K-fold cross-validation test

It is often desirable to obtain an appropriate model that can be capable of delivering high efficiency. Although jackknife test is widely used in the literature for this purpose, now-a-days researchers prefer k-fold cross-validation approach over the jackknife test for verifying their PTM prediction model to reduce the computing time of the model creation (Ju et al., 2016; Ju and Wang, 2020). Therefore, here k-fold cross-validation test is also implemented to evaluate our proposed predictor. We performed this model *M* times and reported the average results to guarantee the stability of our reported results. The M-iterations of k-fold cross-validation were performed according to the following steps:

Step 1. Divide the benchmark dataset randomly into k disjoint sets.
Step 2. Select 1 set as test set and utilize the remaining k-1 sets as training set.
Step 3. Train the predictor with the training set using the LibSVM's default parameters ($C = 1, \gamma = 1/\text{numberoffeatures}$).
Step 4. Perform prediction on the test set.
Step 5. Repeat steps 2 to 4 until all k sets had been used for testing.
Step 6. Merge the prediction outputs and measure the performance with Eq. 7.
Step 7. Repeat steps 1 to 6 for M times.
Step 8. Measure the average performance of M repetitions with corresponding standard deviations.

Several tyrosine nitration prediction systems i.e. pNitro-Tyr-PseAAC (Ghauri et al., 2018), and DeepNitro (Xie et al., 2018) have validated their model using k-fold cross-validation approach where k = 10. The most recent predictor PredNTS (Nilamyani et al., 2021) has utilized 5-fold cross-validation technique for model evaluation. Therefore, we adopted both 5-fold and 10-fold cross-validation to be able to directly compare our results with those reported in the literature. The predictive decision-making workflow of PredNitro is publicly available at https://github.com/Sabit-Ahmed/PredNitro as a GitHub repository.

2.7. Independent test

Existing nitrotyrosine site predictors, particularly, the most recent system PredNTS (Nilamyani et al., 2021) assessed their model using 5-fold cross-validation. However, most researchers emphasize on the necessity of independent test for assessing prediction model in addition to k-fold (e.g. k = 5,10) cross-validation (Xie et al., 2018; Ghauri et al., 2018; Nilamyani et al., 2021). Thus, here an independent test consisting of 203 experimentally annotated positive nitrotyrosine sites and 1022 negative nitrotyrosine sites is generated for further evaluation of our proposed model. It should be mentioned that the proteins in the independent test set have not been used for any parameter tuning.

3. Results and Discussions

3.1. Performance of PredNitro

In this study, the sequence-coupling features have been extracted from the benchmark dataset and the support vector machine (Hasan et al., 2017; Hasan and Ahmad, 2018) has been used as prediction algorithm. We also use Radial Basis Function (RBF) as SVM kernel which demonstrates better results than using other kernels (Hasan and Ahmad, 2018; Hasan et al., 2017). To ascertain the statistical importance of the predicted results of PredNitro, three validation techniques, such as k-fold cross-validation (k = 5 and 10), jackknife test, and an independent test, are extensively used (Hasan and Ahmad, 2018; Ahmed et al., 2021). Therefore, our proposed model is evaluated with different types of validation schemes based on the corresponding datasets used in GPS-YNO2, pNitro-Tyr-PseAAC, DeepNitro, and PredNTS studies. Our results demonstrates that PredNitro can predict nitrotyrosine sites with more than 98.0% accuracy and 96.0% MCC on both jackknife test and k-fold cross-validation schemes. In addition, its sensitivity, specificity and AUC measure crossed a benchmark of 97.0% on both cross-validation methods. Furthermore, it achieves 90.12% accuracy with 99.59% AUC on the independent test. In other words, PredNitro consistently outperforms previous studies using all three evaluation methods which demonstrate the generality of this model. The ROC curves of PredNitro in the training and independent test sets are available in supplementary material S2.

3.2. Performance analysis of different classification techniques

Currently, a wide range of machine learning techniques are available, which can effectively classify any nitrotyrosine site containing sample (Ghauri et al., 2018; Hasan et al., 2018; Xu et al., 2014; Xie et al., 2018; Nilamyani et al., 2021). As our benchmark dataset is comparatively small, we have intended to experiment with a several less data demanding machine learning algorithms such as, random forest (Lv et al., 2020; Shi et al., 2019), k-nearest neighbor (Wang et al., 2020), and support vector machine (Ahmed et al., 2021; Ahmed et al., 2021; Ahmed et al., 2021; Rahman et al., 2020). By utilizing the sequence-coupled features, we have performed the jackknife test, 5-fold and 10-fold cross-validation tests on the benchmark dataset to find the best-suited model for constructing our proposed predictor. From Table 2, it can be observed that the random forest method has obtained the lowest performances in all the cross-validation tests. On the contrary, the k-nearest neighbour algorithm has achieved the highest sensitivity rate of 98.10% in the 10-fold cross-validation. Moreover, its attained sensitivity rates in other validation tests are higher than any other classification method. However, the support vector machine has obtained the highest specificity, accuracy, MCC and AUC measures in all three types of cross-validation tests. Analyzing the outcomes of the three most commonly used classifiers, we have constructed our final prediction system with the SVM classifier because of its high-performance statistics.

3.3. Comparative analysis with existing predictors

At present, there are six main predictors available to predict the nitrotyrosine sites, such as, GPS-YNO2 (Liu et al., 2011), iNitro-Tyr (Xu et al., 2014), pNitro-Tyr-PseAAAC (Ghauri et al., 2018), DeepNitro (Xie et al., 2018), NTyroSite (Hasan et al., 2018), and PredNTS (Nilamyani et al., 2021) for nitrotyrosine site prediction. Particularly, each of these prediction systems has constructed its curated dataset for performance benchmarking. GPS-YNO2, and iNitro-Tyr have used jackknife test while pNitro-Tyr-PseAAC has adopted both the jackknife and 10-fold cross-validation test for model evaluation. DeepNitro and NTyroSite have applied 10-fold cross-validation schemes. On the other hand, PredNTS has validated their model with 5-fold cross-validation technique. For a fair comparison, we have utilized different datasets and validation criteria. The prediction outcome from the jackknife test, k-fold cross-validation, and independent test has been measured with the evaluation metrics described in Eqs. (8)–(12) and reported in Tables 3–5, respectively while corresponding standard deviations where applicable.

As shown in Table 3, the PredNitro achieves significantly better results compared to GPS-YNO2, iNitro-Tyr, pNitro-Tyr-PseAAC in terms of all metrics (i.e accuracy, MCC, sensitivity, specificity) using jackknife cross validation test. For instance, PredNitro outperformed the most recent predictor, pNitro-Tyr-PseAAC, by 7.41% in term of sensitivity, 4.46% in term of specificity, 5.66% in term of accuracy, and 14.37% in term of MCC on the same dataset that they used. AUC which is one of the most important measures has reached above 98.0%. It is important to note that higher sensitivity achieved by our predictor demonstrates that PredNitro is able to predict positive samples significantly better than previous studies.

To scrutinize the results, we further implemented 10-fold cross-validation test for 10 times on the dataset provided by pNitro-Tyr-PseAAC and DeepNitro studies, and again achieved significantly better results compared to these two methods. As shown in Table 4, results obtained by PredNitro are almost similar to the jackknife test performances. This demonstrates the generality of our model for this task. In addition, we have applied 5-fold cross-validation 5 times on the becnhmark dataset used in PredNTS study. When comparing our proposed method with the PredNTS predictor, it can be observed that, the AUC measure has increased from 91.0% to 99.48%. Furthermore, the reported jackknife test results (see Table 3 on the same benchmark dataset are also promising and identical to that of 5-fold cross-validation results. Our proposed method attains over 96.0% for all the prediction metrics.

We also achieve similar results for our independent test set. To conduct a fair comparison, the independent dataset was uploaded to the web-servers of the existing state-of-art predictors (i.e. GPS-YNO2, DeepNitro, NTyroSite and PredNTS) to obtain the prediction outcomes. The predictive performance of PredNitro as well as other predictors are summarized in Table 5 and Fig. 7. As shown in Table 5 and Fig. 7, PredNitro again achieves better results than any other existing predictors. More precisely, it enhance the sensitivity, specificity, accuracy, MCC, and AUC for 47.80%, 7.26%, 14.02% 45.75%, and 13.59% compared to PredNTS as the most recent successful predictor, respectively. Again, significant improvement in sensitivity demonstrates the ability of PredNitro in identifying tyrosine nitration sites. Results represented in Table 5 and Fig. 7 indicate that our proposed predictor PredNitro can be a high throughput tool for the effective identification of the unknown tyrosine nitration sites.

Our results demonstrates that by using effective representation of nitrotyrosine modification in terms of sequence coupling model among the amino acid residues via the conditional probability as well as SVM as our classification technique we are able to significantly enhance tyrosine nitration sites prediction task compared to previous studies (see Fig. 5 and 6).

3.4. Web-server

To increase user accessibility without requiring experimental solutions, we designed an easy-to-use web server for PredNitro which is publicly available at: http://103.99.176.239/PredNitro. Users can enter one or more query protein sequences as text input in Fasta format directly on the web server, or they can upload as a batch to acquire their predictions. There are also more thorough instructions on how to operate the web server as well as the server's operating mechanism. Depending on the availability of server resources, getting the prediction result after submitting a query protein or as a batch may take a few moments. Finally, PredNitro will generate a result page based on the user's input. For example, if protein sequences are entered into the input box, the predictive data will appear on the result page. Otherwise, an email will be sent to the appropriate user.

4. Conclusion

The prediction of nitrotyrosine sites is critically important for attaining a better perception of biological systems. In this study, we develop a novel machine learning tool named PredNitro to accurately predict tyrosine nitrated sites by using a vectorized sequence-coupling model with SVM classifier. By adopting sequence-coupling effect with misclassification cost adjustment, PredNitro acquired extraordinarily higher prediction accuracy compared to the existing nitrotyrosine site predictors. Both in the k-fold cross-validation and jackknife cross-validation test, it obtained a significant improvement in MCC as well as in other crucial metrics (approximately 98.0%, 97.0%, and 0.99 in terms of accuracy, sensitivity, and AUC) which ensures the generality and robustness of our predictor. PredNitro is publicly available as an online predictor at: http://103.99.176.239/PredNitro.

Contributors

A. Rahman, S. Ahmed designed and performed the experiments. A. Rahman, S. Ahmed, and A. Dehzangi wrote the manuscript and validated the results. A. Rahman prepared figures. M. A. M. Hasan and I. Dehzangi mentored and analytically reviewed the paper. S. Ahmad provided the resources for the web-server. All the authors reviewed the article.

Data availability statement

PredNitro as an online predictor and our employed benchmarks are publicly available online at: http://103.99.176.239/PredNitro.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gene.2022.146445.

References

Abello, N., Barroso, B., Kerstjens, H.A., Postma, D.S., Bischoff, R., 2010. Chemical labeling and enrichment of nitrotyrosine-containing peptides. Talanta 80 (4), 1503–1512.

Ahmed, S., Rahman, A., Hasan, M.A.M., Islam, M.K.B., Rahman, J., Ahmad, S., 2021. predPhogly-Site: Predicting phosphoglycerylation sites by incorporating probabilistic sequence-coupling information into PseAAC and addressing data imbalance. Plos One 16 (4), e0249396.

Ahmed, S., Rahman, A., Hasan, M.A.M., Rahman, J., Islam, M.K.B., Ahmad, S., 2021. predML-Site: Predicting Multiple Lysine PTM Sites with Optimal Feature Representation and Data Imbalance Minimization. IEEE/ACM Trans. Comput. Biol. Bioinform. (01), 1–1.

Ahmed, S., Rahman, A., Hasan, M., Mehedi, A., Ahmad, S., Shovan, S., 2021. Computational identification of multiple lysine PTM sites by analyzing the instance hardness and feature importance. Scient. Rep. 11 (1), 1–12.

Bin, Y., Zhang, W., Tang, W., Dai, R., Li, M., Zhu, Q., Xia, J., 2020. Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. J. Proteome Res. 19 (9), 3732–3740.

Blantz, R.C., Munger, K., 2002. Role of nitric oxide in inflammatory conditions. Nephron 90 (4), 373–378.

Brindicci, C., Kharitonov, S.A., Ito, M., Elliott, M.W., Hogg, J.C., Barnes, P.J., Ito, K., 2010. Nitric oxide synthase isoenzyme expression and activity in peripheral lung tissue of patients with chronic obstructive pulmonary disease. Am. J. Respirat. Crit. Care Med. 181 (1), 21–30.

Chandra, A., Sharma, A., Dehzangi, A., Ranganathan, S., Jokhan, A., Chou, K.-C., Tsunoda, T., 2018. Phoglystruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Scient. Rep. 8 (1), 1–11.

Chandra, A.A., Sharma, A., Dehzangi, A., Tsunoda, T., 2019. Evolstruct-phogly: incorporating structural properties and evolutionary information from profile bigrams for the phosphoglycerylation prediction. BMC Genom. 19 (9), 1–9.

Chandra, A., Sharma, A., Dehzangi, A., Shigemizu, D., Tsunoda, T., 2019. Bigram-pgk: phosphoglycerylation prediction using the technique of bigram probabilities of position specific scoring matrix. BMC Mol. Cell Biol. 20 (2), 1–9.

Chandra, A.A., Sharma, A., Dehzangi, A., Tsunoda, T., 2020. Ram-pgk: Prediction of lysine phosphoglycerylation based on residue adjacency matrix. Genes 11 (12), 1524.

Chou, K.-C., 1993. A vectorized sequence-coupling model for predicting hiv protease cleavage sites in proteins. J. Biol. Chem. 268 (23), 16938–16948.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theoret. Biol. 273 (1), 236–247.

Chou, K.-C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11 (3), 218–234.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14 (6), 1188–1190.

Dai, R., Zhang, W., Tang, W., Wynendaele, E., Zhu, Q., Bin, Y., De Spiegeleer, B., Xia, J., 2021. Bbppred: sequence-based prediction of blood-brain barrier peptides with feature representation learning and logistic regression. J. Chem. Inf. Model. 61 (1), 525–534.

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015. Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into chou's general pseaac. J. Theoret. Biol. 364, 284–294.

Dehzangi, A., Lopez, Y., Lal, S.P., Taherzadeh, G., Sattar, A., Tsunoda, T., Sharma, A., 2018. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PloS one 13 (2).

Donnini, S., Monti, M., Roncone, R., Morbidelli, L., Rocchigiani, M., Oliviero, S., Casella, L., Giachetti, A., Schulz, R., Ziche, M., 2008. Peroxynitrite inactivates human-tissue inhibitor of metalloproteinase-4. FEBS Lett. 582 (7), 1135–1140.

Ghauri, A.W., Khan, Y.D., Rasool, N., Khan, S.A., Chou, K.-C., 2018. pnitro-tyr-pseaac: predict nitrotyrosine sites in proteins by incorporating five features into chou's general pseaac. Current pharmaceutical design 24 (34), 4034–4043.

Giasson, B.I., Duda, J.E., Murray, I.V., Chen, Q., Souza, J.M., Hurtig, H.I., Ischiropoulos, H., Trojanowski, J.Q., Lee, V.M.-Y., 2000. Oxidative damage linked to neurodegeneration by selective α-synuclein nitration in synucleinopathy lesions. Science 290 (5493), 985–989.

Hasan, M.A.M., Ahmad, S., 2018. mLysPTMpred: Multiple Lysine PTM Site Prediction Using Combination of SVM with Resolving Data Imbalance Issue. Natural Science 10 (9), 370–384.

Hasan, M.A.M., Ahmad, S., Molla, M.K.I., 2017. iMulti-HumPhos: a multi-label classifier for identifying human phosphorylated proteins using multiple kernel learning based support vector machines. Mol. BioSyst. 13 (8), 1608–1618.

Hasan, M.A.M., Li, J., Ahmad, S., Molla, M.K.I., 2017. predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue. Analytical biochemistry 525, 107–113.

Hasan, M., Khatun, M., Mollah, M., Haque, N., Yong, C., Dianjing, G., et al., 2018. Ntyrosite: Computational identification of protein nitrotyrosine sites using sequence evolutionary features. Molecules 23 (7), 1667.

Ju, Z., Wang, S.-Y., 2020. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. Genomics 112 (1), 859–866.

Ju, Z., Cao, J.-Z., Gu, H., 2016. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. 397, 145–150.

Khan, Y.D., Ahmed, F., Khan, S.A., 2014. Situation recognition using image moments and recurrent neural networks. Neural Comput. Appl. 24 (7), 1519–1529.

Lee, T.-Y., Huang, H.-D., Hung, J.-H., Huang, H.-Y., Yang, Y.-S., Wang, T.-H., 2006. dbptm: an information repository of protein post-translational modification. Nucleic acids research 34 (suppl_1), D622–D627.

Li, S., Lu, J., Li, J., Chen, X., Yao, X., Xi, L., 2016. Hydpred: a novel method for the identification of protein hydroxylation sites that reveals new insights into human inherited disease. Mol. BioSyst. 12 (2), 490–498.

Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J., Xue, Y., 2011. Gps-yno2: computational prediction of tyrosine nitration sites in proteins. Mol. BioSyst. 7 (4), 1197–1204.

Liu, Z., Xiao, X., Qiu, W.-R., Chou, K.-C., 2015. idna-methyl: Identifying dna methylation sites via pseudo trinucleotide composition. Analytical biochemistry 474, 69–77.

Lv, Z., Zhang, J., Ding, H., Zou, Q., 2020. Rf-pseu: A random forest predictor for rna pseudouridine sites. Front. Bioeng. Biotechnol. 8.

McDowell, G., Philpott, A., 2016. New insights into the role of ubiquitylation of proteins. In: International review of cell and molecular biology, Vol. 325, Elsevier, 2016, pp. 35–88.

Nilamyani, A.N., Auliah, F.N., Moni, M.A., Shoombuatong, W., Hasan, M.M., Kurata, H., 2021. Prednts: Improved and robust prediction of nitrotyrosine sites by integrating multiple sequence features. International journal of molecular sciences 22 (5), 2704.

Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C., 2014. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. BioMed research international.

Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C., Chou, K.-C., 2016. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32 (20), 3116–3123.

Qiu, W.-R., Jiang, S.-Y., Sun, B.-Q., Xiao, X., Cheng, X., Chou, K.-C., 2017. iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Med. Chem. 13 (8), 734–743.

Rahman, A., Ahmed, S., Rahman, J., Hasan, M.A.M., 2020. Prediction of formylation sites by incorporating sequence coupling into general pseaac. in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 921–924.

Reddy, H.M., Sharma, A., Dehzangi, A., Shigemizu, D., Chandra, A.A., Tsunoda, T., 2019. Glystruct: glycation prediction using structural properties of amino acid residues. BMC bioinformatics 19 (13), 55–64.

Saraswathy, N., Ramalingam, P., 2011. Concepts and techniques in genomics and proteomics. Elsevier.

Shi, F., Yao, Y., Bin, Y., Zheng, C.-H., Xia, J., 2019. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. BMC medical genomics 12 (1), 81–88.

Singh, V., Sharma, A., Dehzangi, A., Tsunoda, T., 2020. Pupstruct: Prediction of pupylated lysine residues using structural properties of amino acids. Genes 11 (12), 1431.

Vapnik, V., 2013. The nature of statistical learning theory. Springer science & business media.

Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J., Xu, D., 2020. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization, Nucl. Acids Res.

Weissman, J.D., Raval, A., Singer, D.S., 2003. Assay of an intrinsic acetyltransferase activity of the transcriptional coactivator CIITA. In: Methods in enzymology, Vol. 370. Elsevier, pp. 378–386.

Xie, Y., Luo, X., Li, Y., Chen, L., Ma, W., Huang, J., Cui, J., Zhao, Y., Xue, Y., Zuo, Z., et al., 2018. Deepnitro: prediction of protein nitration and nitrosylation sites by deep learning. Genomics, proteomics & bioinformatics 16 (4), 294–306.

Xu, Y., Wen, X., Wen, L.-S., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2014. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PloS one 9 (8), e105018.

Xu, Y., Ding, Y.-X., Ding, J., Wu, L.-Y., Deng, N.-Y., 2015. Phogly-PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity. J. Theor. Biol. 379, 10–15.

Zhang, L., Tan, B., Liu, T., Sun, X., 2019. Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm. In: Journal of Physics: Conference Series, Vol. 1237, IOP Publishing, 2019, p. 022052.