Molecular BioSystems

Dynamic Article Links

Cite this: *Mol. BioSyst.*, 2011, **7**, 1197–1204

www.rsc.org/molecularbiosystems

PAPER

GPS-YNO2: computational prediction of tyrosine nitration sites in proteins†

Zexian Liu,[ab] Jun Cao,[b] Qian Ma,[b] Xinjiao Gao,[b] Jian Ren*[a] and Yu Xue*[c]

Received 15th November 2010, Accepted 21st December 2010
DOI: 10.1039/c0mb00279h

The last decade has witnessed rapid progress in the identification of protein tyrosine nitration (PTN), which is an essential and ubiquitous post-translational modification (PTM) that plays a variety of important roles in both physiological and pathological processes, such as the immune response, cell death, aging and neurodegeneration. Identification of site-specific nitrated substrates is fundamental for understanding the molecular mechanisms and biological functions of PTN. In contrast with labor-intensive and time-consuming experimental approaches, here we report the development of the novel software package GPS-YNO2 to predict PTN sites. The software demonstrated a promising accuracy of 76.51%, a sensitivity of 50.09% and a specificity of 80.18% from the leave-one-out validation. As an example application, we predicted potential PTN sites for hundreds of nitrated substrates which had been experimentally detected in small-scale or large-scale studies, even though the actual nitration sites had still not been determined. Through a statistical functional comparison with the nitric oxide (NO) dependent reversible modification of S-nitrosylation, we observed that PTN prefers to attack certain fundamental biological processes and functions. These prediction and analysis results might be helpful for further experimental investigation. Finally, the online service and local packages of GPS-YNO2 1.0 were implemented in JAVA and freely available at: http://yno2.biocuckoo.org/.

Introduction

The 1998 Nobel Prize in Physiology or Medicine was awarded to Robert F. Furchgott, Louis J. Ignarro, and Ferid Murad for their pioneering discovery of nitric oxide (NO) as a freely-diffusible second messenger that regulates the production of cyclic GMP (cGMP) in the cardiovascular system. Subsequent studies showed that an interaction between excess NO, transition metal centers and oxidants induces protein tyrosine nitration (PTN).[1–6] In NO metabolism, when oxidants such as superoxide radicals ($O_2^{\bullet -}$) or hydrogen peroxide (H_2O_2) are in transition metal centers (*e.g.*, Fe^{2+}), oxo-metal complexes as well as carbonate radicals ($CO_3^{\bullet -}$) are formed which oxidize tyrosines to tyrosyl radicals, which further react with reactive nitrogen species, such as the peroxynitrite anion ($ONOO^-$) and nitrogen dioxide ($^{\bullet}NO_2$), to yield 3-nitrotyrosine (Fig. 1).[4–6] As a ubiquitous and important post-translational modification (PTM), PTN has been implicated in the regulation of protein activity,[7] epitope recognition,[8] and histone modification.[9] Furthermore, evidence suggests that PTN plays critical roles in both physiological and pathological processes, including the immune response, cell death, aging and neurodegeneration.[1–3] Previously, it was widely accepted that PTN is an irreversible event, because of the absence of denitration enzymes. However, Gow *et al.* found that the nitrotyrosine epitope could be removed in a concentration-, time-, and temperature-dependent manner.[10] Subsequent studies reported that PTN could be reversibly regulated in the response to oxygen tension.[11]

The conventional experimental identification of PTN substrates is inefficient, being both laborious and of low-throughput.[7,12] With the development of antibodies which recognize nitrotyrosine, improved methods for the selective enrichment of nitrotyrosine-containing peptides and the new technology of mass spectrometry, the large-scale detection of cellular nitrated proteins was introduced.[13–15] Subsequently, a number of studies systematically investigated the *in vivo* nitrated proteins so as to provide further insight into the biological roles of PTN.[14,16–21] Initially, Souza *et al.* proposed that there was no consensus sequence around the nitration sites.[22] Recently, Elfering *et al.* suggested a PTN substrate

[a] *Life Sciences School, Sun Yat-sen University (SYSU), Guangzhou, 510275, China. E-mail: renjian.sysu@gmail.com; Fax: +86 20-39943788; Tel: +86 20-39943788*
[b] *Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei, Anhui 230027, China*
[c] *Hubei Bioinformatics and Molecular Imaging Key Laboratory, Department of Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China. E-mail: xueyuhust@gmail.com, xueyu@mail.hust.edu.cn; Fax: +86 27-87793172; Tel: +86 27-87793903*
† Electronic supplementary information (ESI) available: Supplementary tables. See DOI: 10.1039/c0mb00279h

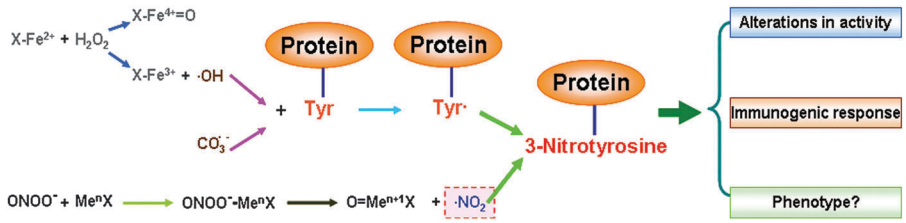

Fig. 1 Biochemical processes of the endogenous NO source and PTN.

motif of [LMVI]-*X*-[DE]-[LMVI]-*X*(2,3)-[FVLI]-*X*(3,5)-*Y* (where *X* is any amino acid and *Y* is the target tyrosine), which can be generalized to *H*-*X*-[DE]-*H*-*X*(2,3)-*H*(2)-*X*(2,4)-*Y* (where *H* represents a hydrophobic residue, such as L, M, V, I, P, A, F, or W).[23] Although numerous studies have made contributions to this area,[15,16,18–21] it is nevertheless still a great challenge to determine the mechanism of PTN. Currently, in addition to the time-consuming and expensive experimental methods, the development of computational approaches has promoted discovery of the PTM sites, since *in silico* prediction is able to rapidly generate useful information for later experimental verification. Although there are ~170 databases and computational tools developed for PTM analysis (http://www.biocuckoo.org/link.php), a program applicable to the prediction of PTN sites is still lacking.

In this work, we collected 1066 experimentally verified PTN sites in 554 unique proteins from the scientific literature and public databases (Table S1, ESI†). Previously, we developed and improved the GPS (Group-based Prediction System) algorithm to predict kinase-specific phosphorylation sites and S-nitrosylation sites.[24,25] Here, the latest GPS algorithm of version 3.0 was applied to predict PTN sites. The leave-one-out validation and 4-, 6-, 8- and 10-fold cross-validations were adopted to evaluate the prediction performance and system robustness. Comparative analysis showed the performance of the GPS 3.0 to be promising, with an accuracy of 76.51%, a sensitivity of 50.09% and a specificity of 80.18%. The novel software package of GPS-YNO2 was then developed to predict PTN sites. As a demonstration application, we also collected hundreds of nitrated substrates from PubMed, for which the *bona fide* nitrated tyrosines had yet to be experimentally determined (Table S2, ESI†). We successfully annotated 325 (~88%) of these proteins having at least one potential PTN site. Furthermore, we systematically compared PTN with another nitric oxide (NO) dependent modification of S-nitrosylation through statistical analyses of their respective gene ontology (GO) terms. It was observed that PTN prefers to attack certain basic biological processes and functions. These predictions and analyses will be useful in future experimental investigation.

Methods

Data preparation

PubMed was searched with the keywords of "nitration" or "nitrated", followed by checking the scientific literature published before July 1st, 2010. A dataset with 1223 experimentally verified PTN sites from 662 proteins were collected. Furthermore, there are also PTN sites in previously constructed public databases, such as dbPTM (50 sites in 39 proteins) and SysPTM (98 sites in 74 proteins).[26,27] All of these datasets were integrated with the protein sequences retrieved from the UniProt database.[28] For the prediction application and functions survey, 370 nitrated substrates from large-scale or small-scale studies were collected, for which the exact PTN sites had not been previously determined (Table S2, ESI†).

As previously described,[24,25] we regarded the nitrated tyrosine (Y) residues as positive data (+), while other tyrosines were taken as negative data (−). It is widely accepted that a redundancy of homologous sites in the positive data (+) leads to overestimated prediction. To avoid such overfitting, we used CD-HIT to cluster the protein sequences,[29] followed by re-alignment with BLAST packages and manual check of the proteins with $\geq 40\%$ identity.[30] The redundant PTN sites at the same position in the homologous proteins according to the alignment result were removed. Finally, the non-redundant data set for training was constructed with 1066 positive sites and 7684 negative sites from 554 unique substrates (Table S1, ESI†). For comparison, the S-nitrosylation dataset was from our previous work.[24]

The algorithms

For prediction of the PTN sites, we employed our recently released GPS 3.0 (Group-based Prediction System) algorithm, which had achieved great success in the prediction of protein S-nitrosylation sites.[24] Based on the hypothesis of similar short peptides bearing similar biochemical properties,[24,25] we defined a *nitration site peptide* NSP(*m*, *n*) as a tyrosine (*Y*) amino acid flanked by *m* residues upstream and *n* residues downstream. Then we scored the similarity of two NSP(*m*, *n*) peptides as:

$$S(A, B) = \sum_{-m \leq i \leq n} \text{Score}(A[i], B[i])$$

Score(*A*[*i*], *B*[*i*]) represented the substitution score of the two amino acid of *A*[*i*] and *B*[*i*] in an amino acid substitution matrix, *e.g.*, BLOSUM62. If $S(A, B) < 0$, we simply redefined it as $S(A, B) = 0$.

For the sake of better performance, we also introduced several performance improvement processes, including *k*-means clustering, motif length selection (MLS), weight training (WT) and matrix mutation (MaM).

(1) *k*-means clustering. Given two NSP(*m*, *n*) peptides *A* and *B*, the similarity was defined and measured as: $s(A, B) = N_s/N$. The *N* is the number of all substitutions, whereas the N_s is the number of conserved substitutions with Score(*a*, *b*) > 0 in the BLOSUM62 matrix. The $s(A, B)$ ranges from 0 to 1.

Thus, the distance between them can be defined as: $D(A, B) = 1/s(A, B)$. If $s(A, B) = 0$, $D(A, B) = \infty$. By exhaustive testing, the k was roughly set to 5, while NSP(7, 7) was adopted. First, five nitration sites from the positive data (+) were randomly chosen as the centroids. Second, the other positive sites were compared in a pairwise manner with the five centroids and clustered into groups with the highest similarity values. Third, the centroid of each cluster was updated with the highest average similarity (HAS). The second and third steps were iteratively repeated until the clusters did not change any longer. After the five clusters for the positive sites had been determined, we put each negative site into the cluster with the HAS.

(2) Motif length selection (MLS). In this step, the optimized combination of NSP(m, n) was determined for better performance. The combinations of NSP(m, n) (m = 1,..., 30; n = 1,..., 30) were extensively tested, while the optimal NSP(m, n) for each cluster with the highest leave-one-out performance was respectively selected. We fixed the Sp at 80% to compare Sn values.

(3) Weight training (WT). We updated the substitution score between two NSP(m, n) peptides A and B as:

$$S'(A, B) = \sum_{-m \le i \le n} w_i \mathrm{Score}(A[i], B[i])$$

The w_i is the weight of position i. Again, if $S'(A, B) < 0$, we simply redefined it as $S'(A, B) = 0$. Initially, the w was defined as 1 for each position. We randomly picked out the weight of any position for +1 or −1, and adopted the manipulation if the Sn value of the re-computed leave-one-out result with the Sp fixed at 80% was increased. The process was repeated until convergence was reached.

(4) Matrix mutation (MaM). As previously described,[24,25] BLOSUM62 was chosen as the initial matrix, and the leave-one-out performance was calculated. Subsequently, we fixed the Sp as 80% to improve the Sn by randomly picking out an element of the matrix for +1 or −1. The procedure was terminated when the Sn value was not increased any further.

For comparison, the GPS 2.0 algorithm and PSSM algorithm were also implemented. The GPS 2.0 algorithm was carried out as previously described.[25] For the PSSM algorithm,[31] the probabilities of twenty amino acids in the positive data (+) and negative data (−) were calculated as P_+ and P_-. Then the score of a given NSP(m, n) was calculated as:

$$\mathrm{Score}[\mathrm{NSP}(m, n)] = \sum_{1 \le i \le m+n} \log_2(P_+[i]/P_-[i])$$

Statistical analysis

In order to analyze the functional abundance and diversity of PTN, we downloaded the gene ontology (GO) (06/29/2010)[32] association files from the GOA database at the EBI (http://www.ebi.ac.uk/goa). There are 18 262 human proteins annotated with at least one GO term, with 408 annotated nitration substrates. Here we defined:

N = number of proteins in human proteome annotated by at least one GO term

n = number of proteins in human proteome annotated by the GO term t

M = number of proteins in human nitrated substrates annotated by at least one GO term

m = number of proteins in human nitrated substrates annotated by the GO term t

Then the enrichment ratio of the GO term t was calculated, while the hypergeometric distribution equation[33] was used to calculate the p-value as below:

$$\mathrm{Enrichment\_ratio} = \frac{\frac{m}{M}}{\frac{n}{N}}$$

$$p\text{-value} = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\mathrm{Enrichment\_ratio} \ge 1), \quad \text{or}$$

$$p\text{-value} = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\mathrm{Enrichment\_ratio} < 1)$$

In this work, we only consider the over-represented GO groups with Enrichment_ratio ≥ 1. From our previous study,[24] we also collected 396 human S-nitrosylated substrates with at least one GO annotation. The statistical procedure was also performed for the GO enrichment analysis of S-nitrosylation.

For comparison of PTN with S-nitrosylation, we performed the Yates' Chi-square (χ^2) test with the 2 × 2 contingency table method.[34]

Performance evaluation

As previously described,[24,25] we used the four measurements of sensitivity (Sn), specificity (Sp), accuracy (Ac), and Mathew's Correlation Coefficient (MCC) to evaluate the prediction performance of GPS-YNO2. The Ac represents the correct ratio between both the positive (+) and negative (−) data sets, whereas Sn and Sp illustrate the correct prediction ratios of the positive (+) and negative data (−) sets, respectively. When the number of positive data and negative data differ too much from each other, the MCC should also be calculated. The value of MCC ranges from −1 to 1, and a larger MCC stands for better performance.

Among the data with positive hits obtained by GPS-YNO2, the real positives are defined as "true positives" (TP), while the others are defined as "false positives" (FP). Among the data with the negative predictions obtained by GPS-YNO2, the real positives are defined as "false negatives" (FN), while the others are defined as "true negatives" (TN). The performance measurements of Ac, Sn, Sp, and MCC are defined as below:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \ \mathrm{Ac} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \text{ and}$$

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

In this work, the leave-one-out validation and 4-, 6-, 8-, 10-fold cross-validations were performed. The Receiver

Operating Characteristic (ROC) curves and AROCs (area under ROCs) were also drawn and analyzed.

Implementation of the online service and local packages

The online service and local packages of GPS-YNO2 1.0 were implemented in JAVA. For the online service, we tested the GPS-YNO2 1.0 on a variety of internet browsers, including Internet Explorer 6.0, Netscape Browser 8.1.3 and Firefox 2 under the Windows XP Operating System (OS), Mozilla Firefox 1.5 of Fedora Core 6 OS (Linux), and Safari 3.0 of Apple Mac OS X 10.4 (Tiger) and 10.5 (Leopard). For the Windows and Linux systems, the latest version of Java Runtime Environment (JRE) package (JAVA 1.4.2 or later versions) of Sun Microsystems should be pre-installed. However, for Mac OS, GPS-YNO2 1.0 can be directly used without any additional packages. For convenience, we also developed local packages of GPS-YNO2 1.0, which worked with the three major Operating Systems, Windows, Linux and Mac.

Results

Development of GPS-YNO2 for prediction of PTN sites

In this work, a training data set of 1066 experimentally verified PTN sites in 554 unique proteins was collected from the scientific literature (Table S1, ESI†). Previously, we developed the GPS (Group-based Prediction System) algorithm for the prediction of kinase-specific phosphorylation sites.[25] Recently, the GPS algorithm was substantially improved to version 3.0 and achieved a considerable success in the prediction of S-nitrosylation sites.[24] Here, we applied the GPS 3.0 algorithm to predict PTN sites. To improve prediction performance, a sequential four-step procedure was adopted, with *k*-means clustering, MLS, WT and MaM. By exhaustively testing, it was found that this training order cannot be changed. Through the *k*-means clustering method, the training dataset was classified into five groups, clusters A, B, C, D, and E, with HAS values of 0.2542, 0.2809, 0.2698, 0.2784 and 0.2521, respectively. Based on the highest leave-one-out performance, the NSP(*m*, *n*) for clusters A, B, C, D and E were determined to be NSP(25, 6), NSP(27, 11), NSP(19, 28), NSP(10, 20) and NSP(8, 4), respectively. Finally, the weight of each position and scoring matrix for each cluster were optimized.

After the training to improve performance, we developed the GPS-YNO2 software to predict the PTN sites. Although the NSP(*m*, *n*) for each cluster is different, NSP(7, 7) is shown for convenience. The prediction results for the mouse 14-3-3 protein epsilon (UniProt ID: P62259) is shown as an example (Fig. 2). Previously, the mouse 14-3-3 protein epsilon was experimentally identified to be nitrated at Y49 and Y214.[15] During data collection, we found that Y9, Y131 and Y214 in the human 14-3-3 protein epsilon are also nitrated.[17] Since the human and mouse sequences of the 14-3-3 protein epsilon are identical, the Y9, Y49, Y131 and Y214 sites were preserved only in the mouse protein after redundant clearing (Table S1, ESI†). Furthermore, the mouse 14-3-3 protein beta/alpha was found to be nitrated at Y84,[35] with a highly conserved site of Y85 in its paralog of the 14-3-3 protein epsilon (Table S1, ESI†). Using GPS-YNO2 1.0, it was predicted that the 14-3-3 protein epsilon can be nitrated at Y9, Y20, Y49, Y85, Y131, Y152 and Y214 (Fig. 2). The positive hits of Y9, Y49, Y85, Y131 and Y214 are consistent with previous experimental studies,[15,17,35] while the prediction of Y20 and Y152 provides useful information for further experimental verification.

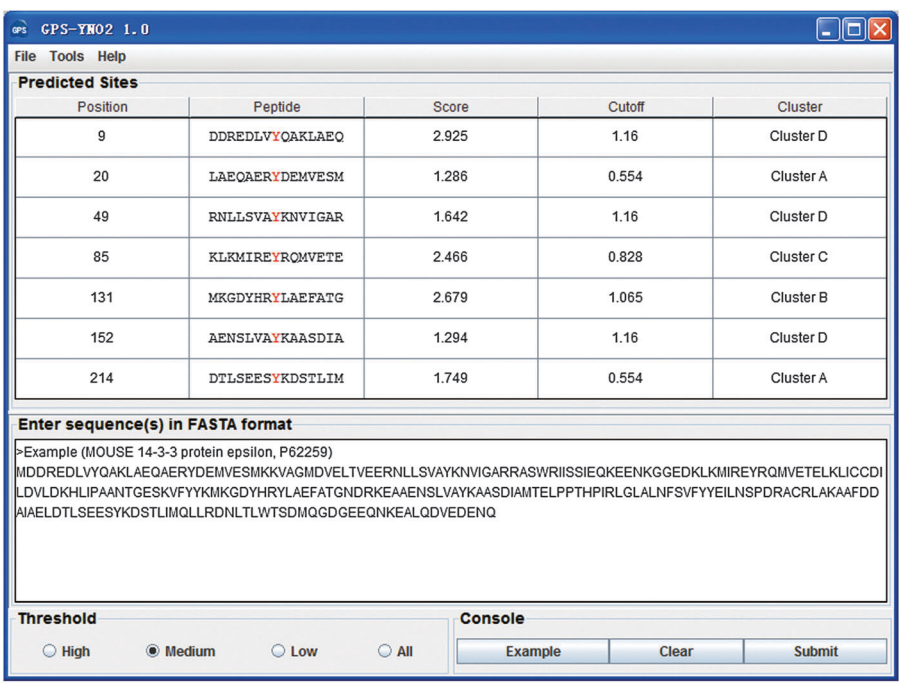

Fig. 2 Screen snapshot of the GPS-YNO2 1.0 software. The medium threshold was chosen as the default threshold. As an example, the prediction results of mouse 14-3-3 protein epsilon (P62259) were shown.

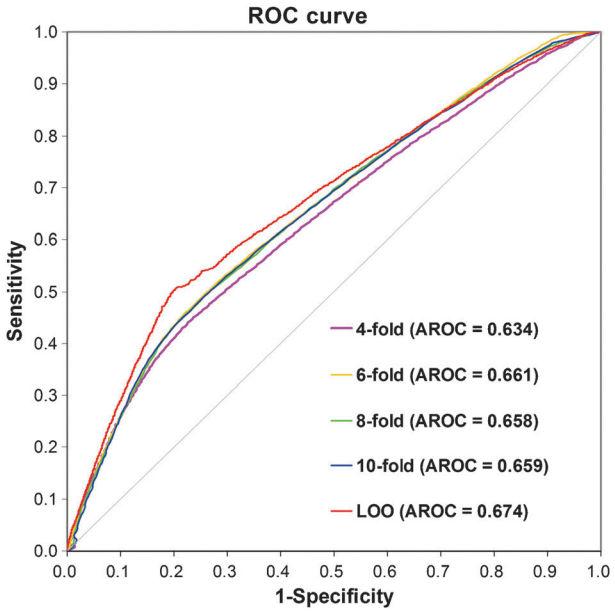

Fig. 3 The prediction performance of GPS-YNO2 1.0. The leave-one-out validation and 4-, 6-, 8-, 10-fold cross-validations were calculated. The Receiver Operating Characteristic (ROC) curves and AROCs (area under ROCs) were also drawn and analyzed.

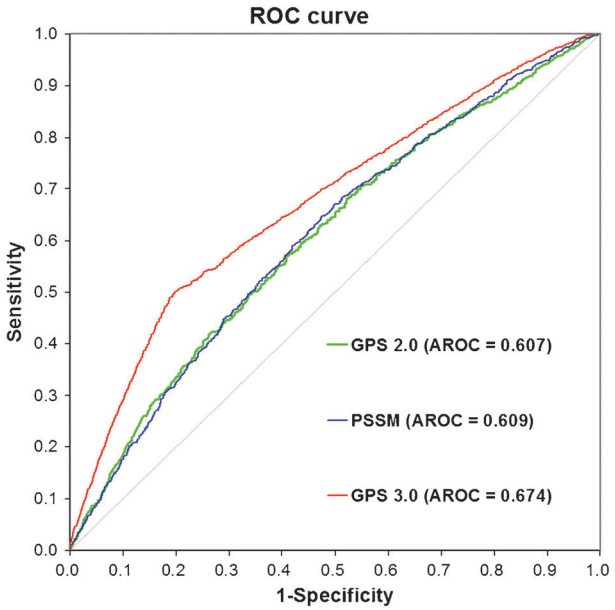

Fig. 4 Comparison of GPS 3.0, GPS 2.0 and PSSM with the leave-one-out performance.

Performance evaluation and comparison

Taking probable over-fitting into consideration, we employed the leave-one-out and 4-, 6-, 8-, 10-fold cross-validations to evaluate the prediction robustness and performance of GPS-YNO2. The ROC curves are presented in Fig. 3, while the AROC values were calculated as 0.674 (leave-one-out), 0.634 (4-fold), 0.661 (6-fold), 0.648 (8-fold) and 0.659 (10-fold), respectively (Fig. 3). Since the 4-, 6-, 8-, 10-fold cross-validations were close to the leave-one-out validation, it was demonstrated that GPS-YNO2 1.0 is a robust predictor of PTN sites and thus of promising performance.

For comparison, we calculated the performance of several other approaches, including GPS 2.0 and position-specific scoring matrix (PSSM).[25,31] To avoid any bias, the dataset used in GPS 3.0 was also employed in these two methods. We calculated the leave-one-out validations for the GPS 3.0, GPS 2.0 and PSSM algorithms and drew the ROC curve (Fig. 4). The AROC values were calculated as 0.607 (GPS 2.0), 0.609 (PSSM) and 0.674 (GPS 3.0). In addition, we compared the Sn values with the fixed Sp values of GPS 3.0 in a manner identical with other methods (Table 1). Through these comparisons, GPS 3.0 was demonstrated to be better than the other methods. Previously, Elfering *et al.* suggested that PTN recognizes the consensus sequences of [LMVI]-*X*-[DE]-[LMVI]-*X*(2,3)-[FVLI]-*X*(3,5)-*Y* (where *X* is any amino acid and *Y* is the target tyrosine) or *H*-*X*-[DE]-*H*-*X*(2,3)-*H*(2)-*X*(2,4)-*Y* (where *H* is a hydrophobic residue).[23] With the same dataset, we calculated the performance of the two motifs as 86.98% of Ac, 1.59% of Sn, 98.83% of Sp and 86.41% of Ac, 2.44% of Sn, 98.06% of Sp, respectively (Table 1). With the same Sp values of 98.83% and 98.06%, the Sn values of GPS 3.0 were 4.78% and 6.85%, respectively (Table 1). In this regard, GPS 3.0 was shown to be superior to the simple linear motif approach.

Table 1 Comparison of the GPS 3.0 algorithm with other approaches. For the construction of GPS-YNO2 software, the three thresholds of high, medium and low were chosen. For comparison, we fixed the Sp values of GPS 3.0 to be identical or similar to other methods and compared the Sn values

Method	Threshold	Ac (%)	Sn (%)	Sp (%)	*MCC*
GPS 3.0	High	82.57	28.89	90.02	0.1884
	Medium	79.60	40.53	85.02	0.2171
	Low	76.51	50.09	80.18	0.2335
		87.37	4.78	98.83	0.0884
		86.95	6.85	98.06	0.0979
GPS 2.0		81.41	18.39	90.15	0.0896
		78.07	27.96	85.02	0.1142
		74.38	33.49	80.05	0.1076
PSSM		81.22	17.46	90.04	0.0806
		77.69	24.86	85.02	0.0877
		74.27	32.46	80.08	0.0999
Motif 1[a]		86.98	1.59	98.83	0.0126
Motif 2[b]		86.41	2.44	98.06	0.0117

[a] Motif 1, [LMVI]-*X*-[DE]-[LMVI]-*X*(2,3)-[FVLI]-*X*(3,5)-*Y* (where *X* is any amino acid and *Y* is the target tyrosine).[23] [b] Motif 2, *H*-*X*-[DE]-*H*-*X*(2,3)-*H*(2)-*X*(2,4)-*Y* (where *H* represents a hydrophobic residue).[23]

Large-scale prediction of PTN sites in proteins

Previously, a large number of nitrated proteins have been reported in small- or large-scale studies, even though the actual PTN sites remain to be elucidated. As a demonstration of GPS-YNO2 1.0, 370 potentially nitrated substrates from the scientific literature were collected (Table S2, ESI†), followed by retrieving the primary sequences from the UniProt database.[28] We successfully predicted 325 (~88%) of the proteins having at least one potential PTN site using GPS-YNO2 under the default threshold (medium) (Table S2, ESI†). These predictions should be useful for further experimental identification. Several examples were randomly picked out, and the results are shown in Fig. 5.

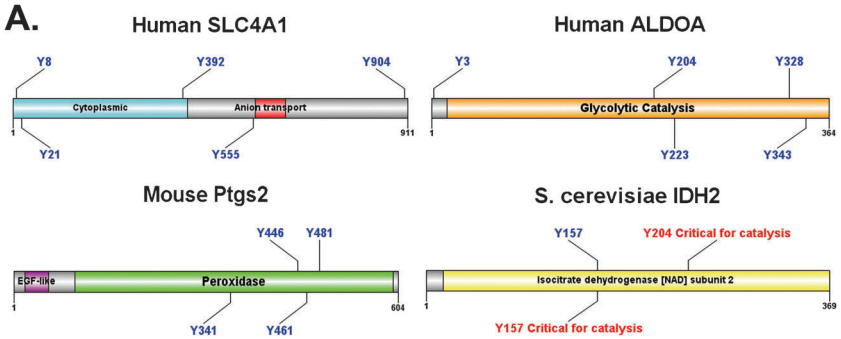

Fig. 5 Applications of GPS-YNO2 1.0. We predicted potential PTN sites in the experimentally identified nitrated substrates with the default threshold. (A) The human band 3 anion transport protein/SLC4A1 (P02730); (B) the human fructose-bisphosphate aldolase A/ALDOA (P04075); (C) the mouse prostaglandin G/H synthase 2/Ptgs2 (Q05769); (D) the yeast isocitrate dehydrogenase/IDH2 (P28241).

It was proposed that PTN of the human band 3 anion transport protein/SLC4A1 (P02730) by peroxynitrite inhibits its phosphorylation by tyrosine kinases in human erythrocytes.[36] However, the corresponding nitrated tyrosines were not experimentally identified. With GPS-YNO2 1.0, we predicted that the human SLC4A1 can be nitrated at Y8, Y21, Y392, Y555 and Y904. Interestingly, from the UniProt annotation information,[28] we found that the three residues of Y8, Y21 and Y904 are known phosphotyrosines (Fig. 5A). Recent work by Sekar *et al.* suggested that the nitration of fructose-bisphosphate aldolase A/ALDOA (P04075) reduces its maximum velocity to regulate human mast cell (MC) phenotype and function.[37] Here, we predicted that Y3, Y204, Y223, Y328 and Y343 in ALDOA might be nitrated, with Y3, Y204 and Y223 being known phosphotyrosines (Fig. 5B). Previously, Schildknecht *et al.* suggested that the mouse prostaglandin G/H synthase 2/Ptgs2 (Q05769) in endotoxin-stimulated RAW 264.7 macrophages was autocatalytically nitrated, although the actual sites were not determined.[38] We predicted that Y341, Y446, Y461 and Y481 might be nitrated (Fig. 5C). Again, the Y446 is a known phosphotyrosine. Taken together, PTN plays an important role in signal transduction by rivaling tyrosine phosphorylation at the same sites. In addition, a proteomics investigation in *S. cerevisiae* identified isocitric dehydrogenase/IDH2 (P28241) as a PTN substrate.[39] Here we predicted that IDH2 might be nitrated at Y157, which would be expected to be critical for its catalytic activity from the UniProt annotations (Fig. 5D).

Discussion

As a ubiquitous PTM critical for a wide array of biological processes in physiology and pathology, PTN has attracted considerable interest and investigative efforts.[1–6] It is widely accepted that cellular nitric oxide has important implications in PTN.[1–6] Since the identification of nitrated substrates with the actual sites is fundamental for dissecting the molecular mechanisms and regulatory functions of PTN,[1–3] a large number of low-throughput and large-scale studies have been performed in this area.[14,16–21] However, compared with time-consuming and expensive experimental methods, the computational prediction of PTN sites is a conveniently rapid method of obtaining useful information for subsequent experimental verification.

Previously, Yang prepared a training data set containing 56 positive and 73 negative 10-mer PTN peptides, and used four machine-learning algorithms for prediction.[40] Although the performance was claimed to be promising, no applicable tool was made available.[40] Recently, He *et al.* obtained 56 positive and 725 negative PTN peptides from NCBI.[41] With a nearest neighbor algorithm, a satisfying performance was reached, although no applicable predictors are available.[41] In this work, we present the novel software GPS-YNO2 for the prediction of PTN sites, using a much larger training data set with 1066 positive and 7684 negative sites. We greatly refined the previously developed algorithm of GPS 2.0, the new program containing a scoring strategy and an approach of matrix mutation (MaM) to improve the performance.[25] In GPS 3.0, the original scoring strategy was adopted as the initial step. Certain performance-improvement procedures, including *k*-means clustering, MLS, WT and MaM, helped the GPS 3.0 algorithm to obtain a promising result. Through annotating the exact PTN sites for those substrates for which precise sites had not been identified in previous large-scale or small-scale studies, the current GPS 3.0 algorithm has already exhibited superiority, although further improvement is still expected.

As a result of the continuing advances made in previous studies, PTN was found to target broad substrates in different biological processes. The collection of PTN substrates from the literature provided an opportunity to analyze the functional abundance and diversity of PTN. With a hyper-geometric distribution,[33] we statistically analyzed the enriched biological processes, molecular functions and cellular components with gene ontology (GO) annotations for the human PTN substrates (Table S3, ESI†). The five most over-represented biological processes of translational elongation (GO:0006414), RNA splicing (GO:0008380), mRNA processing (GO:0006397), translation (GO:0006412) and protein folding (GO:0006457) suggested that PTN regulates transcription, translation and protein stability through post-translational modification, while the statistical results in terms of the molecular functions, *e.g.*, protein binding (GO:0005515), RNA binding (GO:0003723), nucleotide binding (GO:0000166), structural constituent of ribosome (GO:0003735), and unfolded protein binding (GO:0051082), are consistent with this notion (Table S3, ESI†).

Table 2 Statistical comparison of the GO terms of the substrates between PTN and S-nitrosylation

Description of GO term	Nitration		Nitrosylation				
	Num.[a]	Per.[b] (%)	Num.	Per. (%)	*E*-ratio[c]	χ^{2d}	*p*-Value
The most different biological processes							
RNA splicing (GO:0008380)	38	9.84	3	0.76	12.99	30.68	3.04E−08
mRNA processing (GO:0006397)	29	7.51	1	0.25	29.75	26.00	3.42E−07
Translation (GO:0006412)	26	6.74	9	2.27	2.96	8.09	4.44E−03
Translational elongation (GO:0006414)	27	6.99	11	2.78	2.52	6.63	1.00E−02
Glycolysis (GO:0006096)	**10**	**2.59**	**24**	**6.06**	**0.43**	**4.86**	**2.76E−02**
Multicellular organismal development (GO:0007275)	**7**	**1.81**	**18**	**4.55**	**0.40**	**3.87**	**4.91E−02**
The most different molecular functions							
RNA binding (GO:0003723)	63	16.32	17	4.29	3.80	29.50	5.60E−08
Structural constituent of ribosome (GO:0003735)	25	6.48	7	1.77	3.66	9.88	1.67E−03
Receptor binding (GO:0005102)	15	3.89	3	0.76	5.13	7.17	7.40E−03
Single-stranded DNA binding (GO:0003697)	12	3.11	2	0.51	6.16	6.13	1.33E−02
Heme binding (GO:0020037)	12	3.11	3	0.76	4.10	4.56	3.27E−02
Transferase activity (GO:0016740)	**23**	**5.96**	**42**	**10.61**	**0.56**	**4.95**	**2.61E−02**
NAD or NADH binding (GO:0051287)	**2**	**0.52**	**11**	**2.78**	**0.19**	**4.80**	**2.84E−02**
Structural molecule activity (GO:0005198)	**7**	**1.81**	**18**	**4.55**	**0.40**	**3.87**	**4.91E−02**
The most different cellular localizations							
Nucleolus (GO:0005730)	53	13.73	18	4.55	3.02	18.88	1.39E−05
Spliceosomal complex (GO:0005681)	20	5.18	1	0.25	20.52	16.33	5.31E−05
Ribosome (GO:0005840)	27	6.99	8	2.02	3.46	10.18	1.42E−03
Nucleus (GO:0005634)	150	38.86	116	29.29	1.33	7.55	6.00E−03
Extracellular space (GO:0005615)	40	10.36	22	5.56	1.87	5.55	1.85E−02
Nucleoplasm (GO:0005654)	36	9.33	20	5.05	1.85	4.75	2.93E−02
Cytosolic large ribosomal subunit (GO:0022625)	10	2.59	2	0.51	5.13	4.33	3.74E−02
Cytoskeleton (GO:0005856)	**21**	**5.44**	**43**	**10.86**	**0.50**	**6.93**	**8.46E−03**
Synapse (GO:0045202)	**3**	**0.78**	**14**	**3.54**	**0.22**	**5.76**	**1.64E−02**
Mitochondrial outer membrane (GO:0005741)	**2**	**0.52**	**12**	**3.03**	**0.17**	**5.66**	**1.74E−02**
Mitochondrion (GO : 0005886)	**48**	**12.44**	**73**	**18.43**	**0.67**	**4.93**	**2.64E−02**

[a] The number of proteins annotated. [b] The proportion of proteins annotated. [c] *E*-ratio, enrichment ratio, the PTN proportion divided by the S-nitrosylation proportion. [d] The result of the Chi-square (χ^2) test. The entries in bold indicate the Enrichment_ratio $\leq$ 1.

Since PTN and S-nitrosylation are both associated with NO-dependent oxidative stress, we also surveyed the biological roles of S-nitrosylation (Table S4, ESI†). For S-nitrosylation, the over-represented biological processes such as glycolysis (GO:0006096), cellular carbohydrate metabolic process (GO:0044262), proteolysis involved in cellular protein catabolic process (GO:0051603), and the tricarboxylic acid cycle (GO:0006099) suggested that S-nitrosylation plays an essential role in metabolism and catabolism (Table S4, ESI†). Again, the statistical results in terms of cellular components, *e.g.*, mitochondrion (GO:0005739), melanosome (GO:0042470), and mitochondrial matrix (GO:0005759), are consistent with this notion (Table S4, ESI†).

From the individual statistical results for PTN and S-nitrosylation, it was obvious that these two PTMs shared many GO terms, such as anti-apoptosis (GO:0006916), protein folding (GO:0006457) and so on (Tables S3 and S4, ESI†). These results were consistent with previous reports.[13,20,42,43] However, by comparison of PTN and S-nitrosylation with the Yates' Chi-square (χ^2) test,[34] we also observed a number of interesting differences (Table 2). For instance, compared with the enhancement in transcription and translation of the PTN substrates, biological processes such as ion transport, glycolysis and multicellular organismal development were over-represented in the S-nitrosylated proteins (Table 2). The molecular function results were consistent with those of the biological processes. Based on systematic analysis it is proposed that, compared with reversible S-nitrosylation, PTN preferentially attacks fundamental biological processes and functions.

Taken the various lines of evidence together, we propose that GPS-YNO2 1.0 can serve as a useful tool for identifying potential PTN sites. Also, this analysis provides a good start for further investigating molecular mechanisms of PTN. We believe computational predictions followed by experimental verification will help advance the understanding of the mechanisms and dynamics of PTN.

Acknowledgements

We are grateful for two anonymous reviewers, whose suggestions have greatly improved the presentation of this manuscript. The authors thank Dr Chang Chen (IBP) for her helpful comments. This work was supported by grants from the National Basic Research Program (973 project) (2010CB945400), National Natural Science Foundation of China (90919001, 30700138, 30900835, 30830036, 30900835, 31071154), Chinese Academy of Sciences (INFO-115-C01-SDB4-36) and HUST Innovative Program (2010ZD018). Pacific Edit reviewed the manuscript prior to submission.

The authors have declared no conflict of interest.

References

1 R. Radi, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 4003–4008.
2 J. S. Beckman and W. H. Koppenol, *Am. J. Physiol.*, 1996, **271**, C1424–C1437.
3 F. J. Schopfer, P. R. Baker and B. A. Freeman, *Trends Biochem. Sci.*, 2003, **28**, 646–654.

4 D. D. Thomas, M. G. Espey, M. P. Vitek, K. M. Miranda and D. A. Wink, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12691–12696.
5 K. Bian, Z. Gao, N. Weisbrodt and F. Murad, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 5712–5717.
6 S. Goldstein, G. Czapski, J. Lind and G. Merenyi, *J. Biol. Chem.*, 2000, **275**, 3031–3036.
7 R. Zaragoza, L. Torres, C. Garcia, P. Eroles, F. Corrales, A. Bosch, A. Lluch, E. R. Garcia-Trevijano and J. R. Vina, *Biochem. J.*, 2009, **419**, 279–288.
8 L. L. Hardy, D. A. Wick and J. R. Webb, *J. Immunol.*, 2008, **180**, 5956–5962.
9 K. Dixit, M. A. Khan, Y. D. Sharma, M. Uddin and K. Alam, *Int. J. Biol. Macromol.*, 2009.
10 A. J. Gow, D. Duran, S. Malcolm and H. Ischiropoulos, *FEBS Lett.*, 1996, **385**, 63–66.
11 N. Abello, H. A. Kerstjens, D. S. Postma and R. Bischoff, *J. Proteome Res.*, 2009.
12 J. A. Kers, M. J. Wach, S. B. Krasnoff, J. Widom, K. D. Cameron, R. A. Bukhalid, D. M. Gibson, B. R. Crane and R. Loria, *Nature*, 2004, **429**, 79–82.
13 K. S. Aulak, M. Miyagi, L. Yan, K. A. West, D. Massillon, J. W. Crabb and D. J. Stuehr, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 12056–12061.
14 R. Tyther, A. Ahmeda, E. Johns and D. Sheehan, *Proteomics*, 2007, **7**, 4555–4564.
15 Q. Zhang, W. J. Qian, T. V. Knyushko, T. R. Clauss, S. O. Purvine, R. J. Moore, C. A. Sacksteder, M. H. Chin, D. J. Smith, D. G. Camp, 2nd, D. J. Bigelow and R. D. Smith, *J. Proteome Res.*, 2007, **6**, 2257–2268.
16 F. Casoni, M. Basso, T. Massignan, E. Gianazza, C. Cheroni, M. Salmona, C. Bendotti and V. Bonetto, *J. Biol. Chem.*, 2005, **280**, 16295–16304.
17 B. Ghesquiere, N. Colaert, K. Helsens, L. Dejager, C. Vanhaute, K. Verleysen, K. Kas, E. Timmerman, M. Goethals, C. Libert, J. Vandekerckhove and K. Gevaert, *Mol. Cell. Proteomics*, 2009, **8**, 2642–2652.
18 J. Kanski, S. J. Hong and C. Schoneich, *J. Biol. Chem.*, 2005, **280**, 24261–24266.
19 M. Miyagi, H. Sakaguchi, R. M. Darrow, L. Yan, K. A. West, K. S. Aulak, D. J. Stuehr, J. G. Hollyfield, D. T. Organisciak and J. W. Crabb, *Mol. Cell. Proteomics*, 2002, **1**, 293–303.
20 I. V. Turko, L. Li, K. S. Aulak, D. J. Stuehr, J. Y. Chang and F. Murad, *J. Biol. Chem.*, 2003, **278**, 33972–33977.
21 X. Zhan, Y. Du, J. S. Crabb, X. Gu, T. S. Kern and J. W. Crabb, *Mol. Cell. Proteomics*, 2008, **7**, 864–874.
22 J. M. Souza, E. Daikhin, M. Yudkoff, C. S. Raman and H. Ischiropoulos, *Arch. Biochem. Biophys.*, 1999, **371**, 169–178.
23 S. L. Elfering, V. L. Haynes, N. J. Traaseth, A. Ettl and C. Giulivi, *Am. J. Physiol.: Heart Circ. Physiol.*, 2004, **286**, H22–H29.
24 Y. Xue, Z. Liu, X. Gao, C. Jin, L. Wen, X. Yao and J. Ren, *PLoS One*, 2010, **5**, e11290.
25 Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen and X. Yao, *Mol. Cell. Proteomics*, 2008, **7**, 1598–1608.
26 H. Li, X. Xing, G. Ding, Q. Li, C. Wang, L. Xie, R. Zeng and Y. Li, *Mol. Cell. Proteomics*, 2009, **8**, 1839–1849.
27 T. Y. Lee, J. B. Hsu, W. C. Chang, T. Y. Wang, P. C. Hsu and H. D. Huang, *BMC Res. Notes*, 2009, **2**, 111.
28 The UniProt Consortium, *Nucleic Acids Res.*, 2010, **38**, D142–D148.
29 W. Li and A. Godzik, *Bioinformatics*, 2006, **22**, 1658–1659.
30 M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis and T. L. Madden, *Nucleic Acids Res.*, 2008, **36**, W5–W9.
31 D. T. Jones, *J. Mol. Biol.*, 1999, **292**, 195–202.
32 D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan and R. Apweiler, *Nucleic Acids Res.*, 2009, **37**, D396–D403.
33 F. Zhou, Y. Xue, H. Lu, G. Chen and X. Yao, *FEBS Lett.*, 2005, **579**, 3369–3375.
34 D. C. David, N. Ollikainen, J. C. Trinidad, M. P. Cary, A. L. Burlingame and C. Kenyon, *PLoS Biol.*, 2010, **8**, e1000450.
35 C. A. Sacksteder, W. J. Qian, T. V. Knyushko, H. Wang, M. H. Chin, G. Lacan, W. P. Melega, D. G. Camp, 2nd, R. D. Smith, D. J. Smith, T. C. Squier and D. J. Bigelow, *Biochemistry*, 2006, **45**, 8009–8022.
36 C. Mallozzi, A. M. Di Stasi and M. Minetti, *FASEB J.*, 1997, **11**, 1281–1290.
37 Y. Sekar, T. C. Moon, C. M. Slupsky and A. D. Befus, *J. Immunol.*, 2010, **185**, 578–587.
38 S. Schildknecht, K. Heinz, A. Daiber, J. Hamacher, C. Kavakli, V. Ullrich and M. Bachschmid, *Biochem. Biophys. Res. Commun.*, 2006, **340**, 318–325.
39 A. Bhattacharjee, U. Majumdar, D. Maity, T. S. Sarkar, A. M. Goswami, R. Sahoo and S. Ghosh, *Biochem. Biophys. Res. Commun.*, 2009, **388**, 612–617.
40 Z. R. Yang, *Machine Learning Approaches to Bioinformatics*, World Scientific Publishing Co. Pte. Ltd., 2010.
41 Z. He, T. Huang, X. Shi, L. Hu, L. Chen, F. Liu, K. Wang, T. Wen, X. Kong and Y. Cai, presented in part at the Fourth International Conference on Computational Systems Biology (ISB2010), Suzhou, China, 2010.
42 S. Duan and C. Chen, *Cell Mol. Immunol.*, 2007, **4**, 353–358.
43 T. Nakamura and S. A. Lipton, *Apoptosis*, 2009, **14**, 455–468.

