The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Computational Analysis of Protein Tyrosine Nitration

ZhiSong He[4,5,^] Tao Huang[1,2,^] XiaoHe Shi[6] LeLe Hu[3] Lei Chen[9] Fang Liu[3,8] Kai Wang[3] TieQiao Wen[3,8,*] XiangYin Kong[6,7,*] Yudong Cai[3,*]

[1]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2]Shanghai Center for Bioinformation Technology, Shanghai 200235, China
[3]Institute of System Biology, Shanghai University, Shanghai 200444, China
[4]CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China
[5]Centre for Computational Systems Biology, Fudan University, Shanghai 200433, China
[6]Institute of Health Sciences, Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS) and Shanghai Jiao Tong University School of Medicine (SJTUSM)
[7]State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiaotong University, Shanghai 200025, China.
[8]School of Life Sciences, Shanghai University, Shanghai 200444, China
[9]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, 200062, China
^ These authors are considered to be the joint first authors
* Corresponding authors

Abstract As an important covalent post-translational modification, Tyrosine nitration is closely related to the causes of various diseases. However, its mechanism is still largely unknown. Based on experimentally verified tyrosine nitration sites, we introduced a novel computational approach to analyze tyrosine nitration. Nearest Neighbor algorithm armed by maximum relevancy minimum redundancy feature selection approach was used to construct the tyrosine nitration predictor. The problem of the imbalance of dataset sizes was dealt with by dividing large dataset into several smaller ones. Analysis of the selected features shows some interesting phenomena of the tyrosine nitration, which could be helpful for further computational and experimental investigations.

Keywords Tyrosine nitration; feature selection; Nearest Neighbor algorithm

1 Introduction

Tyrosine nitration occurs widely in association with important pathophysiological consequences such as diabetes [1] and neurodegenerative diseases [2]. The nitration of tyrosine residues constitutes the substitution of

hydrogen by a nitro group in the 3-position of the phenolic ring and represents a modification produced by nitric oxide-derived oxidants such as nitrogen dioxide radical and peroxynitrite [3].

A few of progress has been made in detecting nitrated proteins using specific antibodies against protein 3-nitrotyrosine [4]. It is a nearly impossible process to solve every nitration site structure of protein complex through molecular biological methods. Compared to the experimental researches, the computational methods have the advantage of high efficiency and low cost to deal with large scale of data. As no predictor is available for predicting nitrated tyrosine, we developed a novel sequence-based method for predicting protein tyrosine nitration in hopes that it may become a useful computational tool in the relevant areas. Nearest Neighbor algorithm armed by maximum relevancy minimum redundancy feature selection was used to construct the predictor. Our results provide clues of nitration mechanisms, and useful insights to elevate protein tyrosine nitration.

2 Materials and methods

2.1 Data set

Searching NCBI using “nitration” led to 123 protein sequences. To diminish bias caused by high similarities between sequences, we removed sequences which have similarities higher than 70% with other sequences using CD-HIT[5]. Finally, 48 sequences were retained. We used a sliding window to scan the protein sequences to obtain peptides of symmetrical flanking residues of each tyrosine in the proteins. The 30 amino acid residues with 15 ones on each side of each tyrosine were seen as a sample. The peptides in a sliding window formed from an experimentally confirmed nitrated tyrosine was labeled as a positive sample, while the peptides formed from other tyrosine were labeled as negative ones. With the sampling process, we finally got 781 samples, with 56 positive and 725 negative ones, respectively.

2.2 Representing proteins with biochemical and physicochemical features

To develop a method for predicting a protein-related system, one of the most important things is to formulate the sample of a protein with the core features that have intrinsic correlation with the predicted target. Here, we used a set of features based AAIndex (http://www.genome.ad.jp/aaindex/) [6] to represent protein samples in terms of their biochemical and physicochemical properties. 506 indices from AAIndex were used to represent one amino acid. Since each sample contains 30 amino acids in our study, totally $506 \times 30=15180$ features were generated.

2.3 Dealing with the imbalanced data

In this study, the size of negative sample set is much larger than the size of positive one (nearly 13 times of). This imbalance can hamper a training process and bias the classification in favor of the class with more samples. To deal with this problem, we generated 13 different dataset from the original one. All these dataset

contained the 56 positive training samples, while only the 725 negative samples were split equally into 13 portions. Thus 13 training datasets were built with a proportion of around 1:1 between the positive and negative samples in each training dataset. Based on these 13 datasets, 13 independent classifiers were constructed with virtually no data imbalance.

2.4 Classifier construction using Nearest Neighbor Algorithm

In this study, Nearest Neighbor algorithm (NNA) [7-8]was used as the classification model for the nitrated site prediction. It makes its decision based on similarities between the feature vector of the testing sample and all the feature vectors of the training dataset. The class of the sample in the training set, i.e. the nearest neighbor, which has the highest similarity with the training sample, would be designated to be the class of the testing sample. In this study, the distance between two vectors p_i and p_j is defined as:

$$D(p_i, p_j) = 1 - \frac{p_i \cdot p_j}{\| p_i \| \cdot \| p_j \|} \tag{1}$$

where $p_i \cdot p_j$ is the inner product of p_i and p_j, and $\| p \|$ represents the module of vector p.

2.5 Jackknife cross-validation

In Jackknife cross-validation, each sample in the data set would be tested by the classifier based on all the other samples, so every sample would be tested exactly once. Overall accurate prediction rate is used for quantifying the evaluation, which is defined as the rate between the number of corrected predicted samples and the number of the samples in the whole dataset.

2.6 Maximum Relevancy Minimum Redundancy

Maximum Relevancy Minimum Redundancy (mRMR) [9] is used for feature evaluation of nitration prediction. It ranks features in the feature set according to each feature's relevancy to the target variable and redundancy to other features. To quantify relevancy and redundancy, mutual information (MI), which estimates the relationship between different vectors, is calculated and denoted as I. Based on MI, an mRMR function is defined as:

$$\max_{f_j \in \Omega_t} \left[I(f_j, c) - \frac{1}{m} \sum_{f_i \in \Omega_s} I(f_{j,} f_i) \right] (j = 1, 2, ..., n) \tag{2}$$

where Ω_s and Ω_t are the already-selected feature set and to-be-selected feature set, while m, n are the sizes of these two feature sets, respectively. The earlier a feature is selected by this function, the higher rank it would get, and the more important it would be regarded.

For there are 13 independent classifiers built, each of them needs to be processed by mRMR to gain their own feature set. So mRMR was run 13 times.

2.7 Incremental Feature Selection

An incremental feature selection is conducted for each of the independent predictor with the ranked features. Features in a set are added one by one from higher to lower rank. If one feature is added, a new feature set is obtained, then we get N feature sets where N is the number of features, and the i-th feature set is:

$$S_i = \{f_1, f_2, ..., f_i\} (1 \le i \le N) \quad (3)$$

Based on each of the N feature sets, an NNA predictor was constructed and tested with Jackknife cross-validation test. With N overall accurate prediction rates, positive accurate rates and negative accurate rates calculated, we obtain an IFS table with one column being the index i and the other three columns to be the overall accurate rate, positive accurate rate and negative accurate rate, respectively. $S_{optimal}=S_n$ is regarded as the optimal feature set if and only if the row with index n contains the highest overall accurate rate.

2.8 Obtaining important features for analysis

The final feature set, in which the features are regarded as important for the prediction, is obtained as the union of the 13 individual feature sets. This final feature set is used for subsequent analysis.

3 Results

3.1 Results of feature selections and feature set integration

There are two lists in each of the 13 mRMR outputs. The first one is MaxRel table ranking each feature according to the relevancies to the target. Only the second list, the mRMR table that ranks the features according to both relevancies and redundancies, was used for feature selection. Based on the mRMR feature list, IFS was run for each predictor, yielding the IFS tables. Table 1 shows the number of features selected by the IFS processes for the 13 independent predictors and the highest overall accurate rate obtained by Jackknife cross-validation. Combining the 13 individual optimal feature sets, we obtained the union optimal feature set which contained 1014 features.

Table 1**.** The number of features in each individual optimal feature set for each classifier.

The index of classifier	The number of features	The overall accurate rate in Jackknife cross-validation
1	95	0.8125
2	226	0.8304
3	52	0.8928
4	234	0.8036
5	72	0.8125
6	42	0.8125
7	59	0.7321
8	138	0.8750

9	66	0.7321
10	77	0.7768
11	62	0.8571
12	94	0.7857
13	84	0.7523

3.2 Results of feature clustering and the evaluation of its reliability

For further study of nitration mechanisms, a feature clustering was done to the final selected features described above. 5 different feature groups were constructed based on the physicochemical and biochemical properties: alpha and turn propensities, beta propensities, composition, hydrophobicity, physicochemical properties. Among the 1014 features selected, 712 ones were successfully clustered in these groups. Figure 1 shows the number of features located in each feature group.

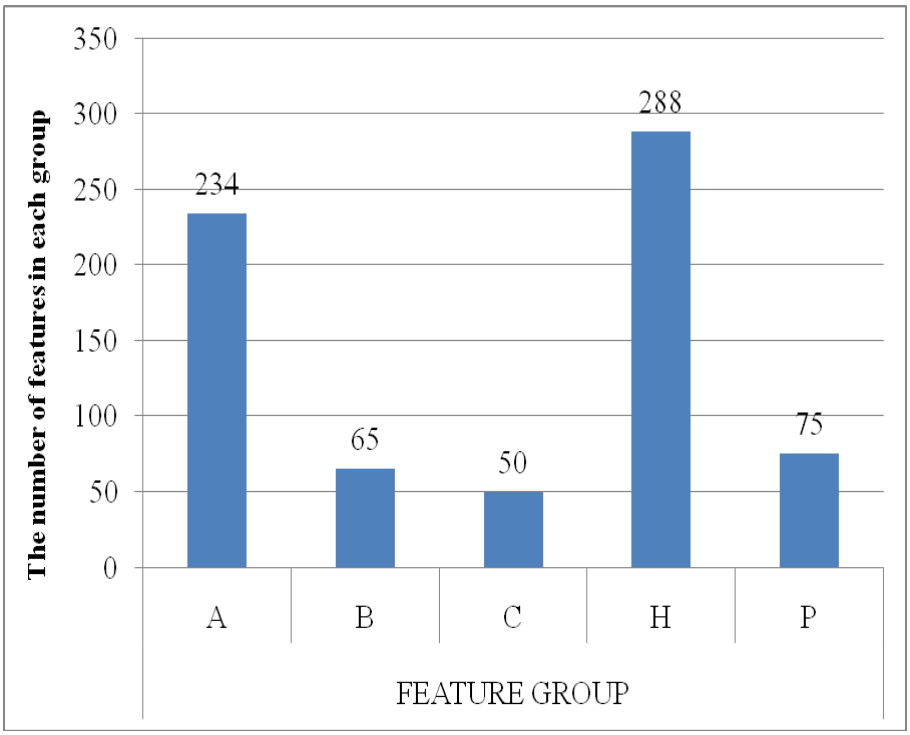

Figure 1. The number of selected features in each feature group. A stands for alpha and turn propensities, while B for beta propensities, C for composition, H for hydrophobicity, and P for physicochemical properties.

To assure the reliability of the union optimal feature set and its clustering, we applied the same method of clustering to the 13 individual optimal feature sets of the 13 dataset, respectively. Based on the grouping results, we got a clustering

vector $\mathbf{V}_i = [n_A, n_B, n_C, n_H, n_O, n_P]^T,\ i \in [1,13]$, where n_A, n_B, n_C, n_H, n_O, n_P represent the number of features grouped into the group of "alpha and turn propensities", "beta propensities", "composition", "hydrophobicity", "others", and "physicochemical properties". We also got the similar vector V for the union optimal feature set. Using a metric of vector angle cosine which is similar to the similarity defined for our NNA classifiers between V and each of the V_i, we could evaluate the similarity of clustering results between union optimal feature set and each of the individual optimal feature sets (see Table 2). It is easy to see that for all the 13 individual dataset, the clustering similarity score was higher than 0.9, indicating the high reliability of our union optimal feature set and its clustering.

Table 2**.** The similarity score of clustering vector between the union optimal feature set and each of the 13 individual optimal feature set.

The index of classifier	The similarity score
1	0.9700998
2	0.9602483
3	0.9059108
4	0.9927712
5	0.972646
6	0.9814104
7	0.95496
8	0.9842654
9	0.9838158
10	0.9829018
11	0.911443
12	0.9817525
13	0.9888647

4 Discussion

According to our analysis, the features about hydrophobicity and α and turn propensities related to the secondary structure of neighboring residues are essential to the microenvironment determining the protein tyrosine nitration (PTN). With a hydropathy index, tyrosine is mildly hydrophilic, a characteristic which is explained by the hydrophobic aromatic ring carrying a hydroxyl group. In consequence tyrosine is often surfaced and exposed in proteins (only 15% of tyrosine residues are at least 95% buried) and should thus be available for nitration [10]. Hydrophobicity of the residues surrounding the target tyrosine seems to play an important role in determining susceptibility towards PTN [10].

It has been studied that many factors favor the selectivity of nitration of Tyr residues: (i) The accessibility of the Tyr residue to nitrating agents; (ii) The presence of the Tyr residue in a loop structure formed by residues Gly or Pro; and (iii) The presence of the Tyr in proximity to a negatively charged residue [10]. Nitration of residues may be favored by their proximity to the negatively charged phosphate backbone of DNA in the nucleosome. A combination of the physi-chemical features such as the positions in the secondary structure (α and turn propensities), the accessibility of Tyr to the nitrating species and their proximities to Cys, or negatively charged residues which are related to hydrophobicity may be responsible for the nitration of Tyr sites [10-11].

Tyrosine nitration has been revealed as a relevant post-translational modification linked to nitro-oxidative stress conditions and pathophysiology. Our investigation may provide some useful insights to elevate protein tyrosine nitration from a biomarker to an important post-translational modification.

5 Conclusion

In this paper, we described a novel computational approach to analyze tyrosine nitration based on experimentally verified tyrosine nitration sites. Nearest Neighbor algorithm is armed by a feature selection process combining mRMR and IFS. Our result may provide insights and knowledge to protein tyrosine nitration and induce further study of the topic.

Acknowledges

This work was funded by the China Ministry of Science and Technology 973 Project (No. 2006CB500702), the Shanghai Commission of Education Science and Technology Innovation Fund (Grant No.08 ZZ41), the Shanghai Committee of Science and Technology (09JC1406600), grant from the Key Research Pro180 gram (CAS) (KSCX2-YW-R-112) and Shanghai Leading Academic Discipline Project (J50101).

References

[1] Turko IV, Li L, Aulak KS, Stuehr DJ, Chang JY, Murad F: Protein tyrosine nitration in the mitochondria from diabetic mouse heart. Implications to dysfunctional mitochondria in diabetes. J Biol Chem 2003, 278(36):33972-33977.

[2] Lee JR, Kim JK, Lee SJ, Kim KP: Role of protein tyrosine nitration in neurodegenerative diseases and atherosclerosis. Arch Pharm Res 2009, 32(8):1109-1118.

[3] Gunaydin H, Houk KN: Mechanisms of peroxynitrite-mediated nitration of tyrosine. Chem Res Toxicol 2009, 22(5):894-898.

[4] Radi R, Peluffo G, Alvarez MN, Naviliat M, Cayota A: Unraveling peroxynitrite formation in biological systems. Free Radic Biol Med 2001, 30(5):463-488.

[5] Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22(13):1658-1659.

[6] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M:

AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008, 36(Database issue):D202-205.

[7] Huang T, Cui W, Hu L, Feng K, Li YX, Cai YD: Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS ONE 2009, 4(12):e8126.

[8] Huang T, Shi XH, Wang P, He Z, Feng KY, Hu L, Kong X, Li YX, Cai YD, Chou KC: Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 2010, 5(6):e10972.

[9] Peng H, Long F, Ding C: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005, 27(8):1226-1238.

[10] Souza JM, Daikhin E, Yudkoff M, Raman CS, Ischiropoulos H: Factors determining the selectivity of protein tyrosine nitration. Arch Biochem Biophys 1999, 371(2):169-178.

[11] Haqqani AS, Kelly JF, Birnboim HC: Selective nitration of histone tyrosine residues in vivo in mutatect tumors. J Biol Chem 2002, 277(5):3614-3621.