THE PREDICTION OF TYROSINE SULFATION SITE IN PROTOEIN BY ANALYZING AMINO ACID COMPOSITION

DỰ ĐOÁN PROTEIN TYROSINE SULFATION DỰA VÀO CÁC PHÂN TÍCH TRÊN AMINO ACID

Khuong T. T. Pham

The University of Danang, University of Technology and Education; pttkhuong@ute.udn.vn

Abstract - Nowadays, there are many post-translational modifications (PTM) to be discovered and it hdelps to explore the secrets of life. PTM, a sulfate group is embedded to a Tyrosine residue called Protein Tyrosine Sulfation. The discoveries of these proteins have related to various biological processes and many different diseases. The most recent findings use traditional experimental methods to explore Sulfation sites; however, few factors such as an high-priced cost and the time taken should be mentioned. This study focuses on developing a bioinformatics method based on AA composition in order to explore Sulfation site. A work builds the training model from 483 experimentally and verifies Sulfation proteins by an inquiry in four features including 20D Binary code, AAC, Blosum62, and PSSM. Selected features will be evaluated by 5-fold cross validation and the model constructed by PSSM feature, show the best result with 4 factors: 94.96% for sensitivity, 95.10% for specificity, 95.09% for accuracy and 77.91% for MCC measurements.

Key words - Tyrosine Sulfation; post-translational modification; PTM; support vector machine; SVM.

Tóm tắt - Ngày nay, có rất nhiều sửa đổi sau dịch thuật (PTM) được phát hiện và điều này giúp khám phá nhiều bí mật của cuộc sống. PTM mà một nhóm sulfate được đính vào amino acid Tyrosine được gọi là Protein sulfation. Những khám phá về các protein này có liên quan đến các quá trình sinh học khác nhau và nhiều bệnh lý khác. Hầu hết các phát hiện gần đây sử dụng các phương pháp thử nghiệm truyền thống để khám phá các vị trí Sulfation; tuy nhiên một số yếu tố cần được đề cập như chi phí cao và thời gian thực hiện. Nghiên cứu này tập trung vào phát triển phương pháp tin sinh học dựa trên thành phần AA để xác định vị trí Sulfation. Mô hình được xây dựng từ 483 protein được xác định bằng các phương pháp thực nghiệm chính xác trong thực tế. Bốn đặc trưng được lựa chọn gồm mã nhị phân 20D, AAC, Blosum62 và PSSM. Mô hình được xây dựng từ PSSM, mang lại hiệu suất tốt nhất với Sn, Sp, Acc và số đo MCC lần lượt là 94,96%, 95,10%, 95,09% và 77,91%.

Từ khóa - Protein Tyrosine Sulfation; biến đổi hậu dịch; PTM; học máy; SVM;

1. Introduction

In 1954 [1], the Tyrosine Sulfation was firstly explored in peptide derived from bovine fibrinogen and it has an effect on a few proteins and peptides. Many researchers explored that the proteins were detected from a number of organisms, including rat, cow, human [2, 3, 4, 5]. The protein has an influence on regulating protein structure, function and physicochemical properties. Moreover, these proteins have been experimentally proved to be crucial to protein proteolytic process regulation, extracellular protein–protein interactions, intracellular protein transportation modulation [6–8], and various path physiological processes, for example atherosclerosis and virus HIV [9–10]. To discover more Sulfation's molecular mechanism, finding the way to define protein Tyrosine Sulfation play an important role. At present, there are several traditional experimental approaches to detect the sites [6]; however, they are time consuming and capable of processing data on a small-scale. Until now, there are no convention to detect the Sulfation proteins with optimized cost and effectiveness, especially for large data [6].

Until now, the drawbacks of traditional practical methods are reasons for the number of identified Sulfation protein to be limited as identification of Sulfated substrate proteins help to delve into valuable knowledge to figure out the molecular mechanism of Sulfation and conquering a major barrier in this area. Computation approaches should be combined to traditional proteomic technologies in order to tackle cost and time taken in laboratory. Analysing and defining the properties of bioinformatics technologies are applied for forecasting the position of Tyrosine. In this paper, the computational approaches are carried out on amino acid compositions to define the position of Sulfation sites effectively and accurately. Four selected feature including BLOSUM62, 20D binary code, amino acid composition, and position specific scoring matrix are used to define Sulfation sites.

2. Material and methods

2.1. Collecting data

As can be seen from table 1, a training data was built from UnitProtKB (release 2018_06) with the keyword “Sulfatyrosine”, “Sulfation protein”, which contains 483 proteins covering 1056 SulfatedTyrosine sites. To generate the training set, the length of short fragments equal to 21 were extracted at which the Sulfated or non-Sulfated Tyrosine residues was at the center. After that, the data including 705 positive data and 6490 negative ones was used as training data. Applying the same way for testing data, the data was gathered from Pre-Sulsite Data [11] as the table. Finally, testing data were 122 Sulfated sites and 515 non-Sulfated ones after detaching sequence fragments from 79 proteins as training data.

Our study focuses on the sequence-based analysis of substrate site specificity of Sulfated Tyrosine. The five-fold cross validation is used to assess training model in the possibility of recognizing the position of Sulfated site. Based on the performance of training models, the best one was further checked by using the independent testing set. Moreover, in both training and testing data could exist the probability of homologous sequences. Consequently, detaching the homologous sequences is required to

enhance the high dependability. In conclusion, there are 705 and 6490 corresponding the numbers of positive and negative fragments for the training data and 122 and 515 ones for the testing data.

Table 1. *Sulfation's Statistical data*

	Data Source	Number of Proteins	Sulfation sites	Non-Sulfation sites
Training Dataset	UniprotKB (release 2018_06)	483	1056	-
Testing dataset	PreSulsite Data	79	130	-
Spliting fragments with removing similarity ones				
Training Dataset	Uniprot DB	483	705	6490
Testing dataset	PreSulsite Data	79	122	515

2.2. Features of the study

Support vector machine was chosen to build the predicting models to detect Sulfation sites based on a few preferred features. The part sequences with length 21 were extracted from our data, which contained the Sulfated Tyrosine in the center. One of the most common solutions is an orthogonal binary coding mechanism, also called 20D binary code, and it is used to encode amino acids into numeric vectors, e. For example, Cysteine (C) was converted to "01000000000000000000". Finally, there was k vectors {xi, i=1, 2, 3, 4 ..., k} to be converted from k fragment sequences. To define the differences between Sulfation and non-Sulfation data, a label was signed as the class for each vector. For AAC feature, the length of vector x_i is 21. Note that happening frequency of 20 AAs was also found in extracted fragments. Based on alignments of amino acid sequences, the BLOSUM62 was constructed from peptides with no more than 62% identity.

PSI-BLAST [12] generates PSSM profiles which show numerous sequence arrangements. The matrix of (2n+1)×20 elements extracted from the PSSM profile on which substrate site was set in the center, and the length of short fragment is 21.

2.3. Evaluating the learning model

The predictive model was built by a public SVM library (LIBSVM) [13] from training data. This way applies a kernel function to transform input data into a higher dimensional space; after that draw a hyper-plane to distinguish the two classes with maximal margin and minimal error based on binary classification. In learning SVM classifier part, the radial basis function (RBF) was used as the kernel function to process data. To enhance the performance, there are two supporting factors including gamma and cost. Each feature was predicted by LIBSVM library. SVM classifier was utilized to evaluate the probability and the result was considered as the best feature.

$$K(S_i, S_j) = \exp(-\gamma \left\| S_i - S_j \right\|^2) \quad (1)$$

To evaluate the predicting efficiency for each feature, five-fold cross validation was executed to find the best final model. Five approximately equal sized subgroups were used to split the processed data. To enhance the results, and these evaluation processes was remade five times.

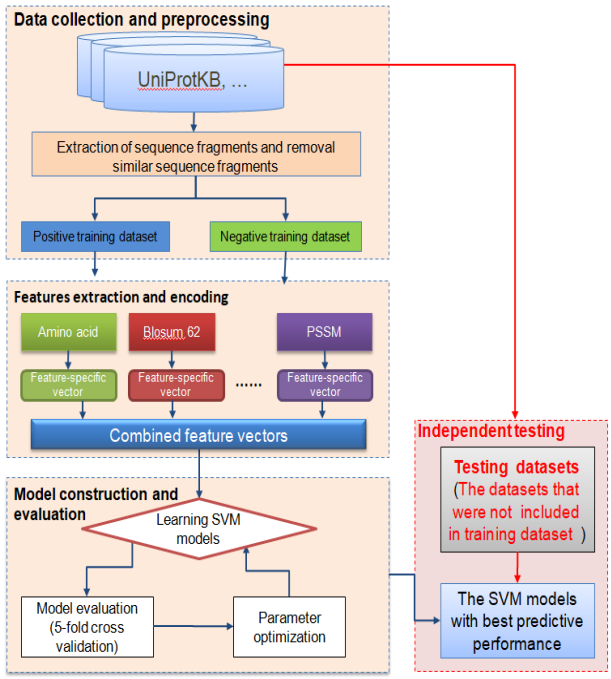

Figure 1. *Analytical flowchart for processing Sulfation data*

A single estimation was generated based on the five validation results. Certainly, the estimated processs will raise ependability of assessment because the whole data was investigated in general. To measure the training model in the predicting performance, accuracy-Acc, sensitivity-Sn, specificity-Sp, and Matthews Correlation Coefficient-MCC are four measures to be used.

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (2)$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (3)$$

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}} \quad (4)$$

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

where four values include TP for the numbers of true positives, TN for the numbers of true negatives, FP for the numbers of false positives and FN for the numbers of false negatives.

Next, a last independent testing was carried out and the results of cross-validation test with greatest performance would be chosen. A proportion of accurate forecast is considered for positive data, or sensitivity, while percentage of negative ones was assessed as specificity. Accuracy represents the general percentage of accurately foretell Sulfation fragments. The MCC reflects a a more accurate measure of the quality of binary classification.

3. Results

3.1. Interaction of 20 amino acids around Tyrosine Sulfation

The research pays attention to identifying the Sulfation sites by investigating common frequency of 20 kind of

amino acids surrounding Tyrosine. As can be seen from the figure 2, there are some curious different things in the occurrence frequency of 20 amino acids of positive and negative data. The figure 2A represents, at Sulfation sites, Aspartic acid (D) and Glutamic acid (E) which has a higher frequency. Associating to figure 2B, the repetition of 20 amino acids at each position was shown apparently. At the position 0, the percentage of amino acid Tyrosine accounts for the highest one and the appearance of Glutamic (E) acid and Aspartic (D) keeps an eye on its recurrence. Interestingly, remarkable things are D residue at position -3, -2, -1, 1 and 2 with over 12% and E at position -3, -1 and 1 with near 10%. Therefore, based on sequence-based analysis results, the amino acid composition has a relationship to the distribution of flanking amino acids near the Sulfation sites in forming of protein Sulfation.

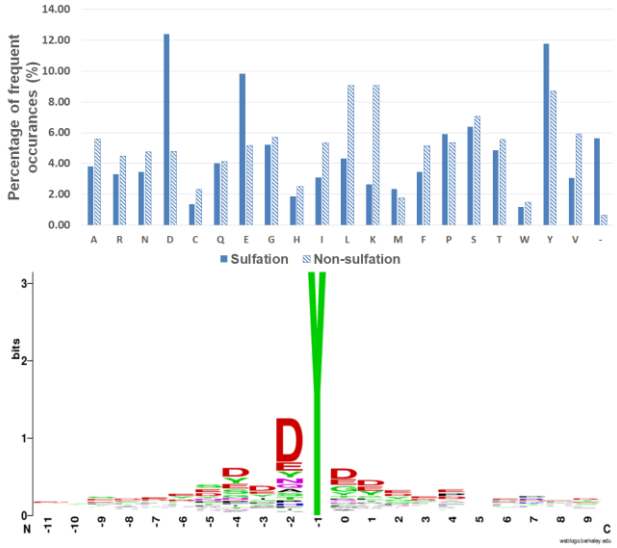

Figure 2. *Influence of 20 amino acid around Sulfation sites A- Analyse in AA composition of Sulfation and non-Sulfation data. B – Frequency of amino acid of Sulfation's training dataset at each position*

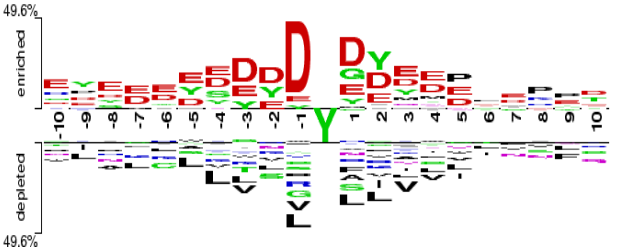

Figure 3. *Comparison of TwoSampleLogo about analyzing occurence frequencies of 20 amino acid*

TwoSampleLogo [14] is a powerful web-based tool to find statistically prominent differences in position-specific symbol compositions between two data set. Note that, the center position of fragment sequences is Tyrosine residue and range value is from -10 to 10 corresponding to the position of flanking amino acid. Comparing between 705 Sulfation sites and 6490 non-Sulfation sites, the negatively charged amino acids including Glutamic (E) and Aspartic (D) combined the highest radio at positions -10 ~ -1, and 1 ~ 5 (with p-value < 0.05). Moreover, Tyrosine(Y) and Glycine (G) appear at few positions -5 ~ -1, 1 ~ 3. It is quite reverse with positive dataset, the beingness of hydrophobic residue in negative one was at position -9 ~ 5 near the non-Sulfation sites. Moreover, the investigation points out that the distance among amino acid characteristic in sequence plays an important role in discriminating between Sulfation sites and non-Sulfation ones.

3.2. Evaluation by Cross-validation test

To detect the Sulfation site, the learning models were estimated by many features to get great ones. There are four predictive measurement units to be mentioned including Acc, Sp, Sn and MCC) to be used to evaluate the best model. From Table 2, we can see that the lowest result of built model with AAC feature is at 93.51% and 93.33% for sensitivity, 93.53% for specificity, and 72.35% for MCC. Besides, the results show that the best model with PSSM feature should be chosen with sensitivity at 94.96 %, specificity at 95.10%, accuracy at 95,09% and MCC at 77,91%. Therefore, PSSM was considered as a potential feature to build the potential predictive model.

Table 2. *Evaluation result by cross validation test on the SVM model*

Learning features	Sn	Sp	Acc	MCC
20D Binary code (AA)	92,70%	95,07%	94,83%	76,68%
BLOSUM62	90,64%	94,61%	94,20%	74,34%
Amino Acid Composition (AAC)	93,33%	93,53%	93,51%	72,35%
Position-specific scoring matrix (PSSM)	**94,96%**	**95,10%**	**95,09%**	**77,91%**

Evaluation of Sulfation's predictive models

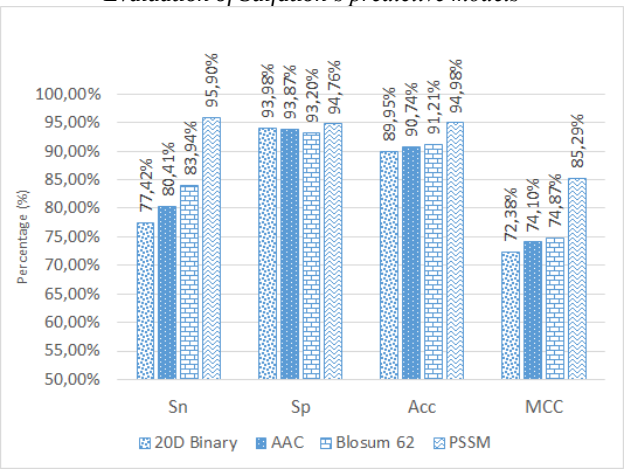

Figure 4. *Comparison of independent testing performance between 20D Binary, AAC, Blosum62 and PSSM*

Based on PSSM feature, an independent test was carried out in order to recheck potential model through five-fold cross-validation. For independent test, there are 122 for positive data and 515 for negative one. It can be seen from Fig. 4, there are widely disparity of four models in the percentage of Sn, Sp, Acc and MCC at 77.42%, 93.98%, 89.95%, 72.38 respectively for 20D Binary code model, 80.41%, 93.87%, 90.74% and 74.10% respectively for BLOSUM62 model, and 83.94%, 93.20%, 91.21%, 74.87% respectively for AAC model and, last one, 95.9%, 94.76%, 94.98%, 85.29%

respectively for PSSM model. Table 3 is the detailed independent testing results. In conclusion, SVM-trained model which is built based on PSSM feature achieved the best one in our research.

Table 3. *The detailed independent testing results between our methods*

Selected features	Sn (%)	Sp (%)	Acc (%)	MCC (%)
20D Binary code (AA)	77,42	93,98	89,95	72,38
BLOSUM62	80,41	93,87	90,74	74,10
Amino Acid Composition	83,94	93,20	91,21	74,87
Position-specific scoring matrix (PSSM)	95,90	94,76	94,98	85,29

4. Conclusion

The Sulfation sites are identified by analyzing amino acid composition. Through image of TwoSampleLogo, the relationship between amino acids surround Sulfated Tyrosine residue is represented between Sulfation and non-Sulfation site. Indeed, the paper shows that the model with PSSM feature is chosen based on the analysis of many steps through four mentioned measurement units.

REFERENCE

[1] F.R. Bettelheim, Tyrosine-O-sulfate in a peptide from fibrinogen, J. Am. Chem. Soc. 76 (1954) 2838–2839.

[2] P.A. Baeuerle, W.B. Huttner, Tyrosine Sulfation of yolk proteins 1, 2, and 3 in Drosophila melanogaster, J. Biol. Chem. 260 (1985) 6434–6439.

[3] P. Rosa, G. Fumagalli, A. Zanini, W.B. Huttner, The major tyrosine-sulfated protein of the bovine anterior pituitary is a secretory protein present in gonadotrophs, thyrotrophs, mammotrophs, and corticotrophs, J. Cell Biol. 100(1985) 928–937.

[4] A. Hille, P. Rosa, W.B. Huttner, Tyrosine Sulfation: a post-translational modification of proteins destined for secretion?, FEBS Lett 177 (1984) 129–134.

[5] M.C. Liu, S. Yu, J. Sy, C.M. Redman, F. Lipmann, Tyrosine Sulfation of proteins from the human hepatoma cell line HepG2, Proc. Natl. Acad. Sci. USA 82 (1985) 7160–7164.

[6] Y.H. Yu, A.J. Hoffhines, K.L. Moore, J.A. Leary, Determination of the sites of tyrosine O-Sulfation in peptides and proteins, Nat. Methods 4 (2007) 583–588.

[7] G. Sahota, G.D. Stormo, Tyrosine Sulfation: a modulator of extracellular protein–protein interactions, Chem. Biol. 7 (2000) R57–R61.

[8] W.B. Huttner, Protein tyrosine Sulfation, Trends Biochem. Sci. 12 (1987) 361–363.

[9] E. Koltsova, K. Ley, Tyrosine Sulfation of leukocyte adhesion molecules and chemokine receptors promotes atherosclerosis, Arterioscler. Thromb. Vasc. Biol. 29 (2009) 1709–1711.

[10] J. Liu, S. Louie, W. Hsu, K.M. Yu, H.B. Nicholas, G.L. Rosenquist, Tyrosine Sulfation is prevalent in human chemokine receptors important in lung disease, Am. J. Respir. Cell Mol. Biol. 38 (2008) 738–743.

[11] S.Y Huang, S.P. Shi, J.D. Qiu, X.Y. Sun, S.B. Suo, R.P. Liang, PredSulSite: Prediction of protein tyrosine Sulfation sites with multiple features and analysis. Anal. Biochem. 428(2012) 16-23

[12] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997. 25(17): p. 3389-402.

[13] Chang, C.-C. and C.-J. Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011. 2(27): p. 1-27.

[14] Vacic, V., L.M. Iakoucheva, and P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics, 2006. 22(12): p. 1536-1537.

(The Board of Editors received the paper on 12/10/2018, its review was completed on 24/12/2018)