***MATCH***
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **71** (2014) 227-240

ISSN 0340 - 6253

# SulfoTyrP: A High Accuracy Predictor of Protein Sulfotyrosine Sites

**Cangzhi Jia[a], Yusen Zhang[b], Zhiping Wang[a,]***

[a]*Department of Mathematics, Dalian Maritime University, Dalian 116026, China;*

[b]*School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China*

(Received February 25, 2013)

**Abstract:** Tyrosine sulfation is a post-translational modification widely distributed in eukaryotic proteins. The prerequisite to reveal its biological role which is largely unknown is identifying more protein sulfotyrosine sites. However, previous computational methods only achieved limited accuracy. In this paper, we propose a novel tool named SulfoTyrP with four designed strategies to predict protein sulfotyrosine sites. Weight parameters in support vector machine (SVM) are optimized for the first time to solve the problem of unbalanced datasets and this approach is proved to perform better than the widely used under-sampling approach for our datasets. Moreover, bi-profile Bayes and composition moment vector (CMV) are used to obtain rationally designed features to highlight the contribution of acidic and hydrophobic amino acids. Using SulfoTyrP, we get a sensitivity of 80.65%, an accuracy of 94.51%, Matthew's Correlation Coefficient (MCC) of 0.779 in jackknife cross-validation evaluations, an average sensitivity of 77.78% and an average ACC of 93.89% in three independent tests. Compared with other published tools, SulfoTyrP can get higher sensitivity and accuracy. We not only propose a high accuracy method to predict protein sulfotyrosine site, but also provide insights into improving the efficiency of other bioinformatics tools.

* Corresponding author: zpwangdlmu@yahoo.com.cn

## 1. Introduction

Tyrosine sulfation is a ubiquitous post-translational modification of eukaryotic proteins [1,2]. It has been found to play vital roles not only in many physiological processes such as leukocyte trafficking, cellular adhension, immune function and glycopeptide hormone activity but also involved in some pathological processes such as virus infection, atherosclerosis and lung disease. This post-translational modification is mediated by tyrosylprotein sulfotransferase (TPST, EC 2.8.2.20), which catalyze the transfer of sulfate from adenosine 3′-phosphate 5′-phosphosulfate to the hydroxyl group of peptidyltyrosine residue (Fig. 1 in [3]). Although some common characteristics of the sequence that favor sulfation have been summarized [4,5], no conserved acceptor sequence motif could be defined for TPST[6,7].

Compared to its wide distribution in nature, the known functions of tyrosine sulfation is only tip of the iceberg. To understand the whole biology of tyrosine sulfation, more proteins with sulfotyrosine sites should be identified. As experimental effort is often laborious and expensive, much effort has been devoted to develop in silico tools for predicting protein sulfotyrosine sites [6,8-13]. Because no consensus sequence motif could be defined, the prediction of protein sulfotyrosine sites is not an easy task. So the performances of existing prediction methods are not satisfied. It is worth noting that the prediction accuracy has great improvement in the recent paper [11] by incorporating the conservation, disorder, and physicochemical properties of amino acids. But the shortcoming of this method is the low sensitivity, which only achieved 66.67% in jackknife test. In the nearest study [12], the sensitivity has been remarkably increased to 92.00%, and the MCC increased to 0.8897 by considering physicochemical properties of amino acids and residue sequence order information. But the hydrophobicity of residues and the effect of amino acid location as well as other information should be considered.

In this study, four strategies are used to improve the prediction accuracy of protein sulfotyrosine sites. Firstly, an updated and non-redundant dataset is established. Secondly, the weight parameters (W1 and W-1) of the support vector machine (SVM) are introduced for the first time to solve the problem of unbalanced dataset. Thirdly, five window sizes (lengths of the sequence segments around the sulfated and non-sulfated tyrosines) are tried and optimized. Fourthly and the most importantly, two potent methods including bi-profile Bayesian amino acid profile (BPBAA), and the composition moment vectors (HCMV and ACMV) are used for sequence feature extraction. In combination of the above strategies, a

novel tool named SulfoTyrP is developed. It proves to significantly outshine all of the existing predictors for protein sulfotyrosine sites.

## 2. Materials and methods

Similar to the research flow used in [6,12], we present the flowchart of the proposed method in Fig.1. The method comprises four major steps: (i) collecting and processing data, (ii) window size optimization, (iii) extraction of features, and (iv) creation and evaluation models.

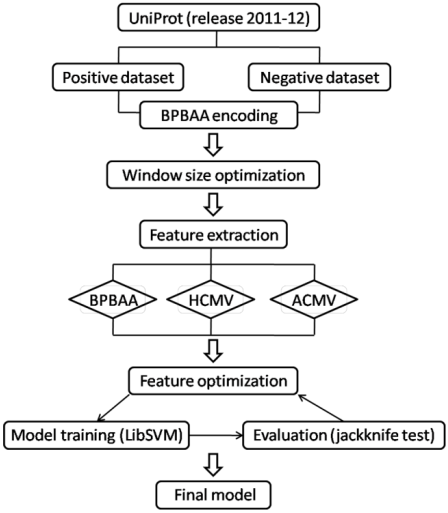


**Figure1.** System flow of SulfoTyrP

### 2.1 Datasets

The protein sequences containing experimentally verified sulfotyrosine sites are collected from UniProt database (release 2011_12 in [13]). Then the sequence segments of different lengths [L=7 (-3 to +3), 9 (-4 to +4), 11 (-5 to +5), 13 (-6 to +6), 15 (-7 to +7)] around sulfotyrosine sites and non-sulfotyrosine sites are extracted as positive and negative training sets. The identical sequence segments are removed to avoid the overestimation of prediction accuracy (Table S1).

As a comprehensive and unbiased comparison with the existing methods, the training dataset [11] and independent test datasets recently constructed in [11,12] are also used. The training dataset in [11] includes 102 sulfotyrosine sites and 629 non-sulfotyrosine sites while the test dataset in [11] contains 27 sulfotyrosine sites and 69 non-sulfotyrosine sites. The

lengths of the sequence segments are changed from 9 to 11. And the final training dataset of length 11 includes 98 sulfotyrosine sites and 624 non-sulfotyrosine sites while the test dataset contains 25 sulfotyrosine sites and 67 non-sulfotyrosine sites. And the final test dataset from [12] contains 17 sulfotyrosine peptide sequences and 51 non-sulfotyrosine peptide sequences of length 11.

## 2.2. Sequence feature extraction

*2.2.1. Bi-profile Bayes profiles*

In this method, each sequence segment can be encoded by a probability vector $P=(p_1,p_2,\cdots,p_n,p_{n+1},\cdots,p_{2n})$. One advantage of bi-profile Bayes method is that the feature vectors are encoded in a bi-profile manner containing information from both positive samples and negative samples. The other advantage of this method is dealing with an unbalanced dataset comprising a small positive subset and a large negative subset. Details about the bi-profile Bayes method are shown in reference [14].

Two kinds of reduced sequences are respectively generated according to the acidity and hydrophobicity of twenty amino acid. Based on the acidity, twenty amino acid residues are classified into three groups: acidic amino acids (A): {D, E}; basic amino acids (K): {K, H, R}; neutral amino acids (N): {A, C, F, G, I, L, M, N, P, Q, S, T, V, W, Y}. Similarly on the hydrophobicity, twenty amino acid residues are classified into another three groups: internal group (F): {F, I, L, M, V}, external group (D): {D, E, H, K, N, Q, R}, ambivalent group (C): {A, C, G, P, S, T, W, Y}. So two types of characteristic sequences can be respectively obtained composed of A, K, N and F, D, C.

*2.2.2. Composition moment vector (CMV)*

CMVs [15-17] are used to reflect the content and position of acidic amino acid residues and hydrophobic amino acid residues around the sulfotyrosine sites and non-sulfotyrosine sites:

$$CMV^N=\frac{\sum_{j=1}^{n_N} n_{Nj}}{\prod_{d}^{k}(L-d)},\ CMV^K=\frac{\sum_{j=1}^{n_K} n_{Kj}}{\prod_{d}^{k}(L-d)},\ CMV^A=\frac{\sum_{j=1}^{n_A} n_{Aj}}{\prod_{d}^{k}(L-d)}$$

$$CMV^F=\frac{\sum_{j=1}^{n_F} n_{Fj}}{\prod_{d}^{k}(L-d)},\ CMV^D=\frac{\sum_{j=1}^{n_D} n_{Dj}}{\prod_{d}^{k}(L-d)},\ CMV^C=\frac{\sum_{j=1}^{n_C} n_{Cj}}{\prod_{d}^{k}(L-d)}$$

where L=11; $n_N$, $n_K$, $n_A$, $n_F$, $n_D$, $n_C$ are the total number of residues N, K A, F, D and C in the corresponding reduced amino acid sequences; $n_{Nj}$, $n_{Kj}$, $n_{Aj}$, $n_{Fj}$, $n_{Dj}$, and $n_{Cj}$ are the $j^{th}$ position of residues N, K, A, F, D and C in the corresponding reduced amino acid sequences and k (k=0,1) is the order of the CMV. If k=0, the CMV reduced to the content of the amino acid.

## 2.3. Support vector machine implementation and parameter selection

Support vector machine (SVM) is a set of related supervised learning methods used for classification and regression based on statistical learning theory. This method has been proven to be powerful in many fields of bioinformatics [14, 16, 18-22]. In this study, SVM is trained with LIBSVM package [23] to build the model and perform the prediction. Radial basis kernel function (RBF kernel) is used in our SVM model. For different input features, penalty parameters C and kernel parameters $\gamma$ are optimized using SVMcg in LIBSVM package based on 15-fold cross-validation.

Optimized weight parameters (W1 and W-1) are searched to solve the problem of unbalanced dataset. The value of W1 denotes the weight for positive samples and the value of W-1 denotes the weight for negative samples. The default values of W1 and W-1 are 1. When Wi=k, the penalty parameter C of class i is set to k*C, which means that the classification is prone to class i (i=1,-1). The main principle can be understood from the following SVM formulations.

The original SVM formulation is defined by:

$$\min \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{L}\xi_i$$

subject to $y_i[\omega^T \varphi(x_i) + b] \geq 1 - \xi_i$

where L is the number of samples and C is the penalty parameter.

After the consideration of the weight parameter, the SVM formulation is modified as:

$$\min \frac{1}{2}\omega^T\omega + C_+\sum_{i=1}^{p}\xi_i + C_-\sum_{j=p+1}^{q}\xi_j$$

subject to $y_i[\omega^T\varphi(x_i) + b] \geq 1 - \xi_i,$

where i=1,2, ..., p are the positive samples and j=p+1, p+2, ..., p+q are the negative samples; And $C_+$ is the cost for positive samples and $C_-$ is the cost for negative samples. The parameters C=22.6274, $\gamma = 2.8284$ and $C_+$=2 are used in our predictor, which resulted in the best predictive performance.

### 2.4. Performance assessments

Since jackknife test is considered as the most objective cross-validation method [24,25], we use it to evaluate our method. Sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthew's Correlation Coefficient (MCC) were used to quantify the performance of our method. They are defined as follows:

$$Sn=\frac{TP}{TP+FN},\ Sp=\frac{TN}{TN+FP},\ ACC=\frac{TP+TN}{TP+TN+FP+FN},$$

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

where TP, TN, FP and FN respectively denote the number of true positives (correctly predicted sulfotyrosine sites), true negatives (correctly predicted non-sulfotyrosine sites), false positives (falsely predicted sulfotyrosine sites) and false negatives (falsely predicted non-sulfotyrosine sites).

## 3. Results

### 3.1. Weight parameter optimization

The weight parameters (W1 and W-1) in SVM are used to solve the problem of unbalanced dataset in this study. For each training process, W1 value is first set to 1, 1.5 and 2, while W-1 value is set to 1. Then the refinement of W1 by setting step size to +0.1/-0.1 is performed around the previously identified W1 with the highest MCC value. It is worth to note that the performances of these models are significantly improved after the optimization of W1 parameter (Table S2).

### 3.2. Window size optimization

The datasets with different window sizes are established by extracting the sequence segments with different lengths [L=7 (-3 to +3), 9 (-4 to +4), 11 (-5 to +5), 13 (-6 to +6), 15 (-7 to +7)] around sulfotyrosine sites and non-sulfotyrosine sites. After the removal of redundant samples, the number of the samples in the datasets with different lengths is not equal (Table S1). To avoid this phenomenon, a common dataset containing 82 positive samples and 546 negative samples are established and used to find the best window size by the features of BPBAA. The detailed results of the jackknife test on different window sizes are shown in Table S2. We choose MCC value as evaluation criterion because it is considered to be more

objective than ACC value in evaluating the performance of a predictor, especially for unbalanced dataset [24]. The best MCC results achieved by different window sizes are listed in Table 1 and the corresponding ROCs of five models are presented in Fig. 2.

**Table 1.**The best jackknife results of models using BPBAA on different window sizes in the same dataset.

| Window size | W1 | W-1 | Sn (%) | Sp (%) | ACC (%) | MCC | AUC |
|---|---|---|---|---|---|---|---|
| 15 | 1.1 | 1 | 73.17 | 97.80 | 94.59 | 0.751 | 0.972 |
| 13 | 1.3 | 1 | 74.39 | 97.99 | 94.90 | 0.756 | 0.968 |
| **11** | **1.4** | **1** | **75.61** | **97.80** | **94.90** | **0.762** | **0.967** |
| 9 | 1.2 | 1 | 73.17 | 97.80 | 94.59 | 0.751 | 0.962 |
| 7 | 1.2 | 1 | 68.29 | 97.62 | 93.79 | 0.710 | 0.956 |

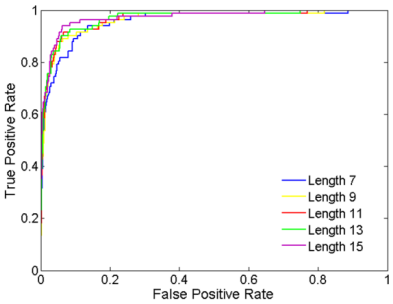


Figure 2. ROCs of five models on different lengths

Along with the window size changing from 7 to 15, the MCC and Sn values firstly increases then decreases, at last it reaches the highest MCC value of 0.762 and Sn value of 75.61% at the same window size of 11. Therefore, the dataset with the window size of 11 containing 93 positive samples and 545 negative samples is used for SVM training and testing.

### 3.3. Feature selection and model performance

Three kinds of features (BPBAA, HCMV, ACMV) were extracted for sequence representation. Among them, BPBAA contains the most abundant information and used as the basic features. Then we evaluated the prediction performances by the combination of these

features with increased complexity (BPBAA, BPBAA+HCMV, BPBAA+ACMV, BPBAA+ACMV+HCMV) and the results of the jackknife test are shown in Table 2. The corresponding ROCs are presented in Fig.3. It is found the prediction performances of sensitivity increased when BPBAA coupled with any of BPBAA+HCMV, BPBAA+ACMV and the best sensitivity of 80.65% achieved by the combination of BPBAA and HCMV. Although the best MCC of 0.794 achieved by the combination of BPBAA+ACMV+HCMV, the sensitivity which plays a crucial role for experiment identification, is a little lower than BPBAA+HCMV. At this point, the SVM-based predictor, SulfoTyrP was built up by using the BPBAA+HCMV feature extraction method with C=22.6274 and $\Upsilon = 2.8284$ to capture potential more potential sulfotyrosine sites.

**Table 2.** Predictive performances of models trained on different sequence encoding schemes in window size 11.

| Features | W1 | Sn (%) | Sp (%) | ACC (%) | MCC | AUC |
|---|---|---|---|---|---|---|
| BPBAA | 1.1 | 76.34 | 97.25 | 94.20 | 0.760 | **0.961** |
| BPBAA+HCMV | 1.5 | **80.65** | 96.88 | 94.51 | 0.779 | 0.959 |
| BPBAA+ACMV | 1.2 | 76.34 | 98.17 | 94.98 | 0.790 | 0.953 |
| BPBAA++HCMV+ACMV | 1.3 | 79.57 | 97.61 | 94.98 | **0.794** | 0.956 |

## 3.4. Comparing with other methods

Since no web server or executable program can be found for Niu *et al.*'s method [11], we use their training and testing dataset to test the SulfoTyrP. Then the jackknife and test results are compared with those of Niu *et al.*'s method shown in their paper (Table 3). Compared with the jackknife results, SulfoTyrP get higher accuracy than Niu *et al.*'s method with the Sn value of 4.76%, and the ACC value of 2.1%. While for the independent test, the prediction results of SulfoTyrP are a little higher than those of Niu *et al.*'s method with the Sn value of 2.52% and the ACC value of 0.61%. For there is no negative training dataset supplied, we use the same test set in Hung *et al.*'s [12] to compare SulfoTyrP with PredSulSite (Table 3). The prediction results of SulfoTyrP are higher than those of PredSulSite with the Sn value of 11.11%, and the ACC value of 0.76%, but it is a little lower with the Sp value of 3.38%.

We compare SulfoTyrP with Sulfinator [9] and SulfoSite [6] indirectly. As reported in Niu *et al.* [11] and Hung *et al.* [12], Niu *et al.*'s method and PredSulSite outperformed Sulfinator

and SulfoSite when tested on some newly identified sulfotyrosine sites (Table 3). From the table we can also find that, SulfoTyrP is more powerful than Sulfinator and SulfoSite.

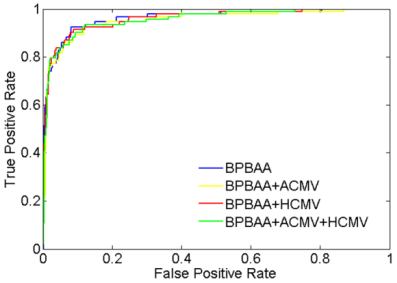


Fig.3 ROCs of different encoding SVM models using jackknife test

### 3.5. Comparison of weight parameter optimization and under-sampling

To compare with the under-sampling method, the dataset containing 93 positive samples and 545 negative samples is randomly divided into two parts: 72 positive samples/435 negative samples as training dataset and 21 positive samples/110 negative samples as testing dataset. After two times of repeat, three pairs of training and testing datasets (Ptrain-1, 2, 3; Ntrain-1, 2, 3; Ptest-1, 2, 3; Ntest-1, 2, 3) are obtained. In under-sampling method, we adapt the optimized ratio of positive/negative dataset as 1: 3 [14, 22]. This ratio is considered to retain the original distribution of negative samples and avoids losing diversity information. In addition, for each positive dataset, five negative datasets containing 216 negative samples are randomly collected from the full negative datasets. The final prediction results are the average of 15 results (Table 3). For the independent test results, the prediction models constructed by weight parameter optimization outshine those constructed by under-sampling with the Sn value of 1.59% and ACC value of 1.52%.

## 4. Discussion

Tyrosine sulfation is a ubiquitous post-translational modification in eukaryotic proteins but its physiological roles are still largely unknown. The process is greatly hindered by lacking an effective method to predict the sulfotyrosine sites. Low prediction sensitivity (the percentage of correctly predicted sulfotyrosine sites) is the main problem existing in previous methods, which is particularly important for crude screening at genomic level. In this study, a high

accuracy tool named SulfoTyrP is developed by using four strategies. Two of them are paramount and generally applicable for other bioinformatics predictors.

## 4.1. Weight parameter optimization in SVM to solve the problem of unbalanced dataset

Since the number of the positive samples is usually much less than that of negative samples [6,14,22], unbalanced dataset is a common problem which many bioinformatics predictors need to face. The consequence of this problem is low value of Sn but high value of Sp. The commonly used approach to solve such problem is under-sampling by reducing the number of negative samples [6,14,22]. However, this approach obviously leads to the loss of sequence information from the negative dataset.

**Table 3.** Comparisons with other methods on different datasets

| Dataset | Method | Sn (%) | Sp (%) | ACC (%) |
|---|---|---|---|---|
| Training dataset[11] | Niu's method | 66.67 | 93.80 | 90.01 |
| | SulfoTyrP | 71.43 | 95.35 | 92.11 |
| Test dataset[11] | SulfoSite | 74.07 | 97.10 | 90.63 |
| | Sulfinator | 77.78 | 95.65 | 90.63 |
| | Niu's method | 81.48 | 100 | 94.79 |
| | SulfoTyrP | 84.00 | 100 | 95.40 |
| Test dataset[12] | Sulfinator | 44.44 | 87.50 | 74.14 |
| | SulfoSite | 83.33 | 87.50 | 86.21 |
| | PredSulSite | 88.89 | 97.50 | 94.83 |
| | SulfoTyrP | 100.00 | 94.12 | 95.59 |
| Test dataset[ours] | Weight parameter | 77.78 | 96.97 | 93.89 |
| | Under sampling | 76.19 | 95.45 | 92.37 |

A novel approach is proposed in this study to overcome the problem of unbalanced dataset by optimizing the weight parameters (W1 and W-1) in SVM. After crude screening (with step size of 0.5) and fine screening (with step size of 0.1/-0.1), the optimized W1 parameters with highest MCC values can be obtained (Table S2). This approach is not only more reasonable in theory but also more effective in practice than under-sampling approach. As shown in Table 3,

the Sn and ACC values of the models established by weight parameter optimization are much higher than the models established by under-sampling.

## 4.2. Rational designed features on the basis of biochemical property of tyrosine sulfation

Bi-profile Bayes feature extraction is an informative method and has been successfully used in various bioinformatics tools for identifying protein methylation sites [14], caspase cleavage sites [22], malaria mitochondrial proteins [20], linear B-cell epitopes [26] and type III secreted effectors [27]. Most of these bioinformatics tools use bi-profile Bayes method to extract features directly from the amino acid sequences. This approach surely contains the most quantity of information but some of unique properties presented in a specific sequence segments are submerged in the complex information. Song *et al.* [22] extracted different types of sequence profiles (BPBRAAs) from the reduced amino acid sequences according to the predicted secondary structure, solvent accessibility and disordered probability of the amino acid residues. By combining these BPBRAAs with BPBAA, the MCC value for caspase cleavage sites obviously increases [22].

Since the acidity and hydrophobicity of amino acid residues adjacent to the tyrosine site is vital for the sulfation process [5, 8, 28], the reduced amino acid sequences are generated according to the acidity and hydrophobicity of them. Then composition moment vector (HCMV and ACMV) are used to extract features from the reduced amino acid sequences to highlight the contribution of the number and the position of the acidic and hydrophobic residues of an individual sequence sample. The prediction results are significantly improved by including these rationally designed features on the basis of the biochemical property of the sequence that favor tyrosine sulfation. The combination of BPBAA and HCMV notably improved the prediction sensitivity to 80.65%, while the combination of BPBAA and ACMV notably improved the prediction specificity to 98.17% (Table 2). Although the MCC of BPBPB+HCMV was not better than BPBAA+HCMV+ACMV, a model trained with BPBPB+HCMV is the most sensitivity of 80.65% as given in Table 2. For the main aim of the bioinformatics predictor is to capture potential post-translational modification sites from bulk data, the best predictive sensitivity model is selected as the final predictor SulfoTyrP.

*Acknowledgements*: The authors would like to appreciate the suggestions from the editor and anonymous reviewer, which contributed to great improvement of the presentation of this manuscript. This work is supported by Liaoning Provincial Natural Science Foundation of China (201102015), and the Fundamental Research Funds for the Central Universities Under contract (N0.2012TD032**,** N100405010, 3132013093).

## References

[1] K. L. Moore, Protein tyrosine sulfation: a critical posttranslation modification in plants and animals, *Proc. Natl. Acad. Sci. USA* **106** (2009) 14741–14742.

[2] P. Onnerfjord, T. F. Heathfield, D. Heinegard, Identification of tyrosine sulfation in extracellular leucine-rich repeat proteins using mass spectrometry, *J. Biol. Chem.* **279** (2004) 26–33.

[3] K. L. Moore, The biology and enzymology of protein tyrosine O-sulfation, *J. Biol. Chem.* **278** (2003) 24243–24246.

[4] J. W. Kehoe, C. R. Bertozzi, Tyrosine sulfation: a modulator of extracellular protein-protein interactions, *Chem. Biol.* **7** (2000) 57–61.

[5] M. J. Stone, S. Chuang, X. Hou, M. Shoham, J. Z. Zhu, Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins, *N. Biotechnol.* **25** (2009) 299–317.

[6] W. C. Chang, T. Y. Lee, D. M. Shien, J. B. Hsu, J. T. Horng, P .C. Hsu, T. Y. Wang, H. D. Huang, R. L. Pan, Incorporating support vector machine for identifying protein tyrosine sulfation sites, *J. Comput. Chem.* **30** (2009) 2526–2537.

[7] Y. 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.

[8] J. R. Bundgaard, J. Vuust, J. F. Rehfeld, New consensus features for tyrosine O-sulfation determined by mutational analysis, *J. Biol. Chem.* **272** (1997) 21700–21705.

[9] F. Monigatti, E. Gasteiger, A. Bairoch, E. Jung, The Sulfinator: predicting tyrosine sulfation sites in protein sequences, *Bioinformatics* **18** (2002) 769–770.

[10] F. Monigatti, B. Hekking, H. Steen, Protein sulfation analysis – A primer, *Biochim. Biophys. Acta* **1764** (2006) 1904–1913.

[11] S. Niu, T. Huang, K. Feng, Y. Cai, Y. Li, Prediction of tyrosine sulfation with mRMR feature selection and analysis, *J. Proteome Res.* **9** (2010) 6490–6497.

[12] S. Huang, S. Shi, J. Qiu., X. Sun, S. Suo, R. Liang, PredSulSite: Prediction of protein tyrosine sulfation sites with multiple features and analysis, *Anal. Biochem.* **428** (2012) 16–23.

[13] C. Uniprot, The universal protein resource (UniProt) in 2010, *Nucleic Acids Res.* **38** (2010), D142–D148.

[14] J. Shao, D. Xu, S.N. Tsai, Y. Wang, S. M. Ngai, Computational identification of protein methylation sites through bi-profile Bayes feature extraction, *PLoS One* **4** (2009) e4920.

[15] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, *BMC Bioinformatics* **9** (2008) 226–226.

[16] T. Liu, C. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, *J. Theor. Biol.* **267** (2010) 272–275.

[17] M. J. Mizianty, L. Kurgan, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, *BMC Bioinformatics* **10** (2009) 414–414.

[18] K. P. Exarchos, C. Papaloukas, T. P. Exarchos, A. N. Troganis, D. I. Fotiadis, Prediction of cis/trans isomerization using feature selection and support vector machines, *J. Biomed. Inform.* **42** (2009) 140–149.

[19] G. Jian, Y. Zhan, P. Qian, Prediction of subcellular localization for apoptosis protein: approached with a novel representation and support vector machine, *MATCH Commun. Math. Comput. Chem.* **67** (2012) 867–878.

[20] C. Jia, T. Liu, A. K. Chang, Y. Zhai, Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction, *Biochimie* **93** (2011) 778–782.

[21] D. M. Shien, T. Y. Lee, W. C. Chang, J. B. Hsu, J. T. Horng, P. C. Hsu, T. Y. Wang, H. D. Huang, Incorporating structural characteristics for identification of protein methylation sites, *J. Comput. Chem.* **30** (2009) 1532–1543.

[22] J. Song, H. Tan, H. Shen, K. Mahmood, S. E. Boyd, G. I. Webb, T. Akutsu, J. C. Whisstock, Cascleave: towards more accurate prediction of caspase substrate cleavage sites, *Bioinformatics* **26** (2010) 752–760.

[23] C. C. Chang, C. J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst.Technol.* **2** (2012) 27–27.

[24] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics* **16** (2000) 412–424.

[25] K. C. Chou, H. B. Shen, Recent progress in protein subcellular location prediction, *Anal. Biochem.* **370** (2007) 1–16.

[26] L. J. K. Wee1, D. Simarmata, Y. W. Kam, L. F. P. Ng, J. C. Tong, SVM-based prediction of linear B-cell epitopes using Bayes feature extraction, *BMC Genomics* **11** (2010) S21–S21.

[27] Y. Wang, Q. Zhang, M. Sun, D. Guo, High-accuracy prediction of bacterial type. III. Secreted effectors based on position-specific amino acid composition profiles, *Bioinformatics* **27** (2011) 777–784.

[28] R. Beisswanger, D. Corbeil, C. Vannier, C. Thiele, U. Dohrmann, R. Kellner, K. Ashman, C. Niehrs, W. B. Huttner, Existence of distinct tyrosylprotein sulfotransferase genes: molecular characterization of tyrosylprotein sulfotransferase-2, *Proc. Natl. Acad. Sc.i USA* **95** (1998) 11134–11139.