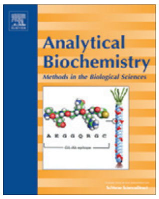
Analytical Biochemistry 428 (2012) 16–23

Contents lists available at SciVerse ScienceDirect

Analytical Biochemistry

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



PredSulSite: Prediction of protein tyrosine sulfation sites with multiple features and analysis

Shu-Yun Huang [a], Shao-Ping Shi [a], Jian-Ding Qiu [a,b,*], Xing-Yu Sun [a], Sheng-Bao Suo [a], Ru-Ping Liang [a]

[a] *Department of Chemistry, Nanchang University, Nanchang 330031, People's Republic of China*
[b] *Department of Chemical Engineering, Pingxiang College, Pingxiang 337055, People's Republic of China*

ARTICLE INFO

Article history:
Received 11 March 2012
Received in revised form 1 June 2012
Accepted 1 June 2012
Available online 9 June 2012

Keywords:
Support vector machine
Prediction of sulfotyrosine sites
10-Fold cross-validation

ABSTRACT

Tyrosine sulfation is a ubiquitous posttranslational modification that regulates extracellular protein–protein interactions, intracellular protein transportation modulation, and protein proteolytic process. However, identifying tyrosine sulfation sites remains a challenge due to the lability of sulfation sequences. In this study, we developed a method called PredSulSite that incorporates protein secondary structure, physicochemical properties of amino acids, and residue sequence order information based on support vector machine to predict sulfotyrosine sites. Three types of encoding algorithms—secondary structure, grouped weight, and autocorrelation function—were applied to mine features from tyrosine sulfation proteins. The prediction model with multiple features achieved an accuracy of 92.89% in 10-fold cross-validation. Feature analysis showed that the coil structure, acidic amino acids, and residue interactions around the tyrosine sulfation sites all contributed to the sulfation site determination. The detailed feature analysis in this work can help us to understand the sulfation mechanism and provide guidance for the related experimental validation. PredSulSite is available as a community resource at http://www.bioinfo.ncu.edu.cn/inquiries_PredSulSite.aspx.

Tyrosine sulfation was considered as a unique posttranslational modification (PTM)[1] that affected only a few proteins and peptides since it was first discovered in a peptide derived from bovine fibrinogen in 1954 [1]. However, during the 1980s, tyrosine sulfation was shown to be a common modification that was observed in secreted proteins from several organisms, including fruit fly [2], rat [3,4], cow [3,5], and human [6]. Evidence suggests that approximately 1% of all tyrosine residues of the total proteins in an organism can be sulfated [7]. As one of the most universal PTMs in secreted and transmembrane proteins, tyrosine sulfation occurs when tyrosylprotein sulfotransferase (TPST) catalyzes the transfer of a negatively charged sulfate from 3′-phosphoadenosine-5′-phosphosulfate (PAPS) to the hydroxyl group of tyrosine residue on a polypeptide [8]. Tyrosine sulfation has been experimentally demonstrated to be essential to extracellular protein–protein interactions, intracellular protein transportation modulation, and protein proteolytic process regulation [9–12], and it has been implicated in various pathophysiological processes, such as atherosclerosis, lung disease, and HIV infection [13–15]. In the overview, identification of protein tyrosine sulfation sites is of fundamental importance to understand the molecular mechanism of tyrosine sulfation in biological systems. Several conventional experimental approaches have been developed to identify sulfotyrosine sites [7,9,16]; however, they are laborious and have low throughput for large-scale data sets. It is because of the large variability of sulfation proteins [9] that there are no consensus features that accurately identify sulfation sites [17,18], and the characterization of this PTM has been hampered by the lack of general definite methods for its site determination. Therefore, the prediction of sulfation sites with computational approaches is desirable and necessary. Moreover, the sites that are predicted by computational models, especially models for performing large-scale predictions, could be of interest with respect to general implications for cell biology and biological experiments. Several predictive models have been developed for identifying the site of tyrosine sulfation.

Initially, Rosenquist and Nicholas discovered a test for acidic amino acids near the target tyrosine that resulted in a very good filtering criterion [19,20]. However, many other sulfotyrosine sites have no acidic residues in their flanking regions; for instance,

[*] Corresponding author at: Department of Chemistry, Nanchang University, Nanchang 330031, People's Republic of China.
E-mail address: jdqiu@ncu.edu.cn (J.-D. Qiu).

[1] *Abbreviations used:* PTM, posttranslational modification; TPST, tyrosylprotein sulfotransferase; PSSM, position-specific scoring matrix; LRR, leucine-rich repeat; ASA, accessible surface area; PWM, positional weighted matrix; SVM, support vector machine; SS, secondary structure; EBGW, encoding based on group weight; ACF, autocorrelation functions; kNN, *k*-nearest neighbor; FLD, Fisher linear discriminant; DT, decision tree; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews correlation coefficient; TP, true positive; FP, false positive; TN, true negative; FN, false negative; ROC, receiver operating curve; AUC, area under the ROC.

0003-2697/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ab.2012.06.003

Tyr30 in mouse Lumican can be sulfated, but no acidic amino acids existed within ±5 residues [9]. Subsequently, Yu and coworkers incorporated context-based rules and logical filters to develop a so-called position-specific scoring matrix (PSSM) to predict tyrosine sulfation sites [21]. Unfortunately, the capacity to find new or unseen PTM patterns is absent because of the static behavior of PSSMs [21,22]. To address such issues, Sulfinator [23] was developed to predict sulfotyrosine sites based on four different hidden Markov models using information from sequence alignment. Despite the high predictive accuracy of Sulfinator's validation test set, it cannot identify certain kinds of sulfated tyrosine sites, such as sulfotyrosine in extracellular class II leucine-rich repeat (LRR) proteins that were identified by mass spectrometry experiment [7,22]. In 2009, Chang and coworkers developed a method called SulfoSite [18] that selected accessible surface area (ASA) and positional weighted matrix (PWM) for predicting tyrosine sulfation sites based on support vector machine. Furthermore, Niu and coworkers incorporated features of sequence conservation, residual disorder, and amino acid factor to predict protein tyrosine sulfation with maximum relevance minimum redundancy method [24]. Although the two methods achieved great progress in predicting sulfotyrosine sites based on several features, there were still some limitations. For example, in the SulfoSite method, Chang and coworkers did not consider physicochemical properties of amino acids and residue sequence order information and had not investigated the effect of features on the occurrence of tyrosine sulfation. Niu and coworkers achieved a total accuracy of 90.01%, but the positive accuracy was just 66.67%. Therefore, it has become a crucial issue to understand the molecular basis of tyrosine sulfation and enhance the quality of predicting protein sulfotyrosine sites by selecting more informative features.

In this article, a new computational method called PredSulSite to predict sulfotyrosine sites from protein primary sequences is presented. The characteristics of protein secondary structure, physicochemical properties of the amino acids, and residue sequence order information were incorporated to extract sequence features. Moreover, we analyzed the individual features and investigated the influence of different features. An evaluation of the trained models based on the support vector machine method in 10-fold cross-validation revealed that the Matthews correlation coefficient, sensitivity, specificity, and accuracy were 84.28%, 92.00%, 93.33%, and 92.89%, respectively, indicating that the proposed method could effectively identify protein tyrosine sulfation.

Materials and methods

According to a recent comprehensive review [25], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein or peptide samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be identified; (iii) introduce or develop a powerful algorithm to operate the prediction and properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (iv) establish a user-friendly web server for the predictor that is accessible to the public. As shown in Fig. 1, the proposed approach, PredSulSite, comprised four main analytical processes: data preprocessing, feature encoding, model learning and evaluation, and sulfotyrosine prediction. Below, we describe how to deal with these steps.

Collecting and preprocessing data

The experimentally validated tyrosine sulfated sites were retrieved from UniProtKB (release 2011_09) [26] with the keyword "sulfotyrosine", which contains 127 proteins covering 184 sulfated tyrosine sites. The sequences with less than 50 amino acids were included because they might be some important peptides. It is well known that if the datasets are highly homologous, the prediction performance would be overestimated. To remove the homologous

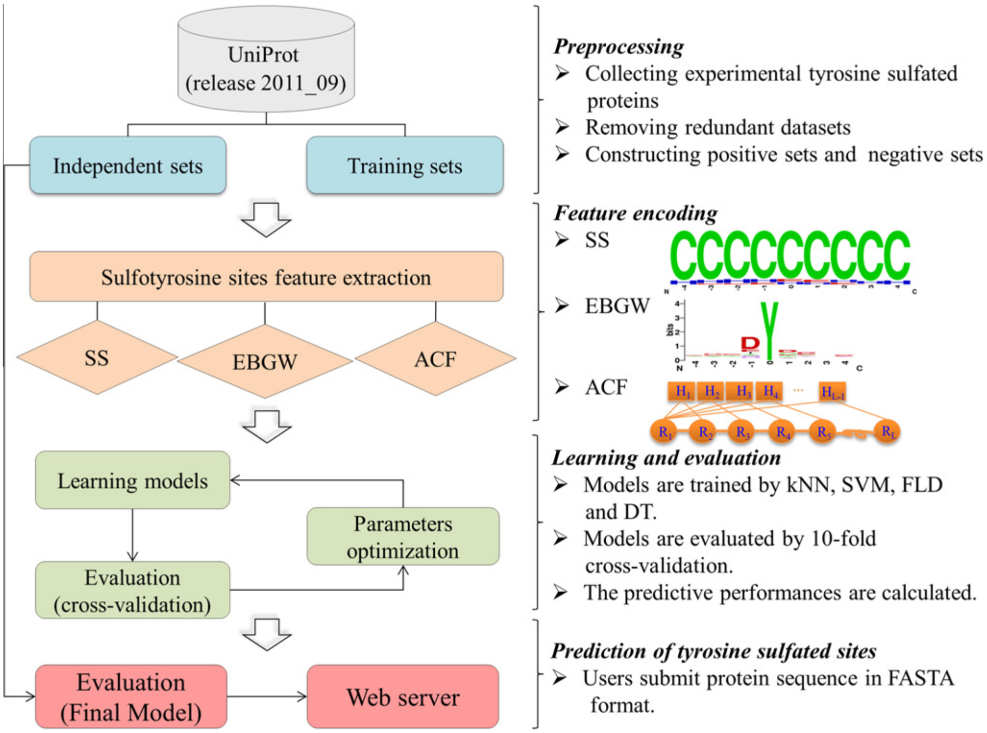

UniProt
(release 2011_09)
Independent sets
Training sets
Sulfotyrosine sites feature extraction
SS
EBGW
ACF
Learning models
Parameters optimization
Evaluation (cross-validation)
Evaluation (Final Model)
Web server
Preprocessing
➢ Collecting experimental tyrosine sulfated proteins
➢ Removing redundant datasets
➢ Constructing positive sets and negative sets
Feature encoding
➢ SS
CCCCCCCCC
➢ EBGW
➢ ACF
H_1 H_2 H_3 H_4 ··· H_{L-1}
R_1 R_2 R_3 R_4 R_5 R_L
Learning and evaluation
➢ Models are trained by kNN, SVM, FLD and DT.
➢ Models are evaluated by 10-fold cross-validation.
➢ The predictive performances are calculated.
Prediction of tyrosine sulfated sites
➢ Users submit protein sequence in FASTA format.

Fig.1. System flowchart of PredSulSite.