Incorporating Support Vector Machine for Identifying Protein Tyrosine Sulfation Sites

WEN-CHI CHANG,[1,2*] TZONG-YI LEE,[2*] DRAY-MING SHIEN,[3,5] JUSTIN BO-KAI HSU,[2] JORNG-TZONG HORNG,[3,6] PO-CHIANG HSU,[2] TING-YUAN WANG,[2] HSIEN-DA HUANG,[1,2] RONG-LONG PAN[4]

[1]*Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu, Taiwan*

[2]*Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-Chu, Taiwan*

[3]*Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan*

[4]*Institute of Bioinformatics and Structural Biology, College of Life Sciences, National Tsing Hua University, Hsin-Chu, Taiwan*

[5]*Department of Electronic Engineering, Chin Min Institute of Technology, Miao-Li, Taiwan*

[6]*Department of Bioinformatics, Asia University, Taichung, Taiwan*

Received 31 October 2008; Revised 21 January 2009; Accepted 2 February 2009
DOI 10.1002/jcc.21258
Published online 16 April 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Tyrosine sulfation is a post-translational modification of many secreted and membrane-bound proteins. It governs protein-protein interactions that are involved in leukocyte adhesion, hemostasis, and chemokine signaling. However, the intrinsic feature of sulfated protein remains elusive and remains to be delineated. This investigation presents SulfoSite, which is a computational method based on a support vector machine (SVM) for predicting protein sulfotyrosine sites. The approach was developed to consider structural information such as concerning the secondary structure and solvent accessibility of amino acids that surround the sulfotyrosine sites. One hundred sixty-two experimentally verified tyrosine sulfation sites were identified using UniProtKB/SwissProt release 53.0. The results of a five-fold cross-validation evaluation suggest that the accessibility of the solvent around the sulfotyrosine sites contributes substantially to predictive accuracy. The SVM classifier can achieve an accuracy of 94.2% in five-fold cross validation when sequence positional weighted matrix (PWM) is coupled with values of the accessible surface area (ASA). The proposed method significantly outperforms previous methods for accurately predicting the location of tyrosine sulfation sites.

© 2009 Wiley Periodicals, Inc. J Comput Chem 30: 2526–2537, 2009

Key words: protein; sulfation; prediction

Introduction

Numerous post-translational modifications (PTMs) of proteins provide the proteome with structural and functional diversity, and regulate cellular plasticity and dynamics. Tyrosine sulfation is one of the most common PTMs in secreted and transmembrane proteins, and has been experimentally demonstrated to be essential to extracellular protein–protein interactions.[1,2] Approximately 1% of all tyrosine residues of the total proteins in an organism can be sulfated.[3] Several proteins that are known to be tyrosine-sulfated play important roles in physiological processes, and a direct link between protein function and tyrosine sulfation has been established in many cases.[4] For example, sulfated tyrosines have an essential role in the immune response,[5] such as promote HIV infection of T-helper lymphocytes,[1,6] leukocyte rollin,[7] and a complementary cascade.[8] Sulfation is also involved in the modulation of intracellular protein transportation, and the regulation of the proteolytic process of proteins.[9] Tyrosine sulfation occurs when tyrosylprotein sulfotransferases (TPSTs) catalyze the transfer of a negatively charged sulfate from 3′-phosphoadenosine-5′-phosphosulfate (PAPS) to the hydroxyl group of tyrosine residue on a polypeptide.[10] Two tyrosylprotein sulfo-

Additional Supporting Information may be found in the online version of this article.

*Both the authors contributed equally to this work.

Correspondence to: H.-D. Huang; e-mail: bryan@mail.nctu.edu.tw or R.-L. Pan; e-mail: rlpan@life.nthu.edu.tw

Contract/grant sponsor: National Science Council of the Republic of China; contract/grant number: NSC 97-2811-B-009-001

Contract/grant sponsor: National Research Program for Genomic Medicine (NRPGM), Taiwan

© 2009 Wiley Periodicals, Inc.

transferases (TPST1 and TPST2) have been identified.[10–13] The use of synthetic peptide acceptors *in vitro* suggests that the two TPST proteins prefer different substrates.[12] For instance, peptides modeled on the N-terminus of P-selectin glycoprotein ligand-1 are sulfated by the two isoenzymes with equal efficiency.[14] Furthermore, peptides that are modeled on sulfation sites of human C4-α-chain and heparin cofactor II are sulfated more efficiently by TPST1 than by TPST2.[12] In contrast, Danan et al. demonstrated that TPST2 is more catalytically efficient than TPST1 in sulfating CCR8 substrates.[15] The two tyrosylprotein sulfotransferases, TPST-1 and TPST-2, have overlapping but not identical substrate specificities *in vitro* and *in vivo*.[2,12,15] However, the relative abundances and distribution of the two isoenzymes have not yet been examined because of lack of suitable analytical reagents and tools.[10,12]

Tyrosine-sulfated protein is important to cellular control, but no clear-cut acceptor sequence of TPSTs, which can be used to predict sulfotyrosine sites. Because of labile sulfotyrosine, the characterization of this PTM has been impeded by the limitation of general, unambiguous methods for its site determination. Eliminating numerous false positive tyrosine sulfations, Rosenquist and Nicholas discovered a test for acidic amino acids close to the target tyrosine, which yielded a good filtering criterion.[16,17] However, no universal features have yet been identified for use in the accurate prediction of sulfation sites. To increase the predictive accuracy and to reduce the number of false-positives, Yu et al. incorporated context-based rules and logical filters to develop a so-called position-specific scoring matrix (PSSM) to predict tyrosine sulfation sites.[18] Unfortunately, the capacity to find new or unseen PTM patterns is absent due to the static behavior of PSSMs.[12,18] Additionally, the methods demonstrated above without a user-friendly interface. Sulfinator[19] was developed to predict sulfotyrosine sites based on four Hidden Markov Models. Despite the high predictive accuracy of Sulfinator's validation test set, it cannot be guaranteed to identify new sulfated tyrosine sites from all of the protein sequence databases.[12] For instance, variable sulfotyrosine in extracellular class II leucine-rich repeat (LRR) proteins were characterized by mass spectrometry, and, Sulfinator[19] does not accurately predict tyrosine sulfation in this class of proteins.[3] Therefore, a computational tool to predict accurately tyrosine sulfation sites in protein sequences is crucial.

This work develops a new method for identifying protein sulfotyrosine residues. The studies cited earlier have been only proposed to analyze the consensus sequences around sulfotyrosine sites and did not take consider other information. Most PTM sites are located at the surface of the protein, and the residues with larger Accessible Surface Areas (ASA) are regarded as surface residues.[20] In this investigation, position-specific amino acid frequencies and the structural property of the protein, in terms of the Accessible Surface Area (ASA), are employed to distinguish between sulfation and non-sulfation sites using a Support Vector Machine (SVM).[21] All tyrosine-sulfated proteins were collected from UniProtKB/SwissProt (Release 53.0).[22] Subsequently, the experimental tyrosine-sulfated proteins were selected to reduce the quantity of the homologous sequences with a given window size and processed for SVM training. Suitable parameters were used to train SVM models. Five-fold cross-validation was adopted to evaluate the accuracy of the models. The estimates made using the models reveal that the solvent accessibility around the sulfation sites can be used to improve their predictive accuracy. Furthermore, during the independent test experiment, the proposed method achieves 84% predictive accuracy. The phosphorylation and sulfation of tyrosine are well known to be isobaric. Hence, differentiating modifications by phosphorylation from those by sulfation is generally difficult. However, the method herein effectively distinguishes sulfotyrosine from phosphotyrosine sites. SulfoSite is now available as a freely accessible web server at http://SulfoSite.mbc.nctu.edu.tw.

Materials and Methods

Various machine learning approaches (such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and others) support the rapid collection, annotation, retrieval, comparison, and mining of specific features from large biological datasets, to solve structure–function problems.[23–25] A similar research flow was also utilized in this study. Figure 1 presents the flowchart of the proposed method. The method comprises three major steps: (i) collecting and preprocessing data, (ii) extraction of features, and (iii) creation and evaluation of models.

Collecting and Preprocessing Data

Collecting and Preprocessing Data to Construct and Test Models

UniProtKB/SwissProt[22] (release 53.0) maintains 269,293 protein entries, of which 313 contain 732 residues, that are annotated as "sulfotyrosine" in the "MOD_RES" fields (and exhibit post-translational modification of a residue). However, only experimentally verified sulfotyrosine sites were collected. Potential sulfotyrosine sites for which the comments in UniProtKB/SwissProt's contained the keywords of "by similarity," "potential," or "probable" were removed. As given in Table 1, 162 experimentally confirmed sulfotyrosine sites within 115 proteins that were extracted from UniProtKB/SwissProt release 53.0 were used for further preprocessing. Thereafter, the sulfotyrosine sites were extracted as positive sets, and the non-sulfotyrosine sites as negative sets from 115 tyrosine sulfated proteins.

Then, the 9-mer sequences (−4 to +4), 11-mer sequences (−5 to +5), 13-mer sequences (−6 to +6), 15-mer sequences (−7 to +7), and 17-mer sequences (−8 to +8) of sulfotyrosine sites are extracted and constructed as training sets. The positive (+) dataset for training may contain some homologous sites of homologous proteins. When the test data were highly homologous with the training data, the prediction accuracy was overestimated. To avoid such overestimation, when two sulfotyrosine sites of the two proteins were in the corresponding positions in the alignment, only one was kept. Thus, non-homologous positive (+) data on high-quality tyrosine-sulfated sites were obtained using different window sizes, as given in Table S1 (see Supporting Information). Nonhomologous positive and negative dataset were merged into a nonhomologous dataset. Moreover, the tyrosine-sulfated proteins in UniProtKB/SwissProt release 55.0, which were not available in release 53.0 were selected to

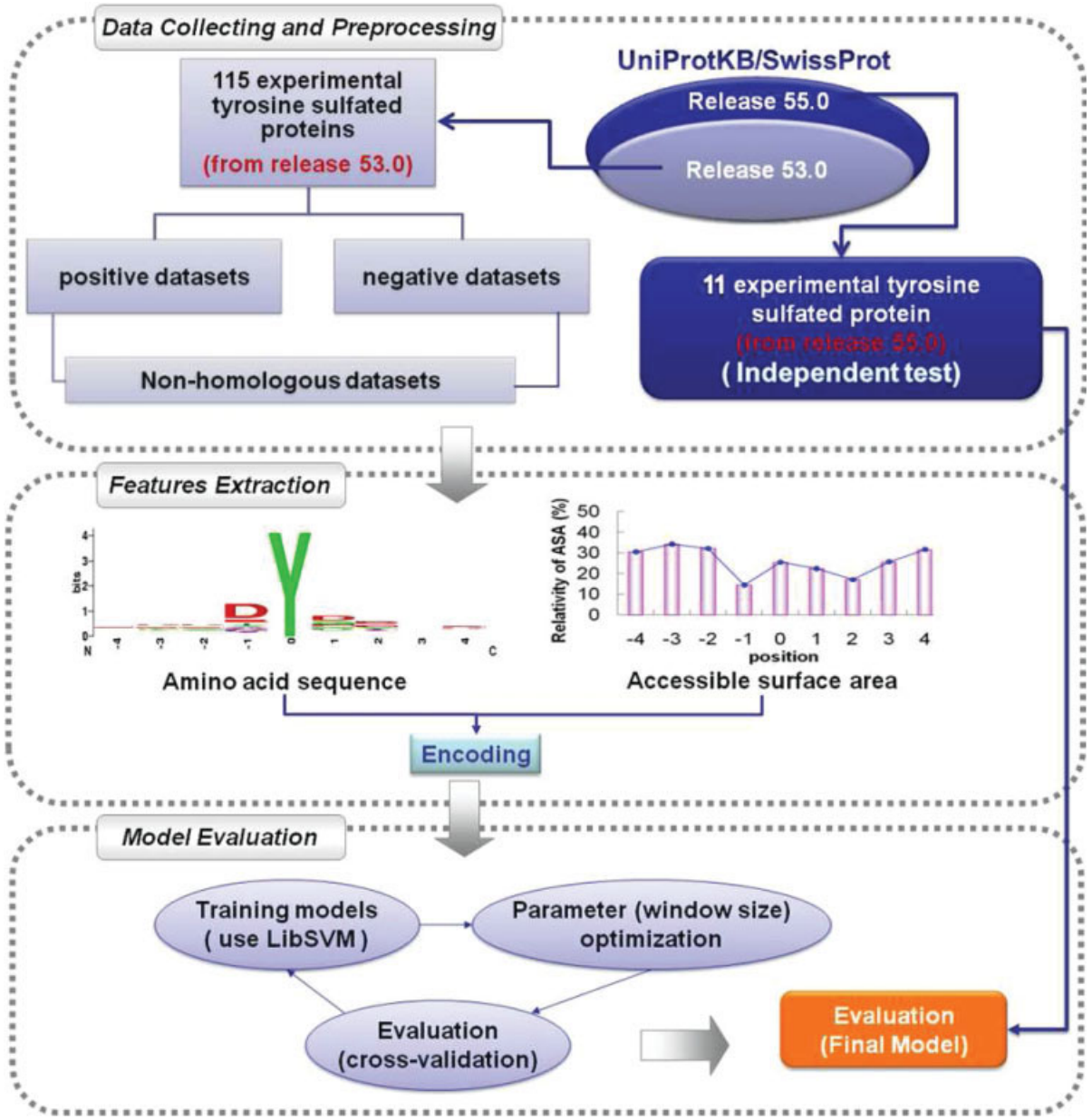


Figure 1. System flow of SulfoSite.

perform an independent test. The statistics of sulfation sites obtained from UniProtKB/SwissProt release 55.0 is given in Table S2 (see Supporting Information).

Collecting Data for Analysis of 3D Structure of Tyrosine-Sulfated Proteins

To elucidate the structural characteristics of sulfated protein, all matched 3D structures were downloaded from the protein dada bank (PDB),[26] as listed in Table S3 (see Supporting Information). Table 2 summarizes structural information on the non-redundant subset of sulfated proteins in PDB. In the case of tyrosine, only nine hit proteins (CXAA_CONPE, CXAB_CONPE, FIBG_HUMAN, GP1BA_HUMAN, HEP2_HUMAN, ITH1_HIRME, ITH3_HIRME, ITHA_HIRME, and TRY1_HUMAN) contained sulfotyrosine, and only 13 sulfated sites had reliable structures, as shown in Table 2. The majority of these proteins are related to the thrombin inhibitor, the others are conotoxins and trypsin, as presented in Table S4 (see Supporting Information). All of them indicated that the absence of sulfation in the critical tyrosine influences biological activity (Table S4, see Supporting Information). Some proteins (small peptides) have pharmacological applications; for example, conotoxins are utilized in the selection of a lead compound in drug design.

Table 1. The Statistics of Sulfation Sites Obtained from UniProtKB/SwissProt Release 53.0.

	Number of proteins	Number of sulfotyrosine sites
Tyrosine sulfated protein	313	732[a]
Experimental tyrosine sulfated protein	115	162[b]
Experimental secreted tyrosine sulfated protein	103	138[b]
Experimental membrane-bound tyrosine sulfated protein	8	17[b]
Experimental Tyrosine sulfated protein annotated in PDB	9	12
Nonhomologous experimental sulfation sites in SulfoSite	–	69[c]

[a]Those sites annotated as "by similarity", "potential," or "probable" are included.
[b]Those sites annotated as "by similarity", "potential," or "probable" are not included.
[c]On the basis of window size 4.

Extracting Features

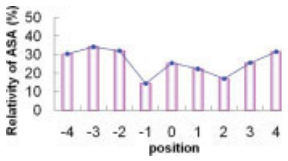
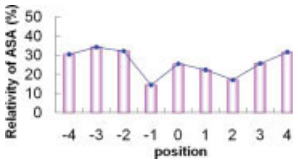
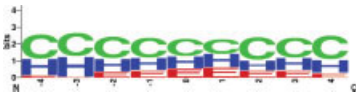
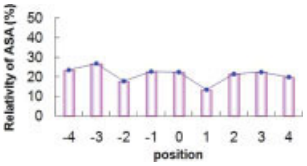
After the nonhomologous training set of tyrosine sulfation sites had been constructed, three characteristics—sequences that surround the sulfotyrosine sites, the secondary structure that surrounds the sulfotyrosine sites, and the solvent accessibility around the sulfotyrosine sites were extracted. However, the experimental ASA and SS information about the sulfotyrosine sites of known 3D structures in the protein data bank (PDB)[26] is not enough to be examined (Table 1). Consequently, RVP-Net[20,27] and PSIPRED[28] were applied to predict the Accessible Surface Area (ASA) and secondary structure (SS), respectively, for each residue of a protein sequence. RVP-Net directly predicts ASA without making exposure categories and achieves a mean absolute error better than 19% mean.[27] PSIPRED[28] is a highly accurate method for predicting the secondary structure, combining two feed-forward neural networks to analyze the output obtained using Position Specific Iterated-BLAST (PSI-BLAST).[29] PSIPRED 2.0 achieved an average accuracy of 80.6% across all 40 submitted target domains with no obvious sequence similarity to structures that are present in PDB, making PSIPRED better than all other evaluated methods.[30] The output of PSIPRED includes three symbols "H," "C," and "E" which stand for helix, coil, and β-sheet respectively. Before SVM models are created, the extracted features should be encoded as real values except the ASA values. The two sequences around the sulfotyrosine site (position 0) and the 9-mer sequences (−4 to +4), 11-mer sequences (−5 to +5), 13-mer sequences (−6 to +6), 15-mer sequences (−7 to +7), and 17-mer sequences (−8 to +8) of the tyrosine sulfation sites were encoded in three ways—two for the sequence and one for the secondary structure. As presented in Table S5 (see Supporting Information), an orthogonal binary coding scheme was adopted to transform amino acids into numeric vectors, as in so-called 20-dimensional vector coding. For example, alanine was encoded as "10000000000000000000";

Table 2. Hits between sulfated proteins from UniProtKB/SwissProt 53.0 and PDB.

			PDB coverage in SwissProt	The position of sulfated tyrosine in SwissProt protein	SS of sulfated tyrosine in crystal structure	ASA percentage of sulfated tyrosine (position = 0) in crystal structure											
SwissProt ID	PDB ID	Chain				−6	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	+6
CXAA_CONPE	1PEN	A	1–16	Y15	coil	0.26	0.74	0.80	0.39	0.73	0.88	0.86	0.50	–	–	–	–	–
CXAB_CONPE	1AKG	A	45–60	Y59	coil	0.28	0.90	0.73	0.41	0.62	0.65	0.83	0.54	–	–	–	–	–
FIBG_HUMAN	2HWL	P	439–452	Y444	coil	–	0.11	0.24	0.39	0.37	0.66	0.51	0.11	0.45	0.29	0.90	0.97	0.78
				Y448	coil	0.37	0.66	0.51	0.11	0.45	0.29	0.90	0.97	0.78	0.87	0.89	–	–
GP1BA_HUMAN	1GWB	B	16–296	Y294	coil	0.52	0.20	0.00	0.15	0.05	0.15	0.07	0.35	0.51	–	–	–	–
	1P8V	A	17–295	Y292	coil	0.97	0.50	0.31	0.65	0.58	0.30	0.17	0.12	0.62	0.75	–	–	–
				Y295	coil	0.65	0.58	0.30	0.17	0.12	0.62	0.75	–	–	–	–	–	–
HEP2_HUMAN	1JMO	A	20–499	Y79	coil	0.33	0.54	0.33	0.50	0.33	0.31	0.26	0.02	0.45	0.20	0.88	0.61	0.12
				Y92	helix	0.18	0.32	0.72	0.19	0.91	0.47	0.95	0.78	0.67	0.52	0.92	0.79	0.89
ITH1_HIRME	1BCU	I	1–65	Y63	helix	0.40	0.78	0.08	0.24	0.89	0.67	0.28	0.64	–	–	–	–	–
ITH3_HIRME	1W7G	I	60–71	Y70	coil	0.38	0.68	0.18	0.42	–	0.89	0.38	–	–	–	–	–	–
ITHA_HIRME	1A46	I	54–64	Y63	coil	0.46	0.25	0.11	0.38	0.29	0.37	0.41	0.60	–	–	–	–	–
TRY1_HUMAN	1TRN	A	24–247	Y154	coil	0.03	0.25	0.54	0.67	0.77	0.45	0.53	0.13	0.47	0.37	0.03	0.11	0.01

More details are shown in Table S3 (see Supporting Information). If SwissProt_ID hits more than one PDB_ID, only one is shown in this table.

Table 3. The Statistics of Sequence Logo, Average Accessible Surface Srea (ASA), and Secondary Structure (SS) Surrounding the Sulfotyrosine Sites (Position = 0) in UniProtKB/SwissProt Release 53.0.

	Sequence logos	Accessible surface area (ASA)	Secondary structure
Entire experimental tyrosine sulfated sites			
Nonhomologous tyrosine sulfated sites			
Nonsulfated tyrosine sites			

cysteine encoded as "01000000000000000000," and so on. The number of feature vectors that correspond to the flanking amino acids that surround the sulfated site was $(2n + 1) \times 20$. Values of n from four to eight were used to determine the optimal window length. The 20-dimensional vector for protein sequence encoding (with each amino acid mapped to a 20-dimensional vector) was named SEQ(20-d). Additionally, the positional weighted matrix (PWM) of amino acids around the sulfated sites was determined for tyrosine using nonhomologous training data. The positional weighted matrix (PWM) specified the relative frequency of amino acids in the sulfated sites, and was used to encode the fragment sequences, namely SEQ(PWM) (Table S5 in Supporting Information). Whenever feature values were missing (such as in case of tyrosine residues near the end of the peptide), they were discarded. Furthermore, the full-length protein sequences with sulfated sites were inputted to PSIPRED to verify the secondary structure of all residues. The orthogonal binary coding scheme was used to transform the three terms that specify the secondary structure into numeric vectors. For instance, helix was encoded as "001," sheet was encoded as "010," and coil was encoded as "100." The 3D vector that encoded the secondary structure (SS) of the protein (each secondary structure was mapped to a 3D vector) was named SS. Figure S1 (see Supporting Information) presents a diagram of the feature dimensions.

Creation and Evaluation of Model

SVM is a machine learning method, which has been utilized to solve pattern identification problems with clear connections to the underlying statistical learning theory.[21] The principle of SVM is to map input vectors to a higher dimensional space where a maximal separating hyperplane is defined. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data into two groups. The separating hyperplane is that which maximizes the distance between the two parallel hyperplanes. Furthermore, SVM can solve the classification problem when the number of training data is very small.[31] LDA is a simpler classifier than SVM that is adopted to solve biological structure–function problems.[32–34] People may argue with our reasons for selecting SVM as the method herein. Therefore, the classification performance of LDA and SVM were compared initially. As given in Table S6 (see Supporting Information), the models trained by SVM greatly outperform those trained using LDA based on the same training and validation data sets. Accordingly, this work employed protein sequences, profiles of Accessible Surface Area (ASA) and secondary structure (SS) in the support vector machine (SVM) for training the models of sulfotyrosine site prediction. An SVM library called LIBSVM,[35] was applied to train the predictive models and a radial basis function (RBF) was selected as the SVM kernel function in this study. The experimental sulfotyrosine sites were defined as the positive dataset, while all other nonsulfated tyrosine sites in the sulfated proteins were treated as the negative dataset. K-fold cross-validation was exploited to evaluate the predictive performance of the models. The sizes of the positive and negative datasets were made equal during the cross-validation processes. After the models were trained, whether they were the models are appropriate had to be determined. The performance of SVM was measured in a five-fold cross-validation analysis, in which each dataset was divided into five parts: four parts were used for model learning (training); the remaining one was used for validation (testing). Four performance measures were employed; they were precision (Pr), sensitivity (Sn), specificity (Sp), and

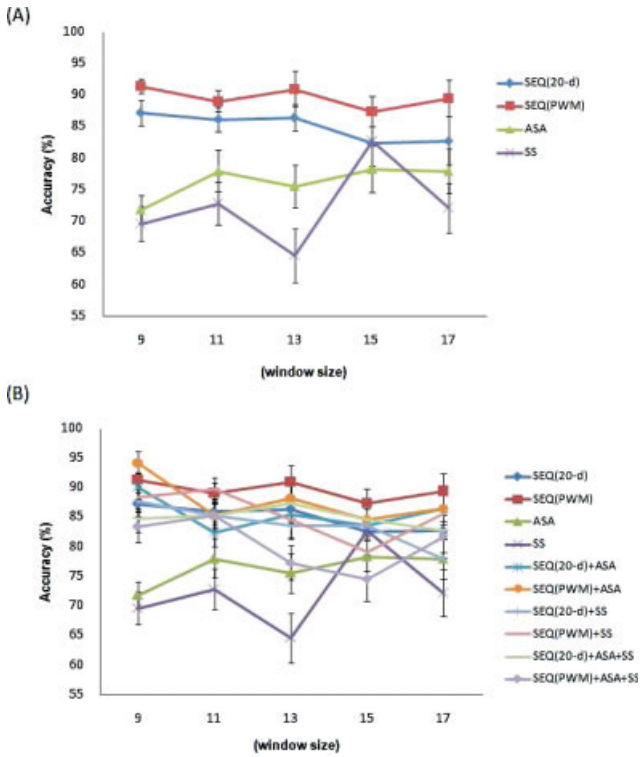


Figure 2. Performance of the models trained with various features and different window sizes in five-fold cross-validation. (A) The predictive accuracies of the models trained with SEQ(20-d), SEQ(PWM), ASA, and SS based on various window sizes. (B) The predictive accuracies of the models trained with SEQ(20-d), SEQ(PWM), ASA, SS, SEQ(20-d)+ASA, SEQ(20-d)+SS, SEQ (20-d)+ASA+SS, SEQ(PWM)+ASA, SEQ(PWM)+SS, and SEQ (PWM)+ASA+SS based on various window sizes. Abbreviation: SEQ(20-d), protein sequence encode by 20-dimentional method; ASA, Accessible Surface Area; SS, secondary structure; SEQ(PWM), protein sequence encode by positional weighted matrix (PWM) method. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

accuracy (Acc), defined later. The Matthews correlation coefficient (MCC) was used to measure the classification performance in the SVM training process to accommodate unbalanced datasets. TP, TN, FP, and FN are true positive, true negative, false positive, and false negative predictions, respectively.

$$\text{Precision (Pr)} = \frac{\text{TP}}{(\text{TP}+\text{FP})}; \quad \text{Sensitivity (Sn)} = \frac{\text{TP}}{(\text{TP}+\text{FN})};$$

$$\text{Specificity (Sp)} = \frac{\text{TN}}{(\text{TN}+\text{FP})};$$

$$\text{Accuracy (Acc)} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})};$$

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

Several parameters of the models, the length of the sequence that surround the sulfotyrosine site, the SVM cost and the gamma values were optimized to maximize predictive accuracy. Additionally, receiver operating characteristic (ROC) curves for different window sizes and features were plotted using LIBSVM.[35] Finally, the parameters of the trained model with the highest predictive accuracy from each dataset were adopted to provide the prediction service on the web. SulfoSite was implemented in PHP, PERL, and hosted by an Apache web server running on a RedHat Linux system.

Results and Discussion

Characteristics of Sulfated Protein

First, all sulfated proteins from UniProtKB/SwissProt[22] (release 53.0) were collected; 115 of them were experimentally annotated. As indicated in Table 1, the numbers of secreted and membrane-bound tyrosine sulfated proteins were 103 and eight, which contained 138 and 17 sulfotyrosine sites, respectively. Table 3 summarizes the sequence logos,[36] average ASA, and secondary structure formed from the 9-mer sulfated and the 9-mer non-sulfated sites in the constructed data set. As the flanking sequences (position -4 to $+4$) of the sulfotyrosine sites (position 0) are graphically visualized as sequence logos[36] of the protein sequence and the secondary structure (SS), examining the sequence logos reveals that the sequences that surround the sulfated sites are more acidic and more strongly conserved than those around non-sulfated sites. Furthermore, Accessible Surface Area (ASA) is higher at the sulfotyrosine site than at the flank sites.

On the basis of the 3D structures from the protein data bank (PDB),[26] Table 2 demonstrates that two sulfotyrosine residues in PDB proteins were observed in helical regions, while the others were observed in the loop. Furthermore, the ASA percentages of some sulfotyrosine residues exceeded those around the sulfated site, indicating that the sulfotyrosine is usually on the protein surface. Additionally, the locations of sulfotyrosines in PDB proteins classified them into three broad types (main chain of protein complex, small peptide, or inhibitor), as displayed in Figure S2 (see Supporting Information). The figure also indicates the sulfotyrosine residues located on the structure surface. The 3D structure is consistent with the sequence analysis that is given in Table 3, potentially explaining why the ASA feature is critical to predicting the sites of sulfotyrosine.

Algorithm

Predictive Performance of Cross Validation–Individual Feature

Although SVM is effectively ignores irrelevant features in high dimensional feature sets, proper attribute selection may further increase the accuracy of the SVM algorithm. In previous studies, methods for predicting sulfotyrosine sites have focused on the conservation of the protein sequence: an optimal discriminative set of conserved protein sequences and structural properties had not been identified.[19] The predictive accuracy was used to evaluate four individual properties (SEQ(20-d), SEQ(PWM), ASA,

Table 4. Prediction Accuracy of the Models Trained with Various Parameters and Different Window Size in Five-Fold Cross-Validation.

	Window size					
Training features	9	11	13	15	17	Average accuracy
SEQ(20-d)	87.13	86.03	86.36	82.45	82.69	84.93
SEQ(PWM)	91.30	88.97	90.91	87.27	89.42	89.58
ASA	71.74	77.94	75.45	78.18	77.88	76.24
SS	69.57	72.79	64.55	82.73	72.12	72.35
SEQ(20-d)+ASA	90.10	82.35	85.45	83.64	86.54	85.62
SEQ(PWM)+ASA	94.20[b]	85.29	88.18	84.55	86.54	87.75
SEQ(20-d)+SS	87.68	85.29	83.64	83.64	77.88	83.63
SEQ(PWM)+SS	88.41	89.71	84.55	79.09	85.58	85.47
SEQ(20-d)+ASA+SS	84.78	85.29	87.27	84.55	82.69	84.92
SEQ(PWM)+ASA+SS	83.33	85.29	77.27	74.55	81.73	80.44
SEQ(20-d)+ASA_5X[a]	87.66	87.82	77.69	86.15	88.52	85.57
SEQ(20-d)+ASA_10X[a]	88.31	87.82	88.46	85.38	80.33	86.06
SEQ(20-d)+ASA_20X[a]	92.21	85.9	82.31	86.15	90.16	87.35
Average accuracy	85.88	84.65	82.47	82.95	83.24	

The ROC curves are given in Figure S4 (see Supporting Inforamation).
[a]ASA_NX: The weight of ASA increase to *N* folds.
[b]The best accuracy.

and SS) of the residues to determine the optimal subset of such properties. The average predictive accuracy of individual attributes, except SEQ(PWM), was low. As shown in Figure 2A and Table 4, that of SEQ(PWM) was 89.58%, that of SEQ(20-d) was 84.93%, that of ASA was 76.24%, and that of SS was only 72.35%. A model trained with a window size 9 (−4 to +4) based on SEQ(PWM) performed best in predicting sulfotyrosine (91.30%).

Predictive Performance of Cross Validation–Combinations of Feature Sets

The predictive performance of models trained with the combined attributes was also estimated. The average prediction accuracy increased to 85.62% when the SEQ(20-d) was coupled with ASA, as given in Table 4. However, the predictive accuracy of SEQ(20-d)+ ASA was not satisfactory. Since the number of dimensions in SEQ(20-d) was 20 times that in ASA, the weight of SEQ(20-d) exceeded that of ASA in the prediction of the sites of sulfated tyrosine. Thus, the predictive performance was dominated by the sequence feature. The weight of ASA was increased five-fold, 10–fold, and 20-fold, and the average accuracy did not significantly change for SEQ(20-d)+ASA_5X (five-fold), but improved to 86.06%, and 87.35% for SEQ(20-d)+ASA_10X (10-fold), and SEQ(20-d)+ASA_20X (20-fold), suggesting that ASA notably improved the predictive accuracy to 92.21% when its weight increased to 20-fold in window size 9 (−4 to +4) (Table 4 and Fig. S3, see Supporting Information). For this reason, the weight of ASA is important in predicting the sites of sulfotyrosine. However, the number of dimensions in SEQ(PWM) equaled that of to ASA, balancing the weights to enable sulfated and non-sulfated tyrosine to be classified. Although the average accuracy of SEQ(PWM)+ASA was not better than that of SEQ(PWM), a model trained with SEQ(PWM)+ASA with window size 9 (−4 to +4) was the most accurate (94.20%), as given in Figure 2B and Table 4. The AUC (area under curve) of the ROC curve was 0.9702 in the model that had been trained with SEQ(PWM)+ASA in window size 9 (−4 to +4), which is the best value among all of the models, as given in Figure S4 (see Supporting Information). Additionally, the secondary structure (SS) feature did not markedly improve the predictive accuracy. The best predictive model

Table 5. The Parameters and Predictive Performance of the Trained Models with Different Features Which Achieve the Highest Accuracy.

Sulfated residue	Number of positive training set	Training features	Window size	Parameters	Pr	Sn	Sp	Acc	MCC
Tyrosine	69	SEQ(PWM) +ASA	−4 to +4	C^a = 32768, G^b = 0.00048828125	95.52	92.75	95.65	94.20	0.88

[a]Cost value.
[b]Gamma value

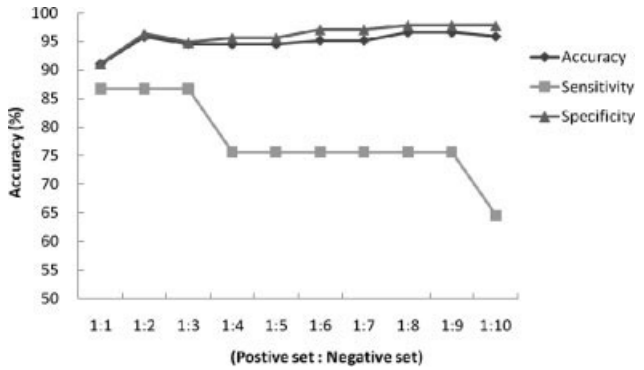


Figure 3. The performance of balanced and unbalanced dataset based on SEQ(PWM)+ASA in window size 9 (−4 to +4).

was trained using a window size 9 (−4 to +4) in sulfotyrosine prediction. Table 5 presents the number of training data, training features, window size, SVM cost and gamma values, precision, sensitivity, specificity, accuracy, and MCC of the best trained model.

Predictive Performance of Balanced and Unbalanced Dataset

In cross-validation, balanced positive and negative datasets were utilized to train models. However, for this purpose, sampling a negative dataset absolutely randomly is difficult because the number of the negative sets is much more than those positive sets. To avoid the tendentiousness in negative set extraction and reduce the false positive rate, the negative set for five-fold cross validation was collected randomly. Its size was separately one, two, three, four, five, six, seven, eight, nine, and 10 times that of the positive set. These unbalanced datasets are for training only, and then testing is conducted with a dataset whose positive/negative ratio is an estimate of the ratio in an unmodified dataset (has a positive/negative ratio of about 1/40). The results suggest that higher accuracy and specificity were related to more negative sets, as given in Figure 3 and Table 6. However, the sensitivity was approximately inversely proportional to the number of negative dataset (see Fig. 3). The accuracy and specificity almost reached their respective maximum values (Ac = 96.55% and Sp = 97.78%), when the ratio of the number of the positive set to that of the negative set was to 1:8. At this ration, the sensitivity had dramatically decreased to 75.56% (see Fig. 3). A trade-off had to be made between specificity and sensitivity. Although the accuracy at the ratio 1:2 was lower than that at 1:8, the sensitively remained at 86.67% and also high accuracy and specificity were maintained. Furthermore, the MCC value was the highest. Therefore, to gain the higher prediction sensitivity and the lower false positive rate, the model that had been trained with 1:2 unbalanced dataset based on SEQ(PWM)+ASA (window size = 9) was selected as the best model herein.

Table 6. The Accuracies of Balanced and Unbalanced Dataset.

Training feature: SEQ(PWM)+ASA				
Window size: 9 (−4 to +4)				
Positive dataset:Negative dataset	ACC	Se	Sp	MCC
1:1	91.03	86.67	91.12	0.72
1:2	95.86	86.67	96.32	0.87
1:3	94.48	86.67	94.85	0.82
1:4	94.48	75.56	95.59	0.77
1:5	94.48	75.56	95.59	0.77
1:6	95.17	75.56	97.06	0.74
1:7	95.17	75.56	97.06	0.74
1:8	96.55	75.56	97.78	0.86
1:9	96.55	75.56	97.78	0.86
1:10	95.87	64.44	97.78	0.78

Table 7. Independent test in Sulfinator[1] and SulfoSite.

SwissProt_ID	Real Sulfotyrosine sites	Sulfotyrosine predicted in Sulfinator	Sulfotyrosine predicted in SulfoSite
CCKN_CANFA	Y52	**Y52**	**Y52**
FMOD_BOVIN	Y20, Y38, Y53, Y55, Y63, Y65	**Y38**, Y42, Y45, Y47, Y62, Y64	**Y38, Y55**
FMOD_HUMAN	Y20, Y53, Y55	Y38, Y39, Y42, Y45, Y47	**Y20**, Y39
HIS1_HUMAN	Y46, Y49, Y53, Y55	**Y46, Y49**	**Y46, Y49**
LUM_MOUSE	Y20, Y21, Y23, Y30	–	**Y23**
OMD_HUMAN	Y39, Y416, Y417	–	Y31, **Y39**, Y51
PSK1_ORYSI	Y80, Y82	**Y80**	**Y80, Y82**
PSK1_ORYSJ	Y80, Y82	**Y80**	**Y80, Y82**
PSK2_ORYSI	Y110, Y112	–	**Y110, Y112**
PSK2_ORYSJ	Y110, Y112	–	**Y110, Y112**
SIAL_HUMAN	Y313, Y314	Y259, Y263, Y265, Y271, Y275, Y278, Y290, Y293, Y297, Y299	Y265, Y271, Y305
Pr	–	23%	73%
Sn	–	19%	52%
Sp	–	80%	94%
Acc	–	65%	84%

The words in bold: the sulfotyrosine sites are recovered by the indicating methods.

Table 8. Comparison Between Sulfinator[1] and SulfoSite.

	Sulfinator	SulfoSite
Number of sulfated protein for training	125	115[a]
Separating experimental and potential sulfotyrosine sites in training dataset	No	Yes
Nonhomologous dataset	No	Yes[b]
Features used for prediction	Protein sequence	Protein sequence, Accessible Surface Area
Method	Hidden Markov Model	Support Vector Machine
Cross-validation performance	Acc = 98%	Acc = 94.2 %
Independent test[c]	Acc = 65% (Sn = 19%, Sp = 80%, Pr = 23%)	Acc = 84% (Sn = 52%, Sp = 94%, Pr = 73%)

[a]Experimental tyrosine sulfated protein. Those sulfotyrosine annotated as "by similarity", "potential," or "probable" are not included.
[b]The number of nonhomologous dataset is shown in Table S1 in Supporting Information.
[c]31 positive sites and 98 negative sites were tested in Sulfiator and SulfoSite.

Testing

Predictive Performance of Independent Test by the Model Trained by SEQ(PWM)+ASA

Eleven sulfated proteins (CCKN_CANFA, FMOD_BOVIN, FMOD_HUMAN, HIS1_HUMAN, LUM_MOUSE, OMD_HUMAN, PSK1_ORYSI, PSK1_ORYSJ, PSK2_ORYSI, PSK2_ORYSJ, SIAL_HUMAN) in UniProtKB/SwissProt release 55.0, which were not available in UniProtKB/SwissProt release 53.0 were used for an independent test in SulfoSite, as given in Table 7. The independent test set involved 31 experimentally sulfated tyrosines and 98 non-sulfated tyrosines. Table 7 reveals that the system herein can predict the number of 16 sulfotyrosines (true positive), and only six predictions are false positive. The sensitivity and specificity of prediction are 52 and 94%, respectively. Generally, the performance in the independent test is just a little lower than those obtained in cross-validation, because that the over-fitting can usually not be completely prevented. Although the predictive accuracy in the independent test is not as high as that in cross-validation, the performance demonstrated in Tables 7 and 8 is acceptable. The precision and accuracy in the independent test are 73 and 84%, respectively, which are much better than test in Sulfinator[19] (precision = 23%, accuracy = 65%), suggesting that combination of sequence (encoded in PWM) and ASA is effective in sulfated protein prediction.

Predictive Performance of Model Trained Only With SEQ(PWM) in Independent Test

To confirm further the effectiveness of ASA in the prediction of sulfotyrosine sites, another model based on SEQ(PWM) with window size 9 (−4 to +4) was trained using un-balanced (positive: negative = 1:2) datasets. Significantly, our final model (trained by SEQ(PWM)+ASA) is also better in the forecasting of sulfotyrosine sites in an independent test than is the

Table 9. Comparison Between Final Model (SEQ(PWM)+ASA) and SEQ(PWM) Model with Independent Test.

SWISS-PROT_ID	Real Sulfotyrosine sites	The model trained by SEQ(PWM)	The model trained by SEQ(PWM)+ASA
CCKN_CANFA	Y52	**Y52**	**Y52**
FMOD_BOVIN	Y20, Y38, Y53, Y55, Y63, Y65	Y42	**Y38, Y55**
FMOD_HUMAN	Y20, Y53, Y55	Y39	**Y20**, Y39
HIS1_HUMAN	Y46, Y49, Y53, Y55	**Y49**	**Y46, Y49**
LUM_MOUSE	Y20, Y21, Y23, Y30	**Y23**	**Y23**
OMD_HUMAN	Y39, Y416, Y417	Y31, **Y39**, Y51	Y31, **Y39**, Y51
PSK1_ORYSI	Y80, Y82	**Y80, Y82**	**Y80, Y82**
PSK1_ORYSJ	Y80, Y82	**Y80, Y82**	**Y80, Y82**
PSK2_ORYSI	Y110, Y112	**Y110, Y112**	**Y110, Y112**
PSK2_ORYSJ	Y110, Y112	**Y110, Y112**	**Y110, Y112**
SIAL_HUMAN	Y313, Y314	Y265, Y271, Y305	Y265, Y271, Y305
Pr	–	63%	73%
Sn	–	39%	52%
Sp	–	93%	94%
Acc	–	80%	84%

The words in bold: the sulfotyrosine sites are recovered by the indicating models.

Figure 4. Web interface of SulfoSite. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

SEQ(PWM) model, as shown in Table 9, revealing that the accuracy can be improved by combining protein sequence (encoded using the PWM method) and ASA, as given in Figure 2 and Tables 4 and 9.

Several Cases Studies to Evaluate Performance of SulfoSite

Recently, tyrosine sulfation has been found to be prevalent in human chemokine receptors and important in human disease.[37] Yu et al., demonstrates Tyr10 and Tyr18 of C-C chemokine receptor type 1 (CCR1) can be sulfated by TPST.[2] The CCR1 protein sequence was inputted into our system to identify sulfotyrosine sites. Subsequently, sulfated Tyr10 and Tyr18 of CCR1 were specifically detected, as shown in Figure S5 (see Supporting Information). Furthermore, the PTM of proteins by phosphorylation is well-known to be the most abundant type of cellular regulation.[38] The tyrosine phosphorylation of proteins in the cytoplasm has critical roles in multiple biological processes.[9,39] It is because phosphotyrosine and sulfotyrosine are isobaric, distinguishing between phosphorylation and sulfation modifications on tyrosine residue is typically difficult, even though the sequence logos and ASA surrounding the sulfotyrosine and phosphotyrosine sites are different.[40] Table S7 (see Supporting Information) shows the amino acid sequences in the flanking of sulfotyrosine are more consensus than phosphotyrosine. It also demonstrates that sulfotyrosine has a higher ASA percentage.

Moreover, 1973 experimental phosphotyrosine sites from UniProtKB/SwissProt release 55.0 were used to evaluate our system. Largely phosphotyrosine sites cannot be miscalculated as sulfotyrosine sites in SulfoSite. The false positive rate is only 4.7%.

Implementation

As mentioned earlier, the best model was trained with sequence (PWM) and Accessible Surface Area in window size 9 (−4 to +4) was selected to utilize in the sulfotyrosine prediction sever. To avoid the biased extraction of the negative set, the un-balanced (positive: negative = 1:2) trained model was applied in SulfoSite web site. The SulfoSite web server provides user-friendly input and output interfaces, as shown in Figure 4. Several proteins in FASTA format could be inputted to the system. The predictions are presented in a diagramed that includes sulfated position, flanking amino acids, and the ASA probability, which were predicted by RVP-Net (see Fig. 4). The training dataset and independent test used in SulfoSite also can be downloaded for analysis.

Conclusion

Because of the levity of sulfotyrosine,[2] the characterization of protein sulfation modification has been hampered by the lack of specific, definite methods for its site determination. This investigation developed the SulfoSite web server, a high-performance sulfotyrosine predictor based on the SVMs method with protein sequence and Accessible Surface Area (ASA). The proposed system is comparable to several previous methods,[12,16,18,19] none of which consider structural information about the protein. Not only was the conservation of the protein sequence considered, but also the solvent accessibility and secondary structure around the sulfotyrosine sites were tested. The 3D structure data (information on ASA and secondary structure) for sulfation proteins were limited, so two effective tools, RVP-Net[20] and PSIPRED,[28] were used to compute the ASA percentage and secondary structure, respectively, based on protein sequence. Finally, the model trained by SEQ(PWM)+ASA was selected as the best predictive model in the SulfoSite web site. The method herein is 94.2% accurate in sulfotyrosine prediction in five-fold cross-validation. Tables 7 and 8 compare the performance of our approach with that of Sufinator.[19] The comparison demonstrates that SulfoSite has much better predictive performance than Sulfinator in an independent test. The use of more training data in SulfoSite may be responsible for its higher accuracy than Sulfinator. The nonhomologous dataset in SulfoSite may also mitigate the over-fitting problem in Sulfinator.

Several cases studies suggest that our method performs well not only in detecting sulfotyrosine sites but also in distinguishing sulfotyrosine from phosphotyrosine sites (Fig. S5 and Table S7 in Supporting Information). On the basis of the above, this developed system can facilitate the characterization of protein tyrosine sulfation.

Availability

The SulfoSite program is available at http://SulfoSite.mbc.nctu.edu.tw/.

References

1. Kehoe, J. W.; Bertozzi, C. R. Chem Biol 2000, 7, R57.
2. Yu, Y.; Hoffhines, A. J.; Moore, K. L.; Leary, J. A. Nat Methods 2007, 4, 583.
3. Onnerfjord, P.; Heathfield, T. F.; Heinegard, D. J Biol Chem 2004, 279, 26.
4. Seibert, C.; Sakmar, T. P. Biopolymers 2008, 90, 459.
5. Lin, H. C.; Tsai, K.; Chang, B. L.; Liu, J.; Young, M.; Hsu, W.; Louie, S.; Nicholas, H. B., Jr.; Rosenquist, G. L. Biochem Biophys Res Commun 2003, 312, 1154.
6. Choe, H.; Li, W.; Wright, P. L.; Vasilieva, N.; Venturi, M.; Huang, C. C.; Grundner, C.; Dorfman, T.; Zwick, M. B.; Wang, L.; Rosenberg, E. S.; Kwong, P. D.; Burton, D. R.; Robinson, J. E.; Sodroski, J. G.; Farzan, M. Cell 2003, 114, 161.
7. Bernimoulin, M. P.; Zeng, X. L.; Abbal, C.; Giraud, S.; Martinez, M.; Michielin, O.; Schapira, M.; Spertini, O. J Biol Chem 2003, 278, 37.
8. Gao, J.; Choe, H.; Bota, D.; Wright, P. L.; Gerard, C.; Gerard, N. P. J Biol Chem 2003, 278, 37902.
9. Zhang, Y.; Jiang, H.; Go, E. P.; Desaire, H. J Am Soc Mass Spectrom 2006, 17, 1282.
10. Moore, K. L. J Biol Chem 2003, 278, 24243.
11. Beisswanger, R.; Corbeil, D.; Vannier, C.; Thiele, C.; Dohrmann, U.; Kellner, R.; Ashman, K.; Niehrs, C.; Huttner, W. B. Proc Natl Acad Sci USA 1998, 95, 11134.
12. Monigatti, F.; Hekking, B.; Steen, H. Biochim Biophys Acta 2006, 1764, 1904.
13. Ouyang, Y. B.; Moore, K. L. J Biol Chem 1998, 273, 24770.
14. Wilkins, P. P.; Moore, K. L.; McEver, R. P.; Cummings, R. D. J Biol Chem 1995, 270, 22677.
15. Danan, L. M.; Yu, Z.; Hoffhines, A. J.; Moore, K. L.; Leary, J. A. J Am Soc Mass Spectrom 2008, 19, 1459.
16. Rosenquist, G. L.; Nicholas, H. B., Jr. Protein Sci 1993, 2, 215.
17. Bundgaard, J. R.; Vuust, J.; Rehfeld, J. F. J Biol Chem 1997, 272, 21700.
18. Yu, K. M.; Liu, J.; Moy, R.; Lin, H. C.; Nicholas, H. B., Jr.; Rosenquist, G. L. Endocrine 2002, 19, 333.
19. Monigatti, F.; Gasteiger, E.; Bairoch, A.; Jung, E. Bioinformatics 2002, 18, 769.
20. Ahmad, S.; Gromiha, M. M.; Sarai, A. Bioinformatics 2003, 19, 1849.
21. Vapnik, V. N. The Nature of Statistical Learning Theory; New York: Springer-Verlag, 1995.
22. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. Nucleic Acids Res 2003, 31, 365.
23. Gonzalez-Diaz, H.; Gonzalez-Diaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics 2008, 8, 750.
24. Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Curr Topics Med Chem 2007, 7, 15.
25. Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. FEBS Lett 2005, 579, 4297.
26. Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z.; Green, R. K.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E. Nucleic Acids Res 2005, 33, D233; Database issue.

27. Ahmad, S.; Gromiha, M. M.; Sarai, A. Proteins 2003, 50, 629.
28. McGuffin, L. J.; Bryson, K.; Jones, D. T. Bioinformatics 2000, 16, 404.
29. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic Acids Res 1997, 25, 3389.
30. Bryson, K.; McGuffin, L. J.; Marsden, R. L.; Ward, J. J.; Sodhi, J. S.; Jones, D. T. Nucleic Acids Res 2005, 33, W36; Web Server issue.
31. Burges, C. J. C. Data Mining and Knowledge Discovery 1998, 2, 121.
32. Gonzalez-Diaz, H.; Aguero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. J Comput Chem 2007, 28, 1049.
33. Gonzalez-Diaz, H.; Perez-Castillo, Y.; Podda, G.; Uriarte, E. J Comput Chem 2007, 28, 1990.
34. Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. J Comput Chem 2008, 29, 2613.
35. Chang, C. C.; Lin, C. J. In http://www.csie.ntu.edu.tw/~cjlin/libsvm 2001.
36. Schneider, T. D.; Stephens, R. M. Nucleic Acids Res 1990, 18, 6097.
37. Liu, J.; Louie, S.; Hsu, W.; Yu, K. M.; Nicholas, H. B., Jr.; Rosenquist, G. L. Am J Respir Cell Mol Biol 2008, 38, 738.
38. Huang, H. D.; Lee, T. Y.; Tzeng, S. W.; Wu, L. C.; Horng, J. T.; Tsou, A. P.; Huang, K. T. J Comput Chem 2005, 26, 1032.
39. Hunter, T. Philos Trans R Soc Lond B Biol Sci 1998, 353, 583.
40. Lee, T. Y.; Huang, H. D.; Hung, J. H.; Huang, H. Y.; Yang, Y. S.; Wang, T. H. Nucleic Acids Res 2006, 34 (Database issue), D622.