Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 203–208
www.elsevier.com/locate/compbiolchem

Predicting *O*-glycosylation sites in mammalian proteins by using SVMs

Sujun Li [a,b], Boshu Liu [a], Rong Zeng [b,c,*], Yudong Cai [d,**], Yixue Li [a,***]

[a] *Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China*
[b] *Research Center for Proteome Analysis, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China*
[c] *Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200436, China*
[d] *Department of Biomolecular Sciences, UMIST, P.O. Box 88, Manchester M60 1QD, UK*

Abstract

O-glycosylation is one of the most important, frequent and complex post-translational modifications. This modification can activate and affect protein functions. Here, we present three support vector machines models based on physical properties, 0/1 system, and the system combining the above two features. The prediction accuracies of the three models have reached 0.82, 0.85 and 0.85, respectively. The accuracies of the three SVMs methods were evaluated by 'leave-one-out' cross validation. This approach provides a useful tool to help identify the *O*-glycosylation sites in mammalian proteins. An online prediction web server is available at http://www.biosino.org/Oglyc.

Keywords: Post-translational modification; Bioinformatics; Prediction; *O*-glycosylation; Support vector machines

1. Introduction

As the annotation of genome sequences becomes more and more important, so does the need to functionally annotate these related proteins. Post-translational modification is one of the functional annotations of the proteins sequences. Of them, glycosylation is one of the most important, frequent and complex post-translational modifications. Glycosylation affect many protein critical functions including cellular communication, half-life and structure (Jenkins and James, 1996). Research into glycosylation can provide huge information for protein function, like recognition, adhesion, protein folding, metabolism, transport, etc.

There exists four types of glycosylation, including *N*-glycosylation, *O*-glycosylation, C-mannosylation and glycophosphatidlyinositol (GPI) anchor attachments (Blom, 2004). This paper will focus on the mucin-type *O*-glycosylation site prediction in mammalian proteins.

Although some high-throughput proteomics experimental methods have been developed to find post-translational modification (PTM) sites, it is still difficult to confirm glycosylation sites by these methods. Therefore, some related computational prediction approaches have been developed in recent years (Hansen et al., 1995, 1998; Eisenhaber and Eisenhaber, 1999; Julenius et al., 2005). For example the methodologies based on weight matrix (Eisenhaber and Eisenhaber, 1999) and artificial neural networks (Hansen et al., 1998; Julenius et al., 2005) help identify these kinds of PTM sites. The former method predicts the GPI-modification sites in proteins by using a weight matrix based on amino acid compositions. The latter method, NetOglyc, predicts the mucin-type *O*-glycosylation sites by using sequence contexts and surface accessibility. The NetOglyc system found 76% of the glycolsylated and 93% of the non-glycosylated serine and threonine sites correctly.

Here, we present three kinds of SVM models, respectively, based on physical properties, 0/1 system, and the system combining these two kinds of features, to predict the *O*-glycosylation sites in protein sequences.

The prediction models in this paper will provide users a helpful tool to identify the *O*-glycosylation sites in the protein sequences and to assist in proteome annotation. An online web server is available at http://www.biosino.org/Oglyc.

* Corresponding author. Tel.: +86 21 54920170; fax: +86 21 54920171.
** Corresponding author. Tel.: +44 161 200 4191; fax: +44 161 236 0409.
*** Corresponding author. Tel.: +86 21 54920089.
E-mail addresses: zr@sibs.ac.cn (R. Zeng), y.cai@umist.ac.uk (Y. Cai), yxli@sibs.ac.cn (Y. Li).

1476-9271/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2006.02.002

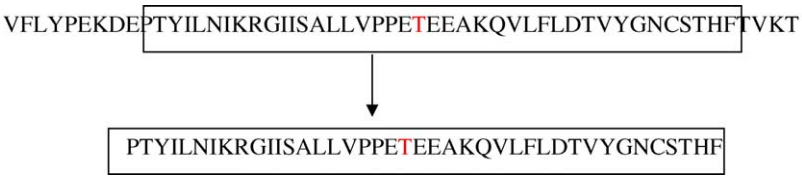


Fig. 1. The sequence was truncated to only include the *O*-glycosylation sites (threonine) region windows. The symmetrical window size is 20, meaning that 41 residues (from left 20 to right 20) in the sequence window.

2. Materials and methods

2.1. Dataset construction

2.1.1. Positive dataset

The glycoprotein sequences come from Swiss-Prot/UniProt6.1 (Bairoch, 1993). We took out most of the sequences with annotations of mucin-type *O*-glycosylation linked to serine or threonine in mammalian proteins sequences, excluding the sequences which have the annotation of "potential" or "probably". We ended up with 170 sequences totally.

These sequences were truncated to only include the verified *O*-glycosylation sites (serine/threonine) region windows. The symmetrical window size is 20, meaning that there are 41 residues with the *O*-glycosylation site (from left 20 to right 20) in the sequence window. The process is shown in the Fig. 1.

Thus we got 503 sequence windows. Before doing cross-validation training, the windows which share over 40% similarities were discarded. This step provides the non-redundant data for SVMs to avoid over fitting. The similarity score of the two sequence windows was calculated using the following formula:

$$\text{score } S =: \sum_{i=1}^{n} f(X_i Y_i)$$

If X is same to Y on the "i" position, $f(X_i Y_i) = 1$, other then $f(X_i Y_i) = 0$. The final score S represents the similarity between the two sequences. If the score S between two windows is larger than 16, which is 40×0.4, one of the two sequence windows is left off from the dataset.

After the deletion process, we have a non-redundant positive dataset containing 261 sequence windows.

2.1.2. Negative dataset

We constructed the negative dataset by randomly choosing the serine (S)/threonine (T) windows from the mammalian protein sequences, which have no annotation related to the glycosylation in the Swissprot/Uniprot6.1. This step is totally independent from the positive dataset construction.

Because the number of the mammalian sequences in Swissprot/Uniprot is too large, we randomly choose 4000 sequences to construct the negative dataset. From the 4000 sequences, we got 99,619 sequence windows including the S/T site. By deleting the similar sequence windows using the method described above, 59,927 windows were selected for the non-redundant negative dataset.

2.1.3. Final dataset

The negative dataset is much larger than the positive dataset. To find the optimal proportion between the positive dataset and the negative dataset, we randomly choose 1, 2, 3, 4, 5 times records to the positive dataset in the negative dataset, meaning that we choose $261, 261 \times 2, 261 \times 3, 261 \times 4, 261 \times 5$ sequence windows in the negative dataset, respectively.

Finally, we got the 261×2, 261×3, 261×4, 261×5, 261×6 records in the final dataset to train SVMs.

2.2. Supported vector machines

Supported vector machine (SVM) is a supervised machine learning technology based on statistical theory for data classification. As well-founded statistical theory, this method is widely applied in the field of biological sciences because of its ability (Rost, 1993; Cai et al., 2000, 2001; Cai, 2002, 2003; Cai and Chou, 2003; Kim et al., 2004; Guo et al., 2004). It has been proven that SVM usually outperforms other machine learning methods in many fields of pattern recognition. More details about SVM can be found in Vapnik's publication (Vapnik, 1992; Vapnik and Vapnik, 1995; Corinna and Vapnik, 1995) or other machine learning publications (Joachims, 1999; Zavaljevski and Reifman, 2002; Chung et al., 2003).

In this paper, the SVM package named SVMLight (Joachims, 1999) is used to perform the task of prediction. This package can be freely downloaded from http://svmlight.joachims.org/. Users can select certain parameters for the different kernel type and SVM type. Further information can be found on their website.

2.3. 'Leave-one-out' cross validation

Cross validation is a method for estimating generalization classification error. It can also be used for model selection. For the method in the paper the popular 'leave-one-out' cross-validation method was chosen for estimation of the generalization error.

2.4. Three kinds of SVMs based on different sequence features

2.4.1. SVM based on amino acid physical properties

There are many kinds of physical properties for the amino acid. Previous work related to 188 amino acid physical properties has been done based on statistical analysis (Scheragat, 1985). Ten orthogonal factors were extracted to represent the 188 amino acid properties in that work. In this paper, we used

the ten orthogonal factors to code the amino acid. The sequence window (excluding the S/T) is represented by 10×40 dimension vectors for the SVM training process.

2.4.2. SVM based on 0/1 system

The amino acid coding scheme, 0/1 system, has been used in many papers (Kim et al., 2004). The common 20 amino acids are coded by 20-D vectors only composed of 0 and 1 (alanine = 00000000000000000001, cysterine = 00000000000000000010 and so on). Then the 40 symbol window (exclude the S/T) is represented by 40×20 dimension vectors as the SVM input.

2.4.3. SVM based on combined features

To avoid losing any information in the features, we combined the above two features: the 0/1 system and the amino acid physical properties. So the sequence window is represented by $10 \times 40 + 20 \times 40 = 1200$ dimension vectors for SVM training.

2.5. Prediction system assessment

Accuracy (Ac), Sensitivity (Sn), Specificity (Sp), and receiver operating characteristics (ROC) are often used to evaluate prediction systems.

Sn, Sp, and Ac are expressed in terms of true positive (TP), false negative (FN), true negative (TN), and false positive (FP) predictions. Each measurement is given as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

The ROC (Johnson, 2004; Bewick, 2004; Sing et al., 2005) curve is an effective method for evaluating the performance of the prediction system. It is commonly defined as a plot of sensitivity on the *Y*-axis versus the false positive rate on the *X*-axis. The area under the ROC curve (AUC) is also important. The bigger the AUC is, the better the overall prediction system performance is. In the paper, ROCR (Sing et al., 2005) and Bioconductor (Gentleman et al., 2004) package in *R* have been used to compute the related work to the ROC.

3. Prediction results

The results of SVMs were summarized in Table 1 and Fig. 2. For the three prediction models, they have the same patterns (Table 1). As the dataset becomes larger, the specificity improves, and the sensitivity gets worse. The best AC is 0.89. Meanwhile, the sensitivity is only 0.51.

The ROC curves and AUC (area under the ROC curves) give us a better assessment for the models performances. The ROC curves and AUCs of the three different features and five datasets are shown in Fig. 2.

Table 1
This table show the overall results from the different models based on different datasets and different amino acid coding schema

	Positive	Negative	Accuracy	Sensitivity	Specificity
Physical-properties					
Dataset1	261	261	0.81	0.85	0.78
Dataset2	261	522	0.83	0.71	0.88
Dataset3	261	783	0.81	0.49	0.91
Dataset4	261	1044	0.84	0.38	0.95
Dataset5	261	1305	0.85	0.34	0.96
0/1 System					
Dataset1	261	261	0.85	0.87	0.83
Dataset2	261	522	0.87	0.78	0.92
Dataset3	261	783	0.87	0.63	0.95
Dataset4	261	1044	0.87	0.56	0.95
Dataset5	261	1305	0.89	0.51	0.97
Combined-features					
Dataset1	261	261	0.85	0.87	0.83
Dataset2	261	522	0.84	0.63	0.95
Dataset3	261	783	0.80	0.21	0.99
Dataset4	261	1044	0.81	0.05	1
Dataset5	261	1305	0.83	0	1

We can find some differences between different datasets (Fig. 2). The model based on dataset 1:2 outperformed other models. And the results of the models based on 0/1 system is better than that of the physical properties and combined system. Fig. 3d shows detailed information about the AUC. It is clear that the model based on 0/1 system and dataset 1:2 has the best performance.

4. Describing the web server

The web server interface is shown in Fig. 3. In the web server, model based on 0/1 system and dataset 1:2 was used. The web application is very easy to use. The user pastes the fasta sequences in the text box area and the result will be given soon, as shown in Fig. 4.

5. Discussion

Normally, most of the machine learning techniques performs very well in this kind of prediction if we can extract enough key messages described by features. Comparing with other machine learning techniques, SVM is one of the best methods to classify this kind of data. Our results show that highly successful prediction results can be obtained if some appropriate features can be given.

In the dataset selection, the negative dataset is the data set in which the serine and threonine residues have not been glycosylated. It's very difficult to construct such data sets because the experiments always have little information about which residue had not been glycosylated. The idea used in the NetOglyc 2.1 is that the sequence windows which did not share similarities to the positive dataset have a lower potential to be glycosylated. This data processing method was not adopted in this paper because we do not have enough evidence to prove that less sim-

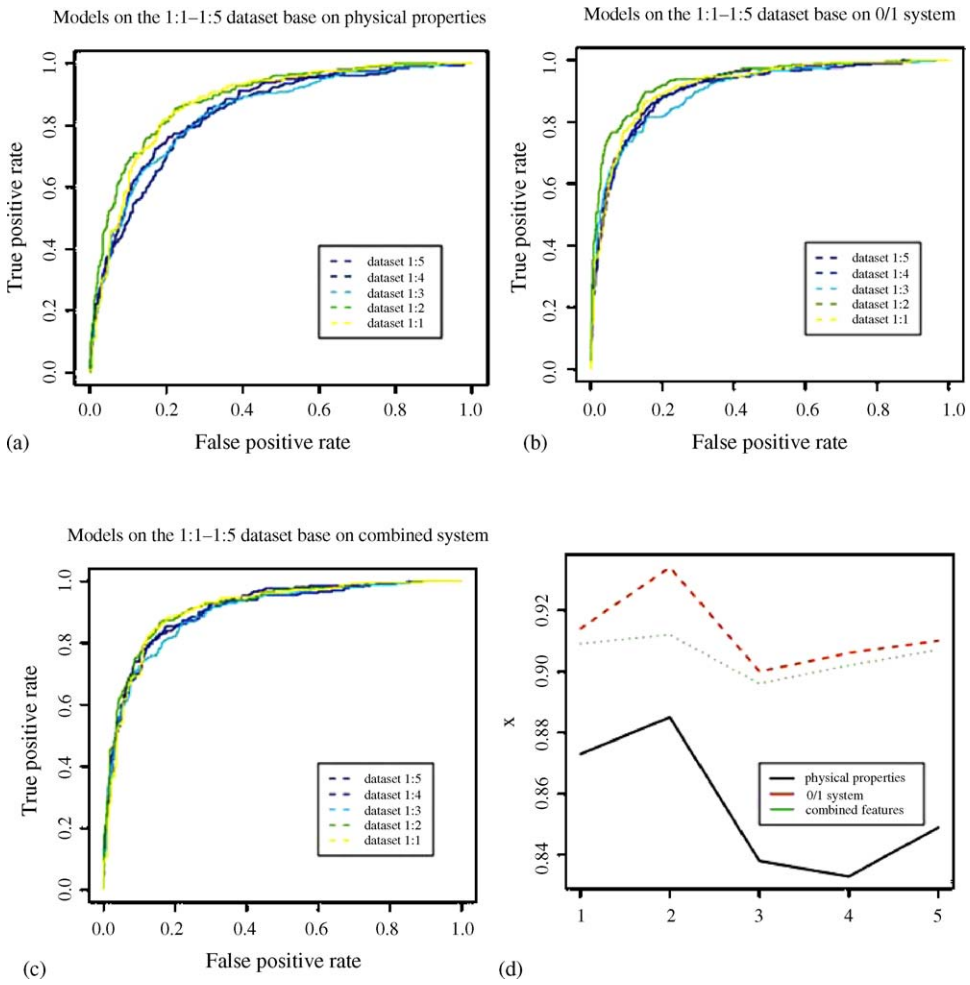


Fig. 2. This figure show the ROC and AUC of the different models. (a) The ROC curves of the different datasets on the physical properties. (b) The ROC curves of the different datasets on the 0/1 system. (c) The ROC curves of the different datasets on the combined features. (d) The AUC based on different datasets and different features.

ilar sequence windows have less potential to be glycosylated. So we constructed the negative dataset by randomly choosing the serine (S)/threonine (T) windows from the abundant mammalian protein sequences in the Swissprot/Uniprot, which have no annotation about the glycosylation, to be as the background dataset.

Different dataset having different positive/negative proportion is another key point of this paper. In biological research,

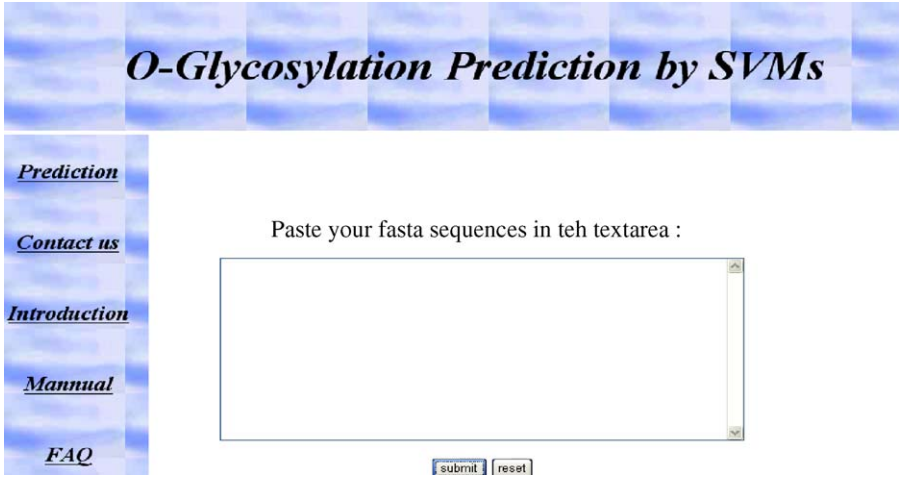
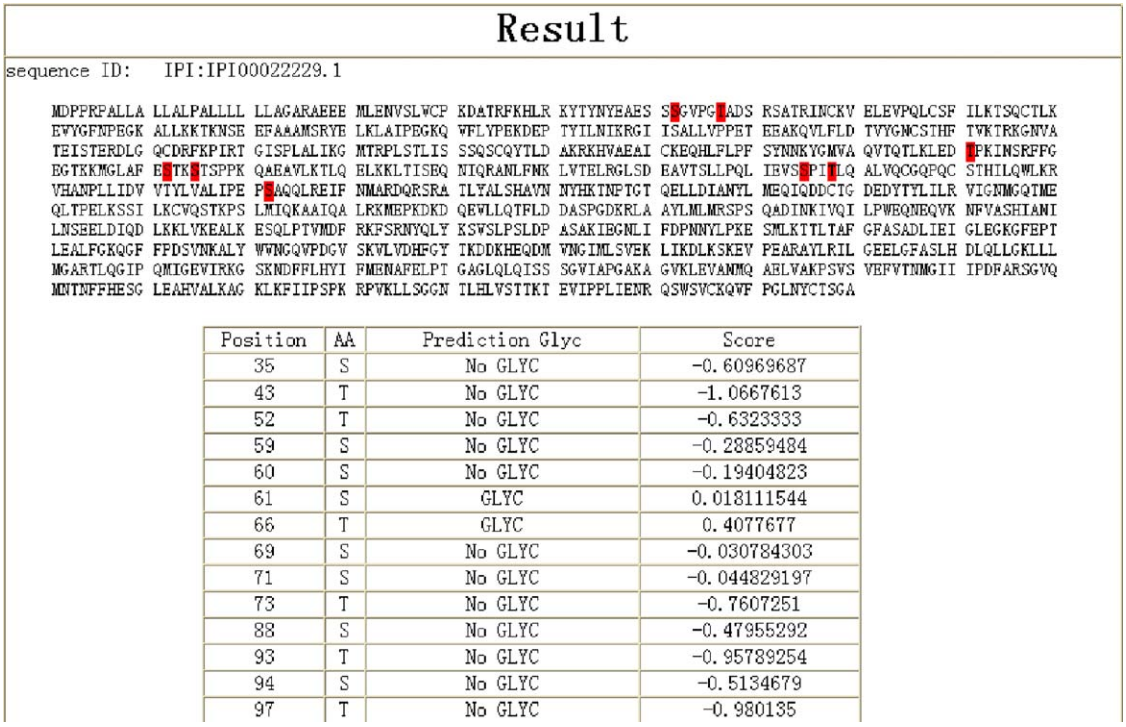


Fig. 3. The interface of the OglycPred web server, which is available at www.biosino.org/Oglyc.



Result

sequence ID: IPI:IPI00022229.1

MDPPRPALLA LLALPALLLL LLAGARAEEE MLENVSLWCP KDATRFKHLR KYTYNYEAES SSGVPGTADS RSATRINCKV ELEVPQLCSF ILKTSQCTLK
EWYGFNPEGK ALLKKTKNSE EFAAAMSRYE LKLAIPEGKQ WFLYPEKDEP TYILNIKRGI ISALLVPPET EEAKQVLFLD TVYGNCSTHF TVKTRKGNVA
TEISTERDLG QCDRFKPIRT GISPLALIKG MTRPLSTLIS SSQSCQYTLD AKRKHVAEAI CKEQHLFLPF SYNNKYGMVA QVTQTLKLED IPKINSRFFG
EGTKKMGLAF ESTKSTSPPK QAEAVLKTLQ ELKKLTISEQ NIQRANLFNK LVTELRGLSD EAVTSLLPQL IEVSSPITLQ ALVQCGQPQC STHILQWLKR
VHANPLLIDV VTYLVALIPE PSAQQLREIF NMARDQRSRA TLYALSHAVN NYHKTNPTGT QELLDIANYL MEQIQDDCTG DEDYTYLILR VIGNMGQTME
QLTPELKSSI LKCVQSTKPS LMIQKAAIQA LRKMEPKDKD QEWLLQTFLD DASPGDKRLA AYLMLMRSPS QADINKIVQI LPWEQNEQVK NFVASHIANI
LNSEELDIQD LKKLVKEALK ESQLPTWMDF RKFSRNYQLY KSWSLPSLDP ASAKIEGNLI FDPNNYLPKE SMLKTTLTAF GFASADLIEI GLEGKGFEPT
LEALFGKQGF FPDSVNKALY WVNGQVPDGV SKVLVDHFGY IKDDKHEQDM VNGIMLSVEK LIKDLKSKEV PEARAYLRIL GEELGFASLH DLQLLGKLLL
MGARTLQGIP QMIGEVIRKG SKNDFFLHYI FMENAFELPT GAGLQLQISS SGVIAPGAKA GVKLEVANMQ AELVAKPSVS VEFVTNMGII IPDFARSGVQ
MNTNFFHESG LEAHVALKAG KLKFIIPSPK RPVKLLSGGN TLHLVSTTKT EVIPPLIENR QSWSVCKQVF PGLNYCISGA

Position	AA	Prediction Glyc	Score
35	S	No GLYC	-0.60969687
43	T	No GLYC	-1.0667613
52	T	No GLYC	-0.6323333
59	S	No GLYC	-0.28859484
60	S	No GLYC	-0.19404823
61	S	GLYC	0.018111544
66	T	GLYC	0.4077677
69	S	No GLYC	-0.030784303
71	S	No GLYC	-0.044829197
73	T	No GLYC	-0.7607251
88	S	No GLYC	-0.47955292
93	T	No GLYC	-0.95789254
94	S	No GLYC	-0.5134679
97	T	No GLYC	-0.980135

Fig. 4. The sample result of the web server.

we will meet many situations in which the negative dataset is much larger than the positive dataset. An optimal proportion is unpredictable. In this paper, we test five different proportions (positive/negative) from 1:1 to 1:5. Our conclusion is that the 1:2 dataset have the best performance.

To this day, we cannot give a reasonable explanation for the behaviour of PTMs. Therefore, we can only test all possible reasons to find some useful rules for prediction. In this paper, we tested about 10 kinds of physical–chemical orthogonal properties which represent 188 kinds of amino acid properties to train the SVMs. In the end, a reasonably high accuracy was reached. It is obvious if given comprehensive information, machine-learning method can be a very powerful tool for enhancing the accuracy of predictions.

Besides physical properties as the SVM input, the 0/1 system is also used. The 0/1 system is commonly used in the protein-related prediction method. This kind of transformation from protein to data signal is efficient. In our prediction test, the 0/1 system-based prediction model performed better than the model based on physical properties features, even though the physical properties have complex information. It is clear that the 0/1 system has not lost any features for the prediction.

The third model based on the combined features of the 0/1 system and properties features was also built. The performance of the combined model is better than the physical properties, and similar to the 0/1 system. We think it is because that the physical-properties features have not supplied adequate complementary information to the 0/1 system.

In conclusion, a new and powerful machine learning method for predicting the *O*-glycosylation sites has been developed here. This method can not only help people to get more function information about *O*-glycosylation sites in mammalian proteins, but also can be integrated into proteomics data analysis systems to assist researchers to annotate proteome comprehensively.

References

Bairoch, A.B., 1993. The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Res. 21, 3093–3096.

Bewick, V., 2004. Statistics review 13: receiver operating characteristic curves. Crit. Care 8, 508–512.

Blom, N., 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4, 1633–1649.

Cai, D.L., Chou, K.C., 2003. Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24, 159–161.

Cai, D.L., Xu, X., Zhou, G.P., 2001. Support vector machines for predicting protein structural class. BMC Bioinformatics 2 (3), 3.

Cai, D.L., Xu, X.B., Chou, K.C., 2000. Support vector machines for prediction of protein subcellular location. Mol. Cell Biol. Res. Commun. 4, 230–233.

Cai, Y.D., 2002. Support vector machines for predicting HIV protease cleavage sites in protein. J. Comput. Chem. 23, 267–274.

Cai, Y.D., 2003. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. 84, 3257–3263.

Chung, M.K., Sun, C.L., Wang, L.L., Lin, C.J., 2003. Radius margin bounds for support vector machines with the RBF kernel. Neural Comput. 15, 2643–2681.

Cortes, C., Vladimir, V., 1995. Support-vector networks. Mach. Learn.

Eisenhaber, B.B., Eisenhaber, F., 1999. Prediction of potential GPI-modification sites in proprotein sequences. J. Mol. Biol. 292, 741–758.

Gentleman, C.C., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Guo, J.C., Sun, Z., Lin, Y., 2004. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins 54, 738–743.

Hansen, E.L., Engelbrecht, J., Bohr, H., Nielsen, J.O., Hansen, J.E., 1995. Prediction of *O*-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyl transferase. Biochem. J. 308, 801–813.

Hansen, E.L., Tolstrup, N., Gooley, A.A., Williams, K.L., Brunak, S., 1998. NetOglyc: prediction of mucin type *O*-glycosylation sites based on sequence context and surface accessibility. Glycoconj. J. 15, 115–130.

Jenkins, N.P., James, D.C., 1996. Getting the glycosylation right: implications for the biotechnology industry. Nat. Biotechnol. 14, 975–981.

Joachims, T., 1999. Making large-scale SVM learning practical. Adv. Kernel Methods.

Johnson, N.P., 2004. Advantages to transforming the receiver operating characteristic (ROC) curve into likelihood ratio co-ordinates. Stat. Med. 23, 2257–2266.

Julenius, K.M., Gupta, R., Brunak, S., 2005. Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 15, 153–164.

Kim, H.L., Oh, B., Kimm, K., Koh, I., 2004. Prediction of phosphorylation sites using SVMs. Bioinformatics 20, 3179–3184.

Rost, B.S., 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584–599.

Scheragat, T.O.H.A., 1985. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Protein Chem. 4, 23–55.

Sing, T.S., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in *R*. Bioinformatics 21, 3940–3941.

Vapnik, M.G.V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory.

Vapnik, V.N., Vapnik, C.V., 1995. The nature of statistical learning theory support-vector networks. Mach. Learn.

Zavaljevski, N.S., Reifman, J., 2002. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics 18, 689–696.