Asian Pacific Journal of Microbiology Research (AJMR)
DOI : http://doi.org/10.26480/ajmr.01.2017.12.14

ISSN: 2329-8251 (Print)
ISSN: 2329-826X (Online)

PREDICTION OF THE PROTEIN O-GLYCOSYLATION SITES BY COMBINING SUPPORT VECTOR MACHINES AND INDEPENDENT COMPONENT ANALYSIS

Xue Mei Yang[1*], Zhen Su[2]

School of Mathematics and Information Science, Xianyang Normal University, Xianyang, China
School of International Business, Southwestern University of Finance and Economicis,Chengdu,China

ARTICLE DETAILS

Article History:

Received 12 November 2017
Accepted 12 December 2017
Available online 1 January 2018

ABSTRACT

O-glycosylation is one of the main types of the mammalian protein glycosylation, it occurs on the particular site of serine and threonine, and has important functions in secretion, antigenicity, and metabolism of glycoproteins, it is very important to predict the O-glycosylation sites for pharmacy, food, and disease control. To improve the prediction accuracy, we proposed a new method of ICA+SVM. The samples(protein sequence) for experiment are encoded by the sparse coding with window size w=21, 120 independent components(feature) are extracted by independent component analysis(ICA), and inputed to the support vector machines(SVM), then the prediction(classification) is done in feature space by SVM. The results of experiment show that the performance of ICA+SVM is better than that of PCA+SVM and SVM. The prediction accuracy is about 88%. Furthermore, we investigated the same protein sequence under various window size, the results indicate that the longer the length of protein sequence, the higher the prediction accuracy.

KEYWORDS

Prediction, Protein, Glycosylation, ICA,SVM.

1. INTRODUCTION

Many Glycosylation is the most common post-translation modification of protein in eukaryotic cells, and has important functions in secretion, antigenicity, and metabolism of glycoproteins. There are four types of glycosylation: N-linked glycosylation to the amide nitrogen of asparagines side chains, O-linked glycosylation to the hydroxyl of serine and threonine side chains(Fig.1), C-linked glycosylation to the tryptophan side chains and GPI. Here we only focus on O-linked glycosylation protein sequence. In fact, not all serine or threonine residue are glycosylated and about 10%-30% protein can't be glycosylated.

There are many factors which affect this process, so it is very important to predict the O-glycosylation sites. Based on a study, many computational methods based on artificial neural networks (ANN) and support vector machines(SVM) have been developed for prediction of O-glycosylation sites [1-3]. The prediction accuracy can be achieved more than 70%. A researcher used a new protein bioinformatics tool, CKSAAP_OGlySite, to predict mucin-type O-glycosylation serine or threonine sites in mammalian proteins, under the composition of k-spaced amino acid pairs (CKSAAP) based encoding scheme, with the assistance of SVM [4]. His method yielded a higher accuracy of 83.1% and 81.4% in predicting O-glycosylated S and T sites, respectively.

Principal component analysis (PCA) is a statistical method for feature extracation, it can reduce the dimension and eliminate relativity of the original data, so it plays a key role in many research areas of science and engineering. In previous study, a researchers used PCA for pattern analysis, and we first found and verified by computational methods that O-glycosylation is abundant near the C terminus for serine [5]. But PCA is an analysis method based on the 2-order statistics feature of the original data, it can't eliminate the high-order relativity of each components of data, and the most part of information of data are included in the high-order statistics feature, so PCA can't recognize the data accurately.

According to a study, independent component analysis (ICA) is a statistics method based on the high-order statistics feature of the original data, it is a linear transform which can eliminate the high-order relativity of each components of data, and make each components independent, so the data after ICA transform can be recognized more accurately [6]. Study showed SVM is a kind of supervised machine learning technology for many two classes of classification problem, in this paper, we proposed a new method of ICA+SVM to predict the O-glycosylation site in protein sequence [7]. We first extracted features of original data by ICA, then used SVM for classification in the feature space. The remainder of the paper is organized as follows. Sec.2 presents protein sequence data and their coding. Section 3 describes the algorithm of ICA+SVM for prediction. Prediction results are shown in Section 4. The conclusions are given in Section 5.

2. PROTEIN SEQUENCE DATA AND ENCODING

Based on a research, the protein sequence data used in this research is from glycosylation database Uniprot (v8.0) [8]. We selected 99 mammalian protein entries, each entry contains some serine and threonine residue sites which are annotated experimentally as being glycosylated, together with other serine and threonine residue sites which have no such annotations. We call the former a positive site(positive S or positive T), while the latter a negative site(negative S or negative T). Each selected protein entry (sequence) is truncated by a window (window size: w) into several subsequences with S or T residues at the center, Fig.2 shows an example of subsequences (w=5). The protein sequence (exclude S or T at the center)with a length of w-1 are used for analysis. We use the sparse coding scheme for representation of the protein sequence. In sparse coding, 21-binary sequence is used to code one site of amino acid or vacancy, for example,

the site of amino acid I is coded as 10000000000000000000, the site of amino acid V is coded as 01000000000000000000. Thus the total length of coded sequence or dimension of sample vector is (w-1) *21. The number of samples for each class is summarized in Table1. Since the number of negative sites is much larger than that of positive sites, we randomly chose 100 samples from each class for training, and 50 samples from each class for testing.

Table 1: Number of samples for experiments

	Total Number	Number for training	Number for testing
Positive S	174	100	50
Positive T	292	100	50
Negative S	693	100	50
Negative T	841	100	50

3. INDEPENDENT COMPONENT ANALYSIS AND SUPPORT VECTOR MACHINES FOR PREDICTION

The prediction can be viewed as a 2-class (positive and negative) classification problem. We first constructd a feature space of protein sequence by using ICA, and projectd the samples into the feature space, then classified the test samples into 2 classes in the feature space by SVM. ICA is a statistics method based on the high-order statistics feature of the original data. In fact, most of the important information is included in the high-order statistics feature. Since each component of data after ICA transform is independent, we can recognize them more accurately. The model of ICA can be showed as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad (1)$$

Where $\mathbf{s}$ is the blind source, we suppose that each component of $\mathbf{s}$ is independent, $\mathbf{A}$ is an unknown hybrid matrix, $\mathbf{x}$ is known signal. We need to find a matrix $\mathbf{W}$, such that

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x} \quad (2)$$

$\tilde{\mathbf{x}}$ should be approached $\mathbf{s}$.

Central limit theorem indicates that the hybrid signal of multiple random variables approaches Gauss distribution, so the stronger the non-Gaussianity of $\tilde{\mathbf{x}}$, the closer it from independent source $\mathbf{s}$, and the closer $\mathbf{W}$ from $\mathbf{A}^{-1}$, the non-Gaussianity of a random variables can be described by kurtosis, kurtosis is a fourth-order cumulant, so we should consider the extremum of kurtosis, this is a problem of optimization.

Before the transform of ICA, we need to preprocess $\mathbf{x}$, we first center $\mathbf{x}$ by the following,

$$\mathbf{x} = \mathbf{x} - \mathbf{E(x)} \quad (3)$$

So as to make $\mathbf{x}$ a zero-mean variable, then we whitening it by the following,

$$\mathbf{z} = \mathbf{D}^{-1/2}\mathbf{E}^{\mathbf{T}}\mathbf{x} \quad (4)$$

So as to eliminate the two-order relativity of data, where $\mathbf{D}$ and $\mathbf{E}$ are the eigenvalue and eigenvector matrix of $\mathbf{C}$ separately, $\mathbf{C}$ is covariance matrix of $\mathbf{x}$.

By retaining the top k($<n$,n is the dimension of $\mathbf{x}$) eigenvectors(which corresponding the top k eigenvalues), we can reduce the dimension.

From the Fast ICA algorithm, the ith row of $\mathbf{W}$ is

$$\mathbf{w_i} = \mathbf{E}\{\mathbf{z}f(\mathbf{w}_i^{\mathrm{T}}\mathbf{z})\} - \mathbf{E}\{f(\mathbf{w}_i^{\mathrm{T}}\mathbf{z})\}\mathbf{w_i} \quad (5)$$

Where $f(u) = \tanh(u)$.

We apply ICA algorithm on $\mathbf{z}$ by (5) and obtain

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{z} \quad (6)$$

SVM SVM is a kind of supervised machine learning technology for many two classes of classification problem. The classification idea of SVM is mapping the training vectors into multidimensional space by a kernel function, and then construct a hyperplane optimally positioned between the positive and negative samples; a testing sample is then projected into the multidimensional space to determine its class affiliation based on its relative position to the hyperplane.

For a given training set $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_M, y_M)\}$, $\mathbf{x}_i \in \mathbf{R}^d$, $y_i \in \{-1,1\}$, the equation of separating hyperplane is

$$\mathbf{w} \cdot \phi(\mathbf{x}) + b = 0$$

Then the objective function of SVM is

$$\min_{w,b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{M}\xi_i^2$$
$$s.t. \quad y_i = \mathbf{w} \cdot \phi(\mathbf{x}_i) + b + \xi_i, i = 1, \cdots, M \quad (7)$$

Where $\frac{2}{\|\mathbf{w}\|}$ is the margin, $\boldsymbol{\xi}$ is the margin slack vector, the parameter C controls the trade-off between the margin and the size of the slack variables.

By using Lagrange multiplier method, we can obtain the solution of (7) $\mathbf{w}^* = \sum_{i=1}^{M}\alpha_i^*\phi(\mathbf{x}_i)$, and $b^* = y_j - \sum_{i=1}^{M}\alpha_i^* \mathrm{K}(\mathbf{x}_j, \mathbf{x}_i) - \frac{\alpha_j}{2C}$, here $\boldsymbol{\alpha}^*$ is Lagrange multiplier, $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is kernel function. So the classifier (decision function)is shown in(8)

$$f(\mathbf{x}) = \sum_{i=1}^{M}\alpha_i^*\mathrm{K}(\mathbf{x}_i, \mathbf{x}) + b^* \quad (8)$$

The choice of kernel function must be subject to the Mercer's theorem.

Usually we use the following kernel function linear kernel function

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}) = \langle \mathbf{x}, \mathbf{x}_i \rangle$$

(a) quadratic kernel function

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}) = \langle \mathbf{x} \cdot \mathbf{x}_i \rangle (\langle \mathbf{x} \cdot \mathbf{x}_i \rangle + 1)$$

(b) polynomial kernel function

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}) = [\langle \mathbf{x}, \mathbf{x}_i \rangle + c]^d$$

(c) sigmoid kernel function

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}) = \tanh[v * \langle \mathbf{x}, \mathbf{x}_i \rangle + c]$$

(d) gauss radial basis kernel function(RBF)

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2})$$

A. Algorithm of ICA+SVM

Step1. Input the training samples $\mathbf{x} = \{\mathbf{x}_1, \cdots \mathbf{x}_M\}$;

Step2. Apply ICA algorithm on $\mathbf{x}$ by(3)(4)(5)(6), obtain the projection $\tilde{\mathbf{x}}$;

Step3. Project the test sample $\mathbf{y}$, get $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$;

Step4. Choose kernel function, take $\tilde{\mathbf{x}}$ as the input of SVM, compute $\boldsymbol{\alpha}^*$ and b^* by using Lagrange multiplier method, and get the decision function, $f(\tilde{\mathbf{x}}) = \sum_{i=1}^{M}\alpha_i^*\mathrm{K}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}) + b^*$;

Step5. Decide the class of $\tilde{\mathbf{y}}$ by decision function.

IV. PREDICTIONS AND RESULTS

We experimented with window size w=21 by using SVM, PCA+ SVM and ICA+ SVM. The dimension n of original data $\mathbf{x}$ is 21*(w-1)=420, since the accumulated variance (eigenvalue) of the top 120 components of ICA is more than 80%, we took $k = 120$, namely we extracted the first 120 independent components($\tilde{\mathbf{x}}$) as the input data of SVM, and 5 kinds of kernel functions are used for projection. The parameter C of SVM is varied from 0.05 to 100.The algorithm is implemented in MATLAB7.8.0. We tested 200 testing samples. The results of prediction are shown in Table2-6.

From Table 2-6, we can see that, when using polynomial kernel function, the best performance of ICA+ SVM is 84%, and the best performance of ICA+ SVM is 88% when using RBF, they are better than the best performance of SVM and PCA+ SVM, it's because that ICA eliminates the high-order relativity of each components of data, and makes the recognition of objective more accurate. We also see that the best parameters for polynomial kernel function and RBF are c=5, d=2 and σ =6 when using ICA+ SVM. When using linear, quadratic, and sigmoid kernel function, the performance of ICA+ SVM is still better than that of SVM and PCA+ SVM, but not better than that of polynomial kernel function and RBF.

Table 2 Prediction results (kernel function: polynomial c=5)

method	SVM				PCA+ SVM				ICA+ SVM			
parameter(d)	2	3	4	5	1	2	3	4	1	2	3	4
Prediction accuracy(%)	73.5	53.5	50	50	83	81.5	77.5	67	68	**84**	80	71

Table 3 Prediction results (kernel function: RBF)

method	SVM				PCA+ SVM				ICA+ SVM			
parameter(σ)	5	6	7	8	4	5	6	7	4	5	6	7
Prediction accuracy(%)	71	71	71	78.5	78.5	80	83	80.5	79	87	**88**	83

Table 4 Prediction results (kernel function: linear)

method	SVM	PCA+ SVM	ICA+ SVM
Prediction accuracy(%)	74	81.5	**82**

Table 5 Prediction results (kernel function: quadratic)

method	SVM	PCA+ SVM	ICA+ SVM
Prediction accuracy(%)	72	77	**79**

Table 6 Prediction results (kernel function: sigmoid)

method	SVM				PCA+ SVM				ICA+ SVM			
parameter(v,c)	C=-1,v=0.2	C=-1,v=2	C=-1,v=4	C=-1,v=6	C=-1,v=0.1	C=-1,v=1	C=-1,v=2	C=-1,v=3	C=-1,v=0.2	C=-1,v=2	C=-1,v=4	C=-1,v=6
Prediction accuracy(%)	53.5	73.5	74	60	79	81.5	79.5	77	68	**82**	80	71

Table 7 Prediction results (kernel function: RBF)

window size	5	7	9	11	21	31	41	51
σ	3.3	3.4	3.7	4	7	8.6	12	15.6
Prediction accuracy(%)	82	81.5	83.5	83.5	88	88.5	88.5	**89.5**

Table 8 Prediction results (kernel function: polynomial)

window size	5	7	9	11	21	31	41	51
c,d	c =4, d =3	c =6, d =3	c =8 , d =2	c =7 , d =2	c =5, d =2	c =5 , d =2	c =6 , d =2	c =4 , d =2
Prediction accuracy(%)	80	83	83.5	83	84	83.5	85	**85.5**

Furthermore, we investigated the same protein sequence under the window size w=5,7,9,11,31,41,51, by using the method of ICA+SVM, the results of experement are shown in Table 7 and Table 8.

With the increasing of window size, this is because that the feature of protein sequence with large window size is more distinctly than that of protein sequence with small window size; meanwhile, for protein sequence with small window size, such as 5 or 7, some input data contradict each other for the prediction, that is, some are positive and the others are negative with the same input sequence, this results in wrong classification. (2)When using polynomial kernel function, the best order d of it is different for different window size, when the window size is smaller, d is larger, and when the window size is larger, d is smaller. (3) Since σ is a width parameter, the parameter σ is increased with the increasing of window size, when using RBF kernel function.

V. CONCLUSIONS

We proposed a new method of ICA+ SVM to realize the prediction of O-linked glycosylated sites in protein sequence. The result of experiments shows that, ICA can eliminate the high-order relativity of each components of data, extract the independent components, and make the recognition more accurately; when using polynomial kernel function and RBF, the proposed method is more effective and accurate than SVM and PCA+ SVM.

In the future, we will try to predict the protein O-linked glycosylated sites by combining kernel independent component analysis (KICA) and support vector machines.

ACKNOWLEDGMENT

This work is partially supported by the Scientific Research Project of Science and Technology Department of Shaanxi Province (No. 11JK1050)

REFERENCES

[1] Nishikawa, I., Sakamoto, H., Nouno, I., Iritani, T., Sakakibara, K., and Ito, M.: Prediction of the O-glycosylation sites in protein by layered neural networks and support vector machines, Lecture Notes in Artificial Intelligence, Springer, LNAI 4252,pp953-960,2006

[2] Kenta Sasaki, Nobuyoshi Nagamine and Yasubumi Sakakibara: Support vector machines prediction of N- and *O*-glycosylation sites using whole sequence information and subcellular localizition[J],IPSJ Transactions on Bioinformatics, Vol.2, pp25-35, 2009

[3] Li, S. et al.: Predicting O-glycosylation sites in mammalian proteins by using SVMs, Computational Biology and Chemistry, Vol.30, pp.203-208 ,2006.

[4] Yong-zi Chen: Prediction of mucin-type O-Glycosylation sites in mammaliam protein using the composition of k-spaced amino acid pairs, BMC Bioinformatics, Vol.9, pp.101, 2008

[5] Xue-mei Yang, Yen-Wei Chen, Masahiro Ito and Ikuko Nishikawa: Principal Component Analysis of O-linked Glycosylation Sites in Protein Sequence[C], IEEE Third International Conference on IIHMSP, Vol.1, pp.121-126, 2007

[6] John Shawe-Taylor, Nello Cristianini: Kernel Methods for Pattern Analysis[M], China Machine Press, Beijing, 2005

[7] Xue-mei Yang: Prediction of the Protein O-Glycosylation by Machine Learning Based on Kernel Principal Component Analysis and Ensemble Classifiers, ICIC Express Letters, Vol.5(8B), pp. 2805-2810, 2011

[8] http://www.ebi.uniprot.