Glycosylation Site Prediction Using Machine Learning Approaches

Cornelia Caragea[1], Jivko Sinapov[2], Adrian Silvescu[3], Drena Dobbs[4], Vasant Honavar[5]

Keywords: Glycosylation site prediction, Machine Learning Approaches

1 Introduction

Glycosylation is one of the most complex post-translational modifications (PTMs) which occurs in many proteins in eukaryotic cells. In general, the result of glycosylation influences protein folding (some proteins cannot fold properly unless they have been glycosylated), protein localization, trafficking, biological activity and half-life, cell-cell interactions, developmental processes, etc. There exist four types of glycosylation, depending on the linkage between sugar and specific residues in a protein: *N-Linked Glycosylation* - the oligosaccharide chain (glycan) is attached to the amide nitrogen of asparagine (N), *O-Linked Glycosylation* - the glycan is attached to the hydroxy oxygen of serine (S) or threonine (T), *C-Mannosylation* - the glycan is attached to the carbon of tryptophan (W), and *GPI Anchor* - the glycan is attached close to the C-terminal of a protein.

2 Problem Formulation

We formulate the problem of glycosylation site prediction as a binary classification problem as follows: given a protein sequence X of length N, $X = x_1x_2\cdots x_N$ over the alphabet Σ of amino acids, $|\Sigma| = 20$, $x_i \in \Sigma$, $i = 1, \cdots, N$ and $x \in \Sigma^*$, the classification task is to predict glycosylation versus non-glycosylation sites from the protein sequence. We show how to apply machine learning algorithms to learn models that can predict three types of glycosylation, O-Linked versus non-O-Linked Glycosylation, N-Linked versus non-N-Linked Glycosylation, and C- versus non-C-Mannosylation. We use an ensemble classifier approach to learn the models.

3 Feature Representation

The dataset used in our experiments comes from O-GlycBase [3], a resource containing experimentally verified glycosylation sites compiled from protein databases and literature. The dataset is available online at http://www.cbs.dtu.dk/databases/OGLYCBASE/.

Glycosylation is a *site-specific process*. It occurs mainly on one of the four residues S, T, N, and W. However, not all of these residues in a protein sequence are actually modified by glycosylation. Therefore, we represent S, T, N, and W sites experimentally verified to be glycosylation sites as positive examples and S, T, N, and W sites not shown experimentally to be either glycosylation or non-glycosylation sites as negative examples.

[1] Department of Computer Science, Iowa State University. E-mail: `cornelia@cs.iastate.edu`

[2] Department of Computer Science, Iowa State University. E-mail: `jsinapov@cs.iastate.edu`

[3] Department of Computer Science, Iowa State University. E-mail: `silvescu@cs.iastate.edu`

[4] Department of Genetics, Department of Cell Biology, Iowa State University. E-mail: `ddobbs@iastate.edu`

[5] Department of Computer Science, Iowa State University. Email: `honavar@cs.iastate.edu`

Glycosylation is also an *enzymatic process.* It has been observed that the enzymes involved (the transferases) recognize a glycosylation site based on the residues surrounding the target. To exploit this observation, we use a local window with each glycosylation or non-glycosylation site in the middle and an equal number of its sequence neighbors on each side to further represent positive and negative examples, respectively.

4 Machine Learning Approaches

Support Vector Machines (SVMs) classifier is a supervised learning algorithm that belongs to the class of discriminative models. In our experiments, we used **SMVs** with different *kernel function*: 0/1 and Substitution Matrix String kernels on indentity window representation, and Polynomial and Radial Basis Function (RBF) kernels on Psi-Blast representation ([2, 7]). **Naive Bayes (NB)** classifier ([5]) is one of the simplest and easy to apply probabilistic approach. It belongs to the class of generative models. **Decision Tree Learning (J48)** classifier ([5]) is among the most widely used methods to extract patterns from data. In our experiments, the input to **NB** and **J48** classifiers is the indentity window. In this work, we trained an *ensemble classifier* ([6]) instead of a *single classifier.*

5 Results

To assess the performance of our classifiers, we report the following measures, defined in [1]: accuracy, correlation coefficient, and Area Under the Receiver Operating Characteristic Curve (AUC). Our methods outperformed the method used in [4] on the glycosylation site prediction problem. Due to space limitation, we report results only for O-Linked.

Classifier/	SVM				NB	J48
Performance Measure	0/1SK	SMK	BlastPoly	BlastRBF	Identity	Identity
Accuracy	0.88	0.85	0.87	0.87	0.88	0.86
Correlation Coefficient	0.58	0.58	0.57	0.57	0.56	0.47
AUC	0.91	0.91	0.90	0.89	0.87	0.84

Table 1: Experimental results for O-Linked glycosylation using SVMs with 0/1 String Kernel (0/1SK) and Substitution Matrix String Kernel (SMK) on identity window, and Polynomial (BlastPoly) and RBF Kernel (BlastRBF) on Psi-Blast representation,and NB and J48 on identity windows.

References

[1] Baldi,P., Brunak, S., Chauvin, Y., Andersen, C., Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. In: *Bioinformatics* 16:412-424.

[2] Burges, C. J. C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. In: *Data Mining and Knowledge Discovery* 2:121-167.

[3] Gupta, R., Birch, H., Rapacki, K., Brunak, S. and Hansen, J. 1999. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. In: *Nucleic Acids Res.* 27:370-2.

[4] Li, S., Liu, B., Zeng, R., Cai, Y., and Li, Y. 2006. Measuring genome evolution. Predicting O-glycosylation Sites in mammalian proteins by using SVMs. *Comput Biol Chem* 30:203-8.

[5] Mitchell, T.M. 1997. Machine Learning, *McGraw Hill*

[6] Russell, S. and Norvig, P. 2003. Artificial Intelligence: A Modern Approach, *Prentice Hall*

[7] Vapnik, V. 1998. Statistical Learning Theory, *New York: Springer Verlag*