A New Probabilistic Model for Glycans

Nobuhisa Ueda ueda@kuicr.kyoto-u.ac.jp
Kiyoko F. Aoki kiyoko@kuicr.kyoto-u.ac.jp
Atsuko Yamaguchi atsuko@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp
Tatsuya Akutsu takutsu@kuicr.kyoto-u.ac.jp
Hiroshi Mamitsuka mami@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: glycans, probabilistic models, labeled ordered trees, Markov property

1 Introduction

Glycans are branched or linear chains of monosaccharides attached to one another via glycosidic linkages, and have become recognized as participants in diverse cellular processes such as modifying structures of proteins and assisting in cell-cell and cell-matrix interactions. Since glycans are generally tree structures and more complex than nucleotide or amino acid sequences, it is impractical to apply methods or algorithms for sequences to glycans directly.

We therefore developed a new probabilistic model called a *probabilistic sibling-dependent tree Markov model* (PSTMM) [1] for glycans. The model is an extension of hidden Markov models which have been extensively used in analyzing biological sequences.

2 Method and Results

We consider the following two features for a probabilistic model for glycans: (i) glycans can be described as labeled rooted trees, and (ii) some monosaccharides are dependent on other ones in a glycan. As in Fig. 1 (a), glycans are basically tree structures, and each of them has one paticular monosaccharide linked to an amino acid of a protein. We can then describe a glycan as a labeled rooted tree, where labels on nodes, edges, and the root correspond to types of monosaccharides, linkages between two monosaccharides, and the monosaccharide linked to the amino acid, respectively. For brevity, we call one of two connected nodes (monosaccharides) close to the root a *parent* and the other a *child*.

For the second feature, we can find two types of dependencies among monosaccharides in a glycan. One is a dependency between a parent and a child. In synthesizing a glycan, monosaccharides are sequentially linked to the glycans with various glycosyltransferases in general. Since each glycosyltransferase usually has its own specificity, the linked monosaccharide can be dependent on its parent. The other type of dependencies is those among children. For example, if two monosaccharides (a parent and a child) are connected in a glycan, any other monosaccharide cannot be attached to the parent with the same binding position as that of the child until the child is separated from the glycan. Such dependencies can be sorted by the order of binding positions of children.

From these observations, one plausible candidate of probabilistic models for glycans is PSTMM whose graphical representation is depicted in Fig. 1 (b). Note that this model explicitly considers dependencies between ordered siblings as well as parent-child relationships in a glycan. Moreover, PSTMM can capture dependencies between distant nodes in a glycan with hidden variables, as shown

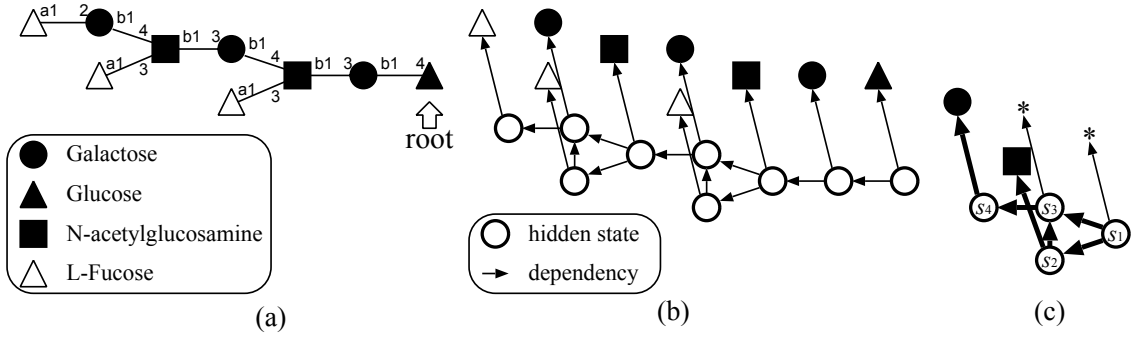

Figure 1: (a) An example of glycans, G00082, from the KEGG-Glycan database [2], (b) Graphical representation of PSTMM for the glycan G00082, (c) A dependency between distant nodes that can be captured with a PSTMM.

in Fig. 1 (c), where an asterisk stands for an arbitrary label and each thick line represents that the corresponding transition or emission probability is 1.

Due to its rich expressiveness, PSTMM becomes an inherently complex model, in other words, PSTMM is categorized into an intractable subclass of Bayesian networks called multiply-connected Bayesian networks. However, by making use of the Markov property, it is possible to efficiently find the likelihood of a glycan in $O(|S|^3|V|)$ time, where $|S|$ and $|V|$ denote the number of nodes and the number of states, respectively.

We compare the performance of PSTMM with that of another probabilistic model (mixture of label-pair model, MLPM) which does not consider any dependencies among children. We first obtained a data set of glycans from the KEGG-Glycan database [2], and selected glycans in two major classes, called 'N-Glycan' and 'O-Glycan' such that each glycan contained at least one sibling pair. For each class, we generated negative test examples so that the distribution of parent-child pair label was equivalent to that of the positive test examples. We evaluated the performance of each model using the five-fold cross validation by prediction accuracies (Acc.) and precisions (Prec.) at recall of 30%. The results are summarized in Table 1, and indicate that there must be some complex patterns that are not limited to parent-child relationships in glycans, and PSTMM successfully captured such complex patterns.

Table 1: Experimental results with glycan data sets

	N-Glycan		O-Glycan	
Methods	PSTMM	MLPM	PSTMM	MLPM
Acc.	**85.5**	64.5	**75.3**	63.8
Prec.	**95.6**	66.8	**84.1**	62.7

References

[1] Aoki, K. F., Ueda, N., Yamaguchi, A., Kanehisa, M., Akutsu, T., and Mamitsuka, H., Application of a new probabilistic model for recognizing complex patterns in glycans, *Proc. 12th Int'l Conf. on Intelligent Systems for Molecular Biology*, to appear, 2004.

[2] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori M., The KEGG resource for deciphering the genome, *Nucleic Acids Research*, 32:D277–D280, 2004.