2005－BIO－2（1）
2005／10／7

糖鎖の木構造からプロファイルを検出する確率木モデルの応用

木下聖子、上田展久、馬見塚拓、五斗進、金久實
京都大学化学研究所バイオインフォマティクスセンター

バイオインフォマティクスにおいては、ノイズの多い生物情報を処理するために確率モデルがよく使われている。以前、PSTMM という木構造の兄弟の依存性を含めた確率木マルコフモデルが紹介された。本研究では、PSTMM を糖鎖木構造からプロファイルを検出する、*profile PSTMM* というモデルに拡張した。パラメータ推定アルゴリズムを修正したことで、最近の糖鎖インフォマティクスの発展で増えた、ノイズの多い糖鎖情報から更に実用的な情報を得られた。糖鎖を認識するガレクチンと言うタンパク質の糖鎖との結合親和力データを用いて実験を行なった。

Application of a probabilistic tree model for determining profiles in glycan tree structures

Kiyoko F. Aoki-Kinoshita, Nobuhisa Ueda, Hiroshi Mamitsuka,
Susumu Goto, Minoru Kanehisa
Bioinformatics Center, Institute for Chemical Research, Kyoto University

Probabilistic models are a common and useful tool in bioinformatics due to the amount of noise known to persist in biological data. A probabilistic tree Markov model, called PSTMM, that captures dependencies between siblings has been introduced previously. In this work, we have modified this model to retrieve profiles in tree structures of carbohydrate sugar chains, or glycans, calling this new model *profilePSTMM*. While the parameter estimation algorithms are only slightly modified, more useful information is obtained to characterize glycan structures, which currently contain much noise because of the recent advances of informatics in the field of glycobiology. We introduce this new profile version of PSTMM, and we illustrate its application to glycan structures known to be recognized by glycan-recognizing proteins called galectins.

1 Introduction

Glycans are the third major class of biomolecules, after DNA and proteins. In contrast to the linear structure of DNA and proteins, glycans are branched, tree structures of monosaccharide units, or sugars. These glycans are mainly found on the cell surface, and they are heavily involved in major biological functions as their structures provide the signals for various developmental processes. This occurs through recognition of glycan structures by proteins; thus the interactions between glycans and proteins and their binding formations control the processes in the cell and the cell matrix. It has been conjectured that the structures at the leaves of the tree (subtrees) are recognized by lectins in order to facilitate binding. Thus, we were interested in capturing such formations using a probabilistic model. (For further basic information regarding glycans and glycobiology, the reader is referred to Varki et al.[16].)

Probabilistic models have been used in bioinformatics since the days of Margaret Dayhoff when she first developed a score matrix for protein sequences to better assess amino acids according to their evolutionary distances[7]. Since then, the BLOSUM score matrix[11] was introduced based on a more computational approach focused on the data at hand and their alignment frequencies, followed by the now popular hidden Markov models (HMM)[5, 9] and profile HMMs[8], which eventually led to the development of databases for motif profiles of amino acid sequences such as Pfam[4] and SCOP[14].

Probabilistic models for trees first came out in 1998 with an application to signal processing[6] and later in 2001 for multiscale image segmentation[13]. Then Ueda *et al.* developed one for labeled ordered trees with a sibling-dependent tree Markov model called PSTMM[15, 1]. This model was

applied to carbohydrate sugar chains, or glycans, and it was able to successfully capture patterns known to exist in one of the largest classes of these glycans.

In this work, we have extended the PSTMM model to one that captures motif profiles in glycans, in a model aptly called profile PSTMM. We illustrate the utility of this new model with the preliminary experimental results of motif profiles found in the glycan structures that are recognized by proteins called galectins.

2 Background

Because of the diversity of fields covered in this paper, we provide a brief explanation of some background information, including basic terminology, introduction to glycobiology and glycans, and the PSTMM model.

2.1 Basic Terminology

Two nodes in a tree x and y are *siblings* if they have the same parent, and a node with no children is a *leaf*. Siblings are ordered, so the first child of a parent is considered the elder sibling of the second, etc.

The following notation is used in this paper. $\mathbf{T} = \{T_1, \ldots, T_{|\mathbf{T}|}\}$ is a set of labeled ordered trees, where $T_u = (V_u, E_u)$, V_u $(=\{x_1^u, \ldots, x_{|V_u|}^u\})$ is a set of nodes, and E_u is a set of edges. x_1^u is the root of tree T_u, $|V| = \max_u |V_u|$, $t_u(i)$ is a subtree of T_u, having x_i^u as the root of $t_u(i)$, $C_u(p) \subseteq \{1, \ldots, |C_u(p)|\}$ is a set of indices of children of x_p^u in T_u, and $|C| = \max_{u,p} |C_u(p)|$. If $x_{\leftarrow}^u(p)$ and $x_{\rightarrow}^u(p)$ are the eldest and youngest child of node p, respectively, then $Y_u(p) = C_u(p) - x_{\leftarrow}^u(p)$. Each node x_j^u has label $o_j^u \in \Sigma$, where $\Sigma = \{\sigma_1, \ldots, \sigma_{|\Sigma|}\}$ is the set of labels (i.e., the alphabet) applied to the nodes. For simplicity, we will often use j for node x_j^u if understood from the context, and for node j, we will use i, k and p to refer to the immediately elder sibling, the immediately younger sibling, and the parent, respectively.

2.2 Glycobiology and Glycans

Glycobiology is the study of glycans, which covers the study of carbohydrate structure as well as function in interaction with proteins and the biological system. There has been much work recently in glycoinformatics, which entails the structural determination of specific glycan structures through such technologies as mass spectrometry and nuclear magnetic resonance (NMR). A term we use for the computational analysis of glycan data at the glycome level is glycome informatics, starting from glycan structure comparison algorithms[3], corresponding score matrices[2], onto expression analysis of glycosyltransferases and glycome level representations of all possible glycan structures known[10].

Glycan structures themselves are labeled ordered trees, with nodes representing monosaccharides and edges representing glycosidic bonds. Glycosidic bonds link two monosaccharides by one of their hydroxyl groups, usually 2, 3, 4, or 6, via one of two types of linkages, α or β.

2.3 PSTMM

The probabilistic sibling-dependent tree Markov model was shown to be able to capture patterns in tree structures, especially glycans[15]. Algorithms were also developed which could estimate parameters and find most likely paths within the bounds of the maximum known limits. In comparison to HMMs and tree Markov models, PSTMM included dependencies between siblings such that the order between them could be maintained.Then, in addition to the classic forward and backward parameters of Baum-Welch, upward and downward parameters were incorporated to estimate parameters most efficiently. Thus, a tree's parameters would be estimated starting from the leaves and traveling up the parents, and forward and backward between siblings, up to the root. Then downward parameters would be estimated in a breadth-first fashion. Finally, most likely state paths were estimated by finding the states with the highest probabilities.

Experiments were performed on glycan structures using PSTMM, and it was shown that patterns could indeed be captured better than previous models. In fact, one of the most popular classes of glycans called N-glycans were analyzed, and PSTMM found the three known subclasses of N-glycans called hybrid, high-mannose, and complex type, which are characterized by patterns at their leaves. Thus the utility of this model in bioinformatics was illustrated.

3 Profile PSTMM

The incorporation of different types of states into PSTMM results in a profile PSTMM where motif profiles may be determined from a set of tree structures. However, this is easier said than done. First, each state in PSTMM is considered a match state. Then insert and delete states are added for each corresponding match state position. Transitions are added appropriately, including insert loops, as well as a begin state to the root state position, and an end state leaving the youngest sibling among all leaves. Also, instead of one type of transition, this new model contains two: one for going down (to match children) and one for going right (to match younger siblings). Figure 1 illustrates this new model. In this figure, match and delete states are combined for simplicity as they always appear at the same positions. Similarly, when both down and right transitions start and end at the same states, they are drawn as a single black transition representing both.

Just as profile HMMs implement insertion and deletion states, profile PSTMMs also implement M, I, and X as match, insertion, and deletion states, respectively. There is also a single BEGIN state and a single END state. BEGIN is a silent state that transitions to the first root state and only goes down. The silent END state transitions from the youngest child leaf states. The profile PSTMM model thus looks like the following:

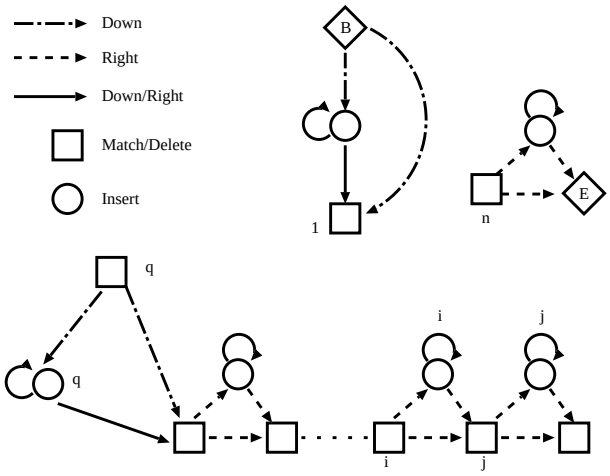


Figure 1: Model sketch of profilePSTMM. Match and delete states are combined for simplicity; they always appear at the same position together. There are also two types of transitions: one that goes down to match children, and one that goes right to match younger siblings. For simplicity, when both down and right transitions start and end at the same states, they are drawn as a single black transition representing both. There are also two special states, BEGIN and END; BEGIN transitions to the first node, and END transitions out of the last (n representing the youngest child leaf state among all leaves in the state model).

The position of each state is fixed and our model has three probability parameters, π, a and b. The initial state probability $\pi[l](= P(z_1 = s_l; \theta)$ is the probability that the state of the root node is s_l. The state transition probability $a[\{s_q, s_l\}, s_m]$ $(= P(z_j^u = s_m | z_p^u = s_q, z_i^u = s_l; \theta))$ is the conditional probability that the state of node z_j is s_m at position m given that the states of the parent (z_p) and immediately elder sibling (z_i) are s_q and s_l, respectively, and the label output probability

$b[s_l, \sigma_h]$ $(= P(o_j^u = \sigma_h | z_j^u = s_l; \theta))$ is the conditional probability that the output of node x_j is σ_h given that the state is s_l at position l. Note that $\sum_l \pi[l] = 1$, $\sum_m a[\{s_q, s_l\}, s_m] = 1$ and $\sum_h b[s_l, \sigma_h] = 1$.

For convenience, we can set the BEGIN state as a MATCH state, and ignore the DELETE state, and we thus allow transitions from the BEGIN state to I_0 and X_1. Correspondingly, the END state collects all transitions from the youngest sibling states.

3.1 Parameter Estimation

These probability parameters are estimated using the same four auxiliary probability parameters as PSTMM, but adjusted for the different types of states fixed at each position. These are the forward probability $F_i(s_q, s_l)$, backward probability $B_i(s_q, s_l)$, upward probability $U_i(s_q)$ and downward probability $D_i(s_q)$. For each probability parameter, i represents node x_i in the given tree, and s_q and s_l are states at position q and l, respectively. s can be either a MATCH (M), an INSERT (I) or a DELETE (X) state. For convenience, for a node j, we usually represent its parent, older sibling, and younger sibling by p, i, and k, respectively.

A straightforward application of the EM algorithm used for estimating and maximizing the likelihood can be applied. Below are the algorithms used for estimation.

3.2 Initialization

We only need to initialize the forward and downward parameters as follows. Note that all values are unlogged values, but the actual implementation uses log calculations.

$$\begin{aligned} F_0(s_q, M_0) &= 0 \\ D_0(M_0) &= 0 \end{aligned}$$

3.3 Forward Probability

$$F_j(s_q, M_l) = \begin{cases} \text{If } \; x_j = x_{\leftarrow}(p) \text{ then } a[\{s_q, -\}, M_l], \\ \text{o.w.} \\ \qquad F_i(s_q, M_k) U_i(M_k) a[\{s_q, M_k\}, M_l] + \\ \qquad F_i(s_q, I_k) U_i(M_k) a[\{s_q, I_k\}, M_l] + \\ \qquad F_i(s_q, X_k) U_i(M_k) a[\{s_q, X_k\}, M_l] \\ \text{where } x_i \text{ is the older brother of } x_j, \; s_k \text{ is the state of } x_i. \end{cases}$$

$$F_j(s_q, I_l) = \begin{cases} \text{If } \; x_j = x_{\leftarrow}(p) \text{ then } a[\{s_q, -\}, I_l], \\ \text{o.w.} \\ \qquad F_i(s_q, M_l) U_i(I_l) a[\{s_q, M_l\}, I_l] + \\ \qquad F_i(s_q, I_l) U_i(I_l) a[\{s_q, I_l\}, I_l] + \\ \qquad F_i(s_q, X_l) U_i(I_l) a[\{s_q, X_l\}, I_l] \\ \text{where } x_i \text{ is the older brother of } x_j. \end{cases}$$

$$F_j(s_q, X_l) = \begin{cases} \text{If } \; x_j = x_{\leftarrow}(p) \text{ then } a[\{s_q, -\}, X_l], \\ \text{o.w.} \\ \qquad F_i(s_q, M_k) a[\{s_q, M_k\}, X_l] + \\ \qquad F_i(s_q, I_k) a[\{s_q, I_k\}, X_l] + \\ \qquad F_i(s_q, X_k) a[\{s_q, X_k\}, X_l] \\ \text{where } x_i \text{ is the older brother of } x_j \text{ and } s_k \text{ is the state of } x_i. \end{cases}$$

3.4 Backward Probability

$$B_i(s_q, s_k) = \begin{cases} \text{If } \; x_i^u = x_{\rightarrow}^u(p) \text{ then } U_i(s_k), \\ \text{o.w. } U_i(M_k)a[\{s_q, s_k\}, M_l]B_j(s_q, M_l)+ \\ \quad U_i(I_k)a[\{s_q, s_k\}, I_k]B_j(s_q, I_k)+ \\ \quad U_i(X_k)a[\{s_q, s_k\}, X_l]B_j(s_q, X_l)) \\ \text{where } x_j \text{ is the younger brother of } x_i \text{ and } s_l \text{ is the state of } x_j. \end{cases}$$

3.5 Upward Probability

$$U_p(s_q) \;\; = \;\; \begin{cases} \text{If } \; C_u(p) = \emptyset \;\; \text{then} \\ \quad \text{if } s_q \text{ is a delete state then } 1 \\ \quad \text{else } b[s_q, o_p], \\ \text{o.w. if } s_q \text{ is a match or insert state then} \\ \quad b[s_q, o_p](F_j(s_q, M_m)B_j(s_q, M_m)+ \\ \qquad\qquad F_j(s_q, I_m)B_j(s_q, I_m)+ \\ \qquad\qquad F_j(s_q, X_m)B_j(s_q, X_m)) \\ \text{else if } s_q \text{ is a delete state then} \\ \qquad\qquad F_j(s_q, M_m)B_j(s_q, M_m)+ \\ \qquad\qquad F_j(s_q, I_m)B_j(s_q, I_m)+ \\ \qquad\qquad F_j(s_q, X_m)B_j(s_q, X_m)) \\ \text{where } s_m \text{ is the state of child } x_j \in C_u(p) \end{cases}$$

3.6 Downward Probability

$$D_j(s_l) \;\; = \;\; \begin{cases} \text{If } j \text{ is the root then } \pi, \\ \text{else if } j = x_{\rightarrow}(p) \text{ then} \\ \quad D_p(M_q)b[M_q, o_p]F_j(M_q, s_l)+ \\ \quad D_p(I_q)b[I_q, o_p]F_j(I_q, s_l)+ \\ \quad D_p(X_q)F_j(X_q, s_l). \\ \text{o.w.} \\ \quad D_p(M_q)b[M_q, o_p]F_j(M_q, s_l) \\ \qquad \{a[\{M_q, s_l\}, M_m]B_k(M_q, M_m) + a[\{M_q, s_l\}, I_l]B_k(M_q, I_l) + a[\{M_q, s_l\}, X_m]B_k(M_q, X_m)\}+ \\ \quad D_p(I_l)b[I_l, o_p]F_j(I_l, s_l) \\ \qquad \{a[\{I_l, s_l\}, M_m]B_k(I_l, M_m) + a[\{I_q, s_l\}, I_l]B_k(I_q, I_l) + a[\{I_q, s_l\}, X_m]B_k(I_q, X_m)\}+ \\ \quad D_p(X_q)F_j(X_q, s_l) \\ \qquad \{a[\{X_q, s_l\}, M_m] \; B_k(X_q, M_m) + a[\{X_q, s_l\}, I_l]B_k(X_q, I_l) + a[\{X_q, s_l\}, X_m]B_k(X_q, X_m)\} \\ \text{where } x_k \text{ is the younger brother of } x_j \text{ and } m \text{ is the younger brother state of } l. \end{cases}$$

3.7 Likelihood Computation

Just as for PSTMM, the likelihood for a given tree can still be calculated by $U_0(M_0)$ as in the following equation:

$$L(T; \theta) \;\; = \;\; \sum_l \sum_s^{M,I,X} \pi[s_l]U_0(s_l)$$

Using these four probability parameters, expectation values are computed in the same manner as when estimating PSTMM.

4 Experimental Motif Profiles

Profile PSTMM was tested on the glycan structures that are believed to be recognized by certain galectins, a protein family known to interact with glycans. Based on a review by Hirabayashi et al.[12],

measurements of galectin affinity were compiled, from which glycan specificity could be estimated. We took the galectins measured in this work to determine the actual glycans that are most likely recognized. These glycans were then profiled using profile PSTMM.

4.1 Data set

The glycan data set consisted of those which had high affinity with galectins. A sample of these structures are listed in Table 1.

Table 1: Glycan structures used for profiling

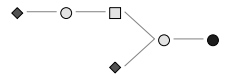
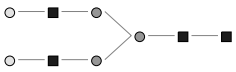
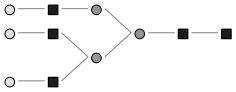
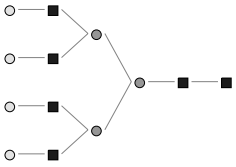
Glycan name	Structure
Galili pentasaccharide	
GD1a	
bi-antennary N-glycan	
tri-antennary N-glycan	
quadra-antennary N-glyca	

Table 2: Legend of monosaccharides, their abbreviations, and their symbols.

Sugar	Abbr.	Sym.
Glucose	Glc	●
Galactose	Gal	○
Mannose	Man	●
N-acetylglucosamine	GlcNAc	■
Fucose	Fuc	▲

When profilePSTMM was run on these structures, disregarding carbon numbers, the model as given in Figure 4.1 resulted. Under a simple cross-validation test, the ROC performance was similar to that of PSTMM, in the 80each node is given next to each monosaccharide in the figure.

The explanation for the mannose at the leaf may be related to the fact that (1) the model contains only one branch and (2) the input structures are not easily alignable. However, it is clear that the ○-■ pattern is learned. It is intriguing that the final node is a ■ as opposed to a sibling ●, but this may be an affect from the other non-N-glycan structures.

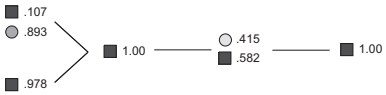


Figure 2: PSTMM model learned from the data in Table 1.

5 Discussion and Future Work

Future work will entail the fine-tuning of this model for certain lectins. Although galectins were studied in this work, these structures are limited in that they are currently characterized by the disaccharide ■/●-○. Thus, the utility of this model would be better demonstrated by other data sets,

which are currently lacking. For example, it is suspected that another group of sialic acid-binding lectins called siglecs recognize a more complex structure at the leaves of glycans. But nothing has been characterized as of yet. Thus, with the increasing data sets of glycans currently available, and with more studies of saccharide affinity, this model can be well used to characterize further structures that may be recognized not only by lectins, but by other biomolecules as well.

Acknowledgments

This work is supported by the Kyoto University 21st Century COE program "Knowledge Information Infrastructure for Genome Science." This is where one acknowledge funding bodies etc. Note that section numbers are not required for Acknowledgments, Appendix or References.

References

[1] K. F. Aoki et al. Application of a new probabilistic model for recognizing complex patterns in glycans. In *Proc. 12th ISMB*, 2004.

[2] K. F. Aoki, H. Mamitsuka, T. Akutsu, and M. Kanehisa. A score matrix to reveal the hidden links in glycans. *Bioinformatics*, 21(8):1457–1463, 2005.

[3] K. F. Aoki, A. Yamaguchi, Y. Okuno, T. Akutsu, N. Ueda, M. Kanehisa, and H. Mamitsuka. Efficient tree-matching methods for accurate carbohydrate database queries. *Genome Informatics*, 14:134–143, 2003.

[4] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, Mhairi Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. Eddie. The Pfam protein families database. *Nucl. Acids Res.*, 32:D138–D141, 2004.

[5] E. Baum and T. Petrie. Statistical inference for probabilistic functions of infinite state Markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

[6] M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. on Sig. Proc.*, 46:886–902, 1998.

[7] M. O. Dayhoff, W. C. Barker, and L. T. Hunt. Establishing homologies in protein sequences. *Methods in Enzymology*, 91:524, 1983.

[8] R. Durbin et al. *Biological sequence analysis.* Cambridge University Press, Cambridge, 1998.

[9] S. R. Eddy. Hidden Markov models. *Current Opinion in Structural Biology*, 6:361–365, 1996.

[10] K. Hashimoto, S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki, and M. Kanehisa. KEGG as a glycome informatics resource. *Glycobiology*, 2005 (in press).

[11] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, 89:10915–10919, 1992.

[12] J. Hirabayashi, T. Hashidate, Y. Arata, N. Nishi, T. Nakamura, M. Hirashima, T. Urashima, T. Oka, M. Futai, W. E. G. Muller, F. Yagi, and K. Kasai. Oligosaccharide specificity of galectins: a search by frontal affinity chromatography. *Biochimica et Biophysica Acta*, 1572:232–254, 2002.

[13] C. Hyeokho and R. Baraniuk. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Trans. on Image Proc.*, 46:886–902, 2001.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[15] N. Ueda, K. F. Aoki-Kinoshita, A. Yamaguchi, T. Akutsu, and H. Mamitsuka. A probabilistic model for mining labeled ordered trees: capturing patterns in carbohydrate sugar chains. *IEEE Transactions on Knowledge and Data Engineering*, 17(8):1051–1064, 2005.

[16] A. Varki et al., editors. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, New York, 1999.