Multiple Tree Alignment with Weights Applied to Carbohydrates to Extract Binding Recognition Patterns

Masae Hosoda, Yukie Akune, and Kiyoko F. Aoki-Kinoshita*

Dept. of Bioinformatics, Faculty of Engineering, Soka University,
1-236 Tangi-machi, Hachioji, Tokyo, Japan 192-8577
{e12d5605,e10d5601,kkiyoko}@soka.ac.jp

Abstract. The purpose of our research is the elucidation of glycan recognition patterns. Glycans are composed of monosaccharides and have complex structures with branches due to the fact that monosaccharides have multiple potential binding positions compared to amino acids. Each monosaccharide can potentially be bound by up to five other monosaccharides, compared to two for any amino acid. Glycans are often bound to proteins and lipids on the cell surface and play important roles in biological processes. Lectins in particular are proteins that recognize and bind to glycans. In general, lectins bind to the terminal monosaccharides of glycans on glycoconjugates. However, it is suggested that some lectins recognize not only terminal monosaccharides, but also internal monosaccharides, possibly influencing the binding affinity. Such analyses are difficult without novel bioinformatics techniques. Thus, in order to better understand the glycan recognition mechanism of such biomolecules, we have implemented a novel algorithm for aligning glycan tree structures, which we provide as a web tool called MCAW (Multiple Carbohydrate Alignment with Weights). From our web tool, we have analyzed several different lectins, and our results could confirm the existence of well-known glycan motifs. Our work can now be used in several other analyses of glycan structures, such as in the development of glycan score matrices as well as in state model determination of probabilistic tree models. Therefore, this work is a fundamental step in glycan pattern analysis to progress glycobiology research.

Keywords: glycomics, glycans, bioinformatics, multiple tree alignment algorithm.

1 Introduction

The purpose of our research is the elucidation of glycan recognition patterns. Glycans are composed of monosaccharides and have complex structures with branches because glycans have more than one binding site compared with the amino acid sequences. Many glycans are bound to proteins and lipids on the

* Corresponding author.

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 49–58, 2012.

cell surface and play important roles in biological processes such as determination of blood type, cellular adhesion, antigen-antibody reactions, and virus infections [15].

Moreover, a family of proteins called lectins are known to recognize and bind to glycans. There are many lectin binding and steric mechanisms involving glycan structures in the control of protein-protein interactions. Many signaling events are also known to be regulated by lectin binding [11].

In general, lectins bind to the terminal monosaccharides of glycans on glycoconjugates. However, it is suggested that some lectins recognize not only terminal monosaccharides, but also internal monosaccharides, possibly influencing the binding affinity [15]. The same may be surmized for other glycan-binding biomolecules such as viruses and bacteria as well. Thus, in order to better understand the glycan recognition mechanism of such biomolecules, glycan arrays were developed [9,2]. Glycan arrays consist of a variety of immobilized glycans on a chip, and are used to assess the binding reaction with fluorescently-labeled proteins, viral glycan binding proteins, antibodies and cells. The Consortium for Functional Glycomics (CFG) [12] has furthermore made their glycan array experimental data available on the web [19]. Therefore, it is now possible to obtain many glycan structures that bind with high affinity to a particular lectin, virus, or bacteria that has been analyzed by the CFG.

With the increasing availability of such glycan binding data, we developed Profile PSTMM [14,4] (probabilistic sibling-dependent tree Markov model) to probabilistically extract recognition patterns of glycans using a probabilistic model similar to HMM [7]. However, the complexity of the algorithm brought forth several challenges that needed to be solved. First, a simplified model with similar probabilistic predictive performance was developed called Ordered Tree Markov Model (OTMM) [10]. However, the development of a "Profile OTMM" model first required the determination of an appropriate state model to learn, which was one of the original challenges of Profile PSTMM.

Therefore, we decided to focus on tree alignment of glycans to obtain glycan profiles that may be recognized by a particular glycan-binding biomolecule. In order to do this, we decided to base our algorithm on ClustalW [13] to progressively build a multiple tree alignment. This was possible due to the existence of a pairwise glycan algorithm which we had previously implemented. Moreover, we also developed a web-based tool on RINGS [1,18] to visualize the resulting glycan profiles on the web such that they could be easily analyzed. We will describe our new multiple glycan alignment algorithm called MCAW (multiple carbohydrate alignment with weights) and briefly introduce some preliminary analytical results.

2 Background

In order for readers to understand our algorithm, we first describe notations that will be used throughout this paper. Glycans are usually classified based on their core structure, which is a particular subtree pattern of monosaccharides

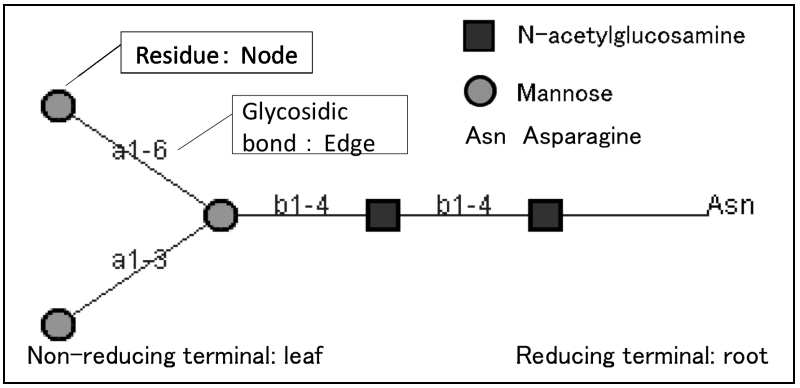

Fig. 1. Example of the *N*-glycan core structure, and description of related terminology

including the root. The glycan in Figure 1 is an *N*-glycan, or *N*-linked glycan structure, which is usually found on asparagine residues on the outer surface of proteins. In mammalian organisms, these glycans on average contain from 10-15 monosaccharides each. As shown in this figure, glycan structures are represented as unordered tree structures, where residues such as monosaccharides and amino acids are nodes, and glycosidic bonds are edges, and the root is usually placed on the right side, branching out towards the left. The right side of the figure is called the reducing terminal, and the left side is the non-reducing terminal end.

2.1 Representation of Glycan Profiles

In order to begin implementing our algorithm, we first needed to define a new text format for representing glycan profiles by expanding the KCF format [5]. PKCF (Profile KCF) contains information indicating alignment order, alignment position, and state (gap, missing, or residue) of each node. The left side of Figure 2 is an example of a glycan alignment of two structures, which is depicted in PKCF format on the right. The ordering of the nodes corresponds to the ordering of the glycans whose names appear in the ENTRY field. Edge information is ordered similarly. PKCF can also represent gaps and missing portions of alignments. In the NODE section, residues are listed by their names, gaps are represented as "-", and missing portions of trees are represented as "0" (the number zero). 'End' in the figure corresponds to "0" in PKCF which indicates that the aligned position does not exist in the indicated glycan structure; that is, it is beyond the terminal residue of the glycan structure.

2.2 KCaM

KCaM [6] is a pairwise glycan alignment algorithm that combines the maximum common subtree and Smith-Waterman local protein sequence alignment

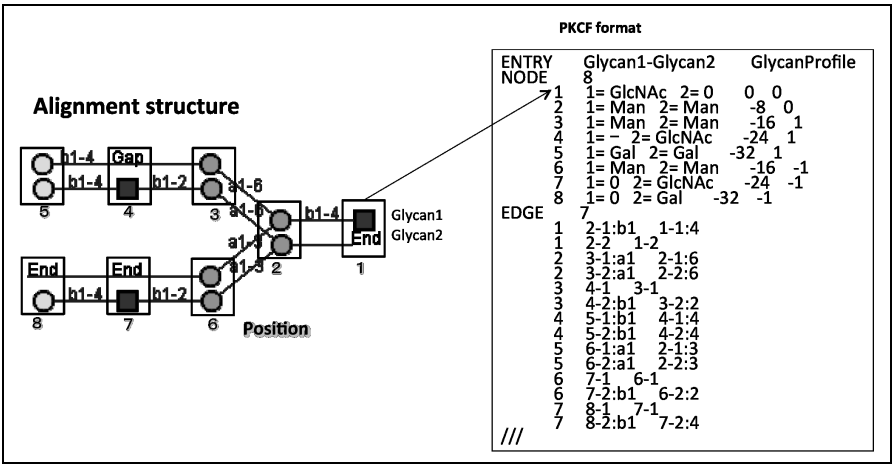

Fig. 2. PKCF format of multiple glycan alignment

algorithms. This algorithm has been preceded by a number of related algorithms, such as the tree edit distance [16] and multiple protein sequence alignment [17]. However, a description of these algorithms are beyond the scope of this manuscript, and the interested reader may refer to the original literature. The dynamic programming algorithm of KCaM is described below.

$$Q[u,v] = \max \left\{ \begin{array}{l} 0, \\ \max_{v_i \epsilon sons(v)} \left\{ Q[u, v_i] + d(v) \right\}, \\ \max_{u_i \epsilon sons(u)} \left\{ Q[u_i, v] + d(u) \right\}, \\ w(u,v) + \max_{\psi \epsilon M(u,v)} \left\{ \sum_{u_i \epsilon sons(u)} Q[u_i, \psi(u_i)] \right\} \end{array} \right\}$$

Here, u and v refer to a particular node u in one tree and node v in the other, and $Q[u, v]$ computes the alignment score of the subtrees rooted at u and v. $sons(x)$ refer to the children of node x, $d(x)$ refers to the gap penalty of deleting node x, $M(u, v)$ refers to the mapping of $sons(u)$ with $sons(v)$, and $w(u, v)$ refers to the score of matching nodes u and v. Thus by computing the scores of all pairs of nodes in the two input glycan structures in breadth-first order, the final score of matching the two glycans can be obtained by finding the pair of nodes with the highest score, and the alignment can be found by backtracking down to the leaves. In most cases, the best score involves the root node of at least one of the input glycans.

With this algorithm, it is possible to align most of the monosaccharides in two glycan structures. However, one must also be careful about the terminal ends. For example, let us assume that the highest scoring node pair are nodes x and y, and that node x is not the root node. Then the parent and further ancestors of node x are not aligned to any other nodes. In this case, we add "missing" nodes to the parent (and possibly grandparent, grand-grandparent, etc.) of node

y to match with the ancestors of x. The same approach is used for nodes at the non-reducing end where the leaf of one glycan is matched to an internal node in the other.

3 Methods

3.1 MCAW Algorithm

Multiple glycan alignment is based on comparing the nodes and edges of glycan profiles, similar to KCaM. In order to distinguish between single glycans and glycan profiles, we use the term *position* to indicate a node of a profile.

The procedure of the MCAW algorithm is as follows.

1. Calculate a distance matrix from the pairwise alignments (using KCaM) of all vs. all of the input glycans.
2. Create a guide tree based on the distance matrix.
3. Calculate weights of each glycan based on distance as indicated from the guide tree.
4. Add glycans to the alignment in the order of the guide tree, adding the most similar glycans first.

We generated the guide tree using the Fitch-Margoliash method [8]. The weights of each glycan structure is computed from the distance to the root of the guide tree. This is performed in order to avoid the inclusion of too many gaps in the alignment by first aligning the most similar glycan structures. Glycans are aligned according to the guidetree. Aligned glycan structures become a single profile structure. However, an alignment may also be performed on a glycan and a profile. Therefore, for the MCAW algorithm, we consider even a single glycan as a profile (containing one glycan), and thus progressively align two profiles with one another with this algorithm. The dynamic programming algorithm of MCAW is described below.

$$Q[u,v] = \max \left\{ \begin{array}{l} 0, \\ \max_{v_i \epsilon sons(v)} \left\{ Q[u, v_i] + d(v) \right\}, \\ \max_{u_i \epsilon sons(u)} \left\{ Q[u_i, v] + d(u) \right\}, \\ \frac{1}{|A||B|} \left\{ \sum_{n=1}^{|A|} \sum_{m=1}^{|B|} w(u_n, v_m) a_n b_m \right\} + \\ \qquad \max_{\psi \epsilon M(u,v)} \left\{ \sum_{u_i \epsilon sons(u)} Q[u_i, \psi(u_i)] \right\} \end{array} \right\}$$

$Q[u, v]$ is the glycan alignment score for positions u and v in profiles A and B, respectively. $|A|$ (resp. $|B|$) is the number of glycans in profile A (resp. B). a_n (resp. b_m) signifies the weight of the nth glycan in profile A (resp. mth glycan in profile B). $w(u_n, v_m)$ is the score between the nodes in positions u and v of the nth and mth glycans of profiles A and B, respectively. $sons(u)$ (resp. $sons(v)$) are the child positions of u (resp. v), and $M(u, v)$ is the mapping between the children of position u and those of position v.

3.2 MCAW Tool

We implemented steps 1 through 3 of the MCAW procedure in Perl, and step 4 was implemented in Java. The Perl program stores the resulting guide tree as a text file including the weights computed for each glycan structure. The Java program then reads in this file to progressively build up the multiple alignment. The resulting alignment is output in PKCF format. CGI-Perl was used to implement the web interface for reading in the input glycan structures and alignment parameters and also to display the results, which is a Java applet that takes the PKCF results from the MCAW program and draws the profile graphically.

4 Results

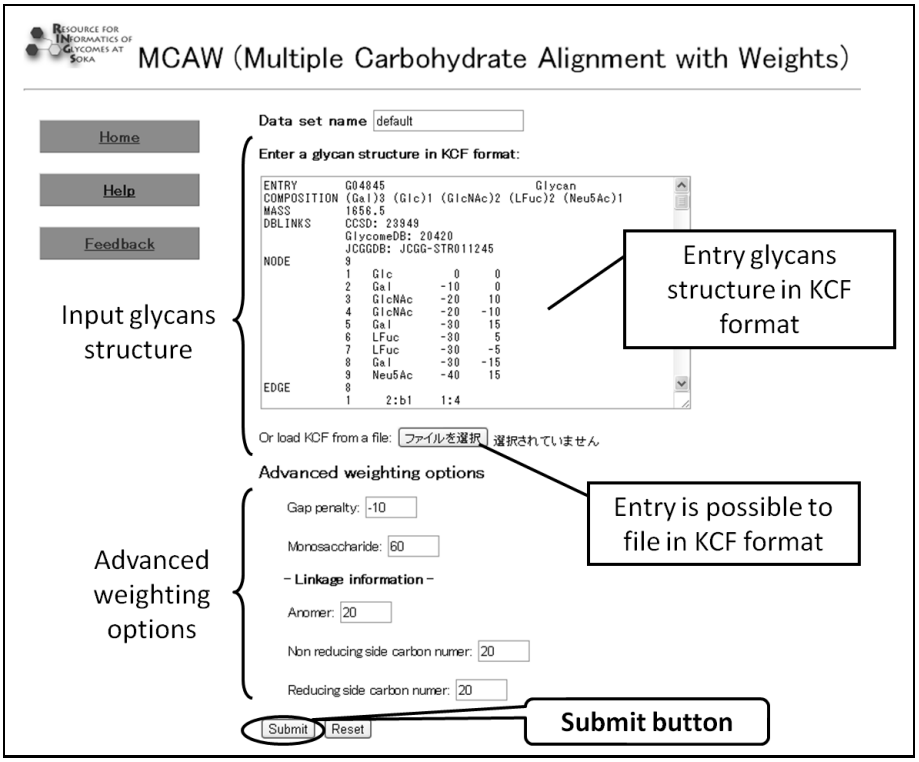

Fig. 3. A snapshot of the input screen for the MCAW tool, where glycan structures are specified in KCF format. Input glycans can also be specified as a file. There are also options to weight the gaps, residues (monosaccharides), and glycosidic bond information (anomers, non-reducing side carbon number and reducing side carbon number) in the "Advanced weighting options" which are provided with default values.

4.1 MCAW Tool

We implemented MCAW as a web tool in RINGS to output a multiple glycan alignment of an input data set of glycans on the web. The URL of the MCAW Tool is http://www.rings.t.soka.ac.jp/cgi-bin/tools/MCAW/mcaw_index.pl. Figure 3 is a snapshot of the input screen, where glycan structures are specified in KCF format.

Input glycans can also be specified as a file. Additionally, it is possible to add weighting options when calculating the alignment score. There are options to weight the gaps, residues (monosaccharides), and glycosidic bond information (anomers, non-reducing side carbon number and reducing side carbon number) in the "Advanced weighting options" in which default values are provided.

4.2 CFG Array Experiment

We have analyzed several data sets of glycan structures from binding affinity data which we obtained from the CFG. Here we present one example of our analyses. We performed an alignment of high-affinity glycan structures (illustrated in Figure 4) from glycan array data of Siglec-F, which belongs to the Siglec family that are well known to bind to sialic acids [15]. As shown in this figure, different glycan structures bound to Siglec-F with various binding affinities, which are

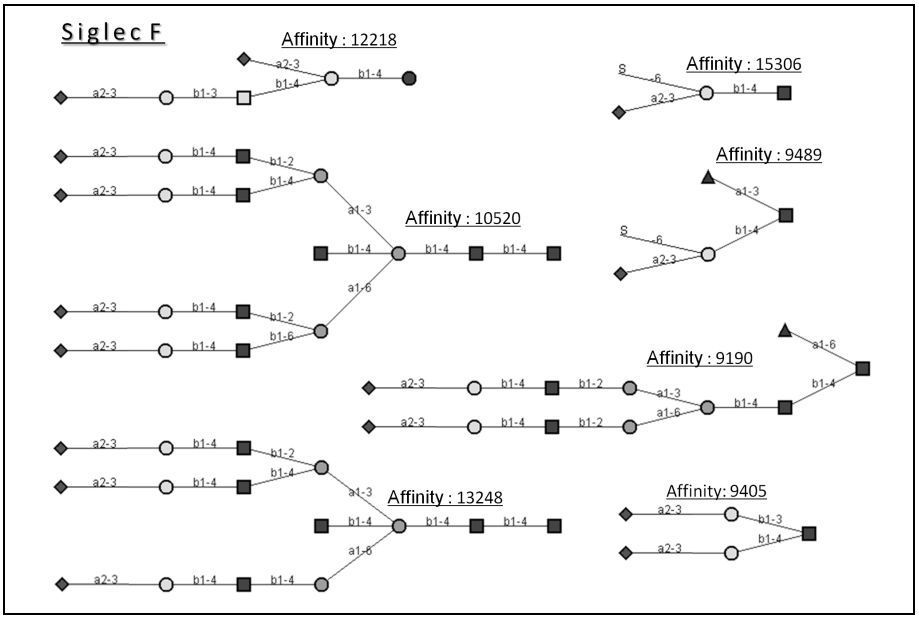

Fig. 4. Input data of Siglec-F with corresponding binding affinity values in relative fluorescence units (RFU) of each glycan structure. Each glycan structure was repeated in the data set according to these values (see text for details).

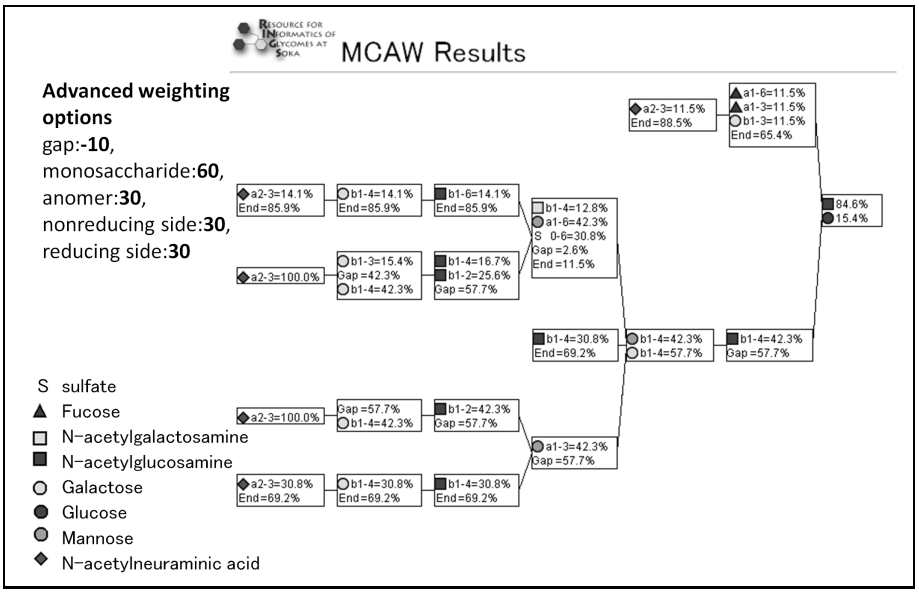

Fig. 5. Resulting glycan profile for glycans with high binding affinity to Siglec-F

provided as average relative fluorescence units (RFU). Therefore, we added multiple copies of the same glycan structures according to binding affinity; those with higher affinities were made to be more prevalent than those with lower affinities. In particular, we divided the binding affinity by 1000 and rounded to the nearest integer. Thus, the structure with affinity 10520 RFU was repeated 11 times, and the structure with affinity 15306 RFU was repeated 15 times. This resulted in a total of 87 glycan structures in the input data set. Moreover, we adjusted the scoring option configurations such that sialic acid residues (NeuAc, purple diamonds in CFG notation) are aligned at the non-reducing end. Our results are illustrated in Figure 5, which also lists the advanced weighting options that we used. It is clear from this figure that not only NeuAc a2-3, but also Gal b1-4 was very highly aligned. Moreover, we find that N-acetyglucosamine comes quite often following this series of NeuAc a2-3 Gal b1-4, forming a sialylated lactosamine structure, which is a well-known glycan motif.

5 Discussion and Conclusions

We have presented the first algorithm to perform multiple glycan alignments as well as a web-based tool so that users can quickly visualize glycan profiles from a group of glycans. In order to weigh input glycan structures based on the strength of binding affinity, weights can be incorporated into the calculation by repeating higher-affinity glycan structures in the input.

We have shown that biologically significant glycan patterns could be extracted from our tool by illustrating that sialic acids as well as other related

monosaccharides could be extracted from our Siglec-F experimental results. By performing further experiments with other Siglecs and various other lectins, more patterns in glycan structure recognition can be obtained. Further work will focus on finding relationships between these patterns and protein sequence/structure. We also plan on studying the most appropriate parameters for the advanced weighting options such that users can select predefined sets of parameters that are most appropriate for their input data.

With the development of this algorithm and tool, we can also compute glycan scoring matrices [3] to analyze similarities in terms of physico-chemical properties of monosaccharides and glycosidic linkages, and further work is now possible for state model determination of Profile PSTMM and Profile OTMM. Therefore, this work is a fundamental step towards glycan recognition analysis to progress glycoinformatics research.

References

1. Akune, Y., Hosoda, M., Kaiya, S., Shinmachi, D., Aoki-Kinoshita, K.F.: The RINGS resource for glycome informatics analysis and data mining on the Web. OMICS 14(4), 475–486 (2010)
2. Alvarez, R.A., Blixt, O.: Identification of ligand specificities for glycan-binding proteins using glycan arrays. Methods Enzymol. 415, 292–310 (2006)
3. Aoki, K.F., Mamitsuka, H., Akutsu, T., Kanehisa, M.: A score matrix to reveal the hidden links in glycans. Bioinformatics 21(8), 1457–1463 (2005)
4. Aoki-Kinoshita, K.F., Ueda, N., Mamitsuka, H., Kanehisa, M.: ProfilePSTMM: capturing tree-structure motifs in carbohydrate sugar chains. Bioinformatics 22, e25–e34 (2006)
5. Aoki-Kinoshita, K.F.: Glycome Informatics: Methods and Applications. CRC Press (2009)
6. Aoki, K.F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., Kanehisa, M.: KCaM (KEGG carbohydrate matcher): a software tool for analyzing the structures of carbohydrate glycans. Nucleic Acids Research 32, 267–272 (2004)
7. Eddy, S.R.: Profile hidden Markov models. Bioinformatics 14, 755–763 (1998)
8. Fitch, W.M., Margoliash, E.: Construction of phylogenetic trees. Science 155, 279–284 (1967)
9. Fukui, S., Feizi, T., Galustian, C., Lawson, A.M., Chai, W.: Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. Nat. Biotechnology 20(10), 1011–1017 (2002)
10. Hashimoto, K., Aoki-Kinoshita, K.F., et al.: A new efficient probabilistic model for mining labeled ordered tree. In: Proc. KDD, pp. 177–186 (2006)
11. Ohtsubo, K., Marth, J.: Glycosylation in cellular mechanisms of health and disease. Cell 126(5), 85–867 (2006)
12. Ramakrishnan, S., Lang, W., Raguram, S., Raman, R., Venkataraman, M., Sasisekharan, R.: Advancing glycomics: Implementation strategies at the Consortium for Functional Glycomics. Glycobiology 16, 82–90 (2006)
13. Thompson, J.D., Higgins, D.G., Gibson, T.J.: Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22(22), 4673–4680 (1994)

14. Ueda, N., Aoki-Kinoshita, K.F., Yamaguchi, A., Akutsu, T., Mamitsuka, H.: 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)
15. Varki, A., et al. (eds.): Essentials of Glycobiology second edition. Cold Spring Harbor Laboratory Press (2009)
16. Bille, P.: A survey on tree edit distance and related problems. Theoretical Computer Science 337(1-3), 217–239 (2005)
17. Shatsky, M., Nussinov, R., Wolfson, H.J.: A method for simultaneous alignment of multiple protein structures. Proteins: Structure, Function, and Bioinformatics 56(1), 143-156, 1097-0134(2004)
18. RINGS, `http://www.rings.t.soka.ac.jp`
19. Consortium for Functional Glycomics, `http://www.functionalglycomics.org`