AN IMPROVED SCORING SCHEME FOR PREDICTING GLYCAN STRUCTURES FROM GENE EXPRESSION DATA

AKITSUGU SUGA suga@kuicr.kyoto-u.ac.jp
YOSHIHIRO YAMANISHI yoshi@kuicr.kyoto-u.ac.jp
KOSUKE HASHIMOTO khashimo@kuicr.kyoto-u.ac.jp
SUSUMU GOTO goto@kuicr.kyoto-u.ac.jp
MINORU KANEHISA kanehisa@kuicr.kyoto-u.ac.jp

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

The prediction of glycan structures from gene expression of glycosyltransferases (GTs) is a challenging new area in computational biology because the biosynthesis of glycan chains is under the control of GT expression. In this paper we developed a new method for predicting glycan structures from gene expression data. There are two main original aspects of the proposed method. First, we proposed to increase the number of predictable glycan structure candidates by estimating missing glycans from a global glycan structure map, which enables us to predict new glycan structures that are not stored in the database. Second, we proposed a more general scoring scheme based on real-valued gene expression intensity rather than converting it into binary information. In the result we applied the proposed method to predicting cancer-specific glycan structures from gene expression profiles for patients of acute lymphocytic leukemia (ALL) and acute myelocytic leukemia (AML). We confirmed that several of the predicted glycan structures successfully correspond to known cancer-specific glycan structures according to the literature, and our method outperforms the previous methods at a statistically significant level.

Keywords: glycosyltransferase; glycan structure; DNA microarray; gene expression.

1. Introduction

Glycans are carbohydrate chains attached to lipids or proteins and are notable as the third type of biological chain next to DNA and proteins, since they have a huge variety of structures and play key roles in a wide variety of biological processes, such as immunity and disease pathogenesis. Pathogens have evolved to exploit host lineage-specific glycans and are constantly shaping the glycomes of their hosts [3]. It is well known that some N-linked glycans are necessary in proper protein folding in eukaryote and specific glycan structures are expressed in carcinoma samples [8]. In addition, some glycans are involved in cell adhesion [1]. Understanding glycan functions requires determining glycan structures, as well as genome and amino-acid sequences.

Some powerful experimental instruments for glycan purification and analysis have been developed and successively improved, such as high-performance liquid chromatography, capillary electrophoresis, mass spectrometry and nuclear magnetic resonance technology [9]. In addition, a variety of computational tools have recently been developed, such as automatic annotation tools for mass spectrometry [4], glycan

structure matching methods [2], glycan composite structure maps [6] and glycan structure prediction methods [7]. However, even with these advances, the experimental determination and computational analysis of glycan structures is still difficult. This is because glycans have more complicated structures than DNA and proteins. While nucleotide and amino acid chains are linear and consist of 4 and 20 elementary components, respectively, glycan chains are branched structures and consist of a number of monosaccharides. In addition, they are multivalent, and linkages have anomeric configurations (alpha and beta).

Recently, Kawano *et al*. developed a method for predicting glycan structures based on microarray gene expression data [7]. The basic idea of their method stems from the fact that glycan biosynthesis is under the control of the expression of glycosyltransferases (GTs). If the expression level of GTs is known in the transcriptome or in the proteome of a given organism, it should be possible to predict the repertoire of glycan structures related with the experimental conditions of expression data such as tissues, organs, and diseases. In their method, the gene expression information of GTs is used in the prediction process. However, there are some limitations in Kawano's method. First, the number of predictable glycans depends on the number of glycans stored in the database, because their prediction is based on a database search. Secondly, the prediction accuracy is far from ideal at practical levels, because their method can treat only binary value information of microarray gene expression data.

In this study, we propose a new method to predict glycan structures from gene expression profiles by improving on the framework of Kawano's method. First, we introduced a strategy of predicting missing glycans, which are not stored in the glycan database, in order to add new glycan structures into our candidate set using the glycan composite structure map. Next, we proposed a new scoring scheme to use the original real-valued expression values, the so-called 'signal', from the microarray data, rather than using binary values, the so-called 'detection', because gene expression levels are observed with real-valued signals in most microarray data in nature. Finally, we applied the proposed method to an experimental gene expression dataset from acute lymphocytic leukemia (ALL) and acute myelocytic leukemia (AML) in order to predict cancer specific glycan structures. As a result, we found that the proposed method outperform Kawano's method in terms of the number of correctly predicted cancer-specific glycan structures.

2. Materials and Methods

2.1. Data

2.1.1. Glycosyltransferase reaction

To construct a GT reaction pattern library, GT genes were obtained from the human genome in the KEGG GENES database based on their annotations [1]. The reaction

specificity of each GT was determined according to the published literature and was characterized by the following three features: (1) the acceptor monosaccharide residue in the glycan chain, (2) the donor monosaccharide residue, and (3) the linkage between them. Fig. 1 shows an illustration of GT-related reactions. 186 GT genes are currently annotated in the human genome. The reaction pattern library consists of nine kinds of monosaccharides: glucose (Glc), galactose (Gal), mannose (Man), N-acetyl-galactosamine (GalNAc), fucose (Fuc), xylose (Xyl) glucuronic acid (GlcA) and N-acetyl-neuraminic acid (sialic acid, Neu5Ac).

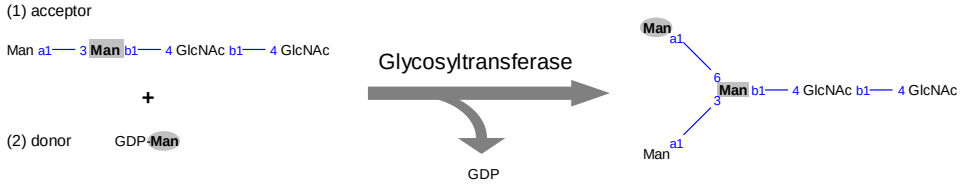


Fig. 1. An example of GT related reactions. This reaction is catalyzed by a GT. The substrates are shown in the left and the product is shown in the right. The acceptor monosaccharide is represented by a gray square and the donor monosaccharide is represented by a gray oval. In this case, the reaction component is 'Man a1-4 Man'. The numbers in the reaction component represent the positions of covalent bonding in the monosaccharides, and 'a' and 'b' represent anomeric configurations 'alpha' and 'beta', respectively.

2.1.2. Glycan structures

All glycan structures were collected from the KEGG GLYCAN database which contains 13022 entries as of this writing. Non-carbohydrate residues in the entries, such as Cer (ceramide), Asn, Ser/Thr, S (sulfate) and P (phosphate) were deleted to obtain glycan entries consisting of only carbohydrates, and duplicate structures were merged. Furthermore, glycan entries, including monosaccharides that are not present in the reaction library, were removed. In this study we focused on the analysis of N-glycans stored in the database, where the number of N-glycans is 1723.

2.1.3. Microarray expression data

Human DNA microarray expression data was obtained from the previous study [5]. The expression dataset consists of 5357 genes for 48 ALL and 21 AML patients. Note that gene expression is measured using real-valued intensity data. We prepared two types of gene expression datasets: ALL and AML datasets. In this study, the gene expression values in ALL and AML datasets are averaged over 48 and 21 patients, respectively. We used these datasets in the prediction process.

The previous method cannot handle real-valued gene expression data [7]. To carry out the previous method, we also prepared a binary type of gene expression dataset in the following manner. The gene expression values were transformed into binary values using a threshold of 1. If the gene expression signal is greater than the threshold, we assign 1 to the corresponding gene. Otherwise, if the gene expression signal is not greater than the

threshold, we assign 0 to the corresponding gene. The binary-valued version is used as predictor data in the previous method.

2.2. Methods for predicting glycan structures

2.2.1. Previous method

The previous method is based on a combination of the co-occurrence score of the GT-related reactions and candidate score for the database search [7]. Here we make a brief review of the previous method.

The co-occurrence score is designed to represent how frequently two GT-related reactions occur simultaneously. Suppose that we have a candidate set consisting of n glycans in the database as $G = \{g_i\}_{i=1}^{n}$. Each glycan g_i is broken down into reaction components, where a reaction component consists of two adjacent monosaccharides and their linkage. For example, the reaction component in Fig. 1 is 'Man a1-4 Man'. Here a set of all possible reaction components is represented by $R = \{r_j\}_{j=1}^{m}$, where m is the number of reaction components. Let us define a reaction pattern vector $\mathbf{x}_j$ for the j-th reaction component r_j as $\mathbf{x}_j = (x_{1j}, x_{2j}, ..., x_{nj})^T$, where x_{ij} is the frequency of r_j in the i-th glycan g_i. Then, we represent all the reaction components by the reaction pattern vector as $\{\mathbf{x}_j\}_{j=1}^{m}$. The co-occurrence score between two reaction components r_j and r_k is obtained by computing the correlation coefficient between the corresponding reaction pattern vectors $\mathbf{x}_j$ and $\mathbf{x}_k$, for example, using Pearson's correlation coefficient defined as follows:

$$C(r_j, r_k) = \frac{\mathrm{cov}(\mathbf{x}_j, \mathbf{x}_k)}{\sqrt{\mathrm{var}(\mathbf{x}_j)\,\mathrm{var}(\mathbf{x}_k)}}.$$

This score represents how likely it is that two GT genes involving r_j and r_k are co-expressed. The prediction of glycan structures is performed by looking for glycans associated with the GT gene expression. Suppose that we are given a gene expression profile $\mathbf{z} = (z_1, z_2, ..., z_p)^T$ from a microarray experiment, where p is the number of GT genes in the microarray experiment. Note that the gene expression profile consists of binary-valued gene expression data in the previous method, where the presence or absence of GT gene transcripts is coded as 1 or 0, respectively. Then, we represent the corresponding reaction components of GT genes in the microarray as $R^* = \{r_l^*\}_{l=1}^{p}$.

Glycans associated with gene expression data are selected from our candidate set G. To select the most appropriate glycan out of the candidate set, Kawano *et al.* proposed to compute the following candidate score for all the candidate glycans g_i $(i = 1,2,...,n)$:

$$S(g_i) = \frac{1}{\sum_{j=1}^{m} I(r_j \in g_i)} \sum_{l=1}^{p} \sum_{j=1}^{m} I(z_l = 1) I(r_j \in g_i) C(r_l^*, r_j),$$

where $C(\cdot,\cdot)$ is the co-occurrence score function, $r_j \in g_i$ means that glycan g_i contains the reaction component r_j, and $I(\cdot)$ is an indicator function which returns 1 if the event is true and returns 0 if the event is false. After the candidate scores are calculated for all glycans stored in the database, high scoring glycan structures are predicted.

2.2.2. Proposed method

There are several limitations and drawbacks in the previous method. First, we cannot predict glycan structures which are not stored in the database, because the prediction procedure is only able to select high scoring glycans from the set of glycans stored in the database. Therefore, the number of predictable glycan candidates is very limited in practical applications. Secondly, we have to transform the original expression values into binary values by taking an appropriate threshold. However, the gene expression is observed with real-valued intensity in most microarray experiments, so the discretization process might lead to a loss of information. Therefore, it is more natural to use the original real-valued intensity rather than converting it into binary information. In this study, we develop a new prediction method in order to overcome the two problems.

First, we propose to increase the number of glycan structure candidates to be predicted. A global glycan composite structure map enables us to suggest missing glycan structures and to estimate possible intermediate structures for the corresponding missing glycans [6]. Fig. 2 shows an illustration of missing glycans and possible intermediate structures. The structures of such intermediate glycans have not been determined experimentally so far, but they are likely to exist in nature. We propose to add such intermediate glycan structures as candidates for predictable glycans. Therefore, our candidate glycan set is composed of not only known glycan structures stored in the

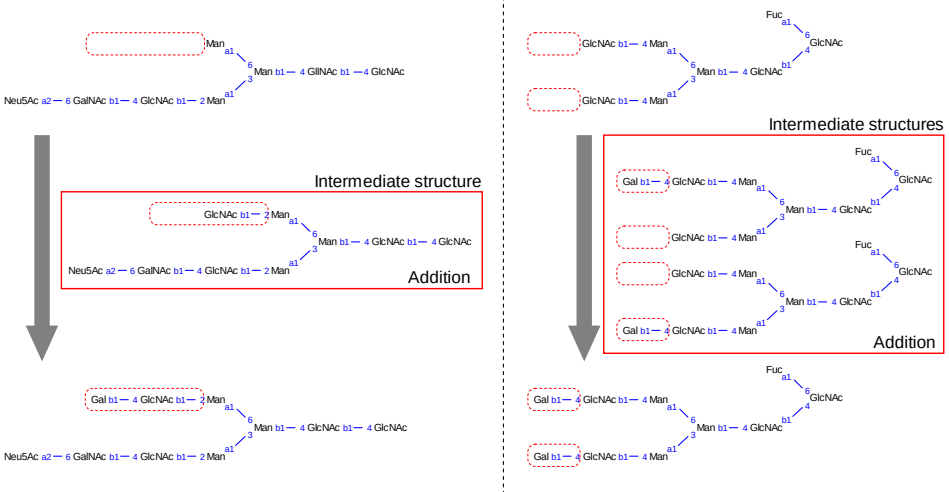


Fig. 2. An illustration of missing glycans and possible intermediate structures. When the intermediate glycans are not stored in the database, they are added into our candidate set for the prediction.

database but also newly identified glycan structures. We represent the number of known glycans and newly identified glycans by n and N, respectively.

Second, we propose a new scoring scheme based on real-valued intensities of microarray gene expression data. Suppose that we are given a GT gene expression profile $\mathbf{z}^* = (z_1^*, z_2^*, ..., z_p^*)^T$ from a microarray experiment. Unlike the previous method, the gene expression profile consists of real-valued expression intensities in this context. For the database search, we propose the following candidate score based on the original real-valued expression intensities for all the candidate glycans g_i $(i = 1,2,..., n+N)$:

$$S(g_i) = \frac{1}{\sum_{j=1}^{m} I(r_j \in g_i)} \sum_{l=1}^{p} \sum_{j=1}^{m} z_l^* I(r_j \in g_i) C(r_l^*, r_j),$$

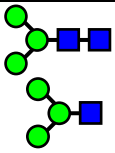
where $C(\cdot,\cdot)$ is the co-occurrence score function, p is the number of GT genes in the microarray experiment, and m is the number of all GT reaction components in a similar manner to the previous method. After the candidate scores are calculated for all glycans in the database and for newly identified glycans, then high scoring glycan structures are predicted.

3. Results

3.1. Amplification of predictable glycan structure candidates

We constructed a global composite structure map for all the glycans in the database, which enabled us to identify many missing glycans [6]. We estimated the intermediate glycan structures for the corresponding missing glycans from the global composite structure map in order to increase the number of predictable glycan structure candidates. As a result, we identified 1291 new intermediate glycans which contain N-glycan core structures. Table 1 shows the numbers of N-glycan structures before and after applying

Table 1. The number of predictable glycan structures before and after applying the global composite structure map.

	Before	After
All N-Glycan	1723	3014
	1164	2094
	218	372
Others	341	548

■ GlcNAc, ● Man

this process. There are two major core structures (80%) shared by N-glycans. Table 1 also shows the numbers of N-glycan structures containing two major N-glycan core structures before and after applying this process. In the following prediction process, we focused on two types of glycans and used not only glycans stored in the database but also newly identified intermediate glycans.

3.2. Prediction of cancer-specific glycan structures

We performed the prediction of glycan structures using two types of gene expression data: ALL and AML dataset. To evaluate the prediction results, we collected known cancer-specific glycan epitopes from the literature [2]. Fig. 3 shows three epitopes which are known to be cancer-specific, referred to as Lewis^a, Lewis^x and sialyl Lewis^x, respectively. Glycans containing the above cancer-specific epitopes are considered to be the gold standard for the performance evaluation below. In this study we are interested in comparing the prediction performance between the previous method and our proposed method.

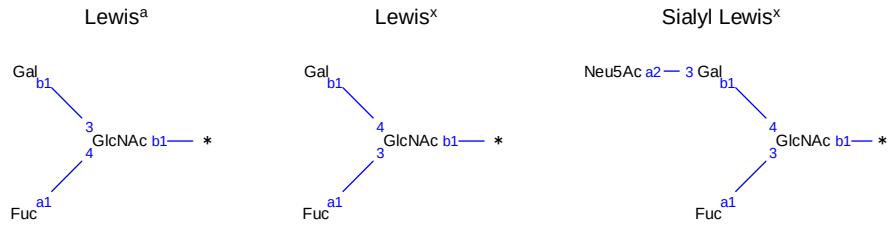


Fig. 3. Cancer-specific glycan epitopes used in this study.

First, we applied both the previous and the proposed prediction methods to ALL gene expression data. If predicted glycans contain cancer-specific substructures, they are considered to be correctly predicted cancer-specific glycans. The left and right panels in Fig. 4 show the index-plots of the sorted scores of the previous and proposed methods,

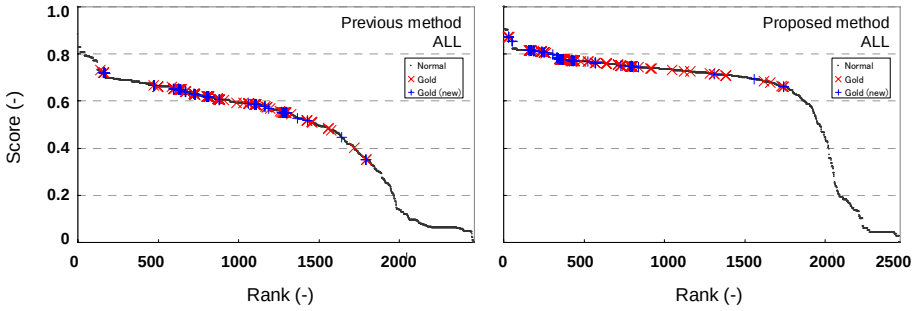


Fig. 4. Performance comparison between the previous method (left) and the proposed method (right): ALL data. Cancer-specific glycans stored in the database are indicated by red crosses, and cancer-specific glycans which are newly identified by our procedure are indicated by blue cross joints.

respectively. In each panel, cancer-specific glycans stored in the database are indicated by red crosses, and cancer-specific glycans that are newly identified by our procedure are indicated by blue cross joints. It seems that high scoring glycans tend to correspond to gold standard cancer-specific glycans. The proposed method seems to catch more information in the prediction process compared to the previous method. This suggests that the use of real-valued gene expression data is meaningful. It also seems that many newly identified glycans have high scores, which also attests to the usefulness of our data amplification procedure.

Next, we applied both prediction methods to AML gene expression data. The left and right panels in Fig. 5 show the index-plots of the sorted scores of the previous and proposed methods, respectively. As in the results for the ALL data, we observed that high scoring glycans correspond to gold standard cancer-specific glycans and many newly identified glycans have high scores. By comparison, the proposed method seems to outperform the previous method in terms of the score ranks of correctly predicted glycan structures. These results also suggest the possibility of predicting glycan structures from observed gene expression patterns in actual applications.

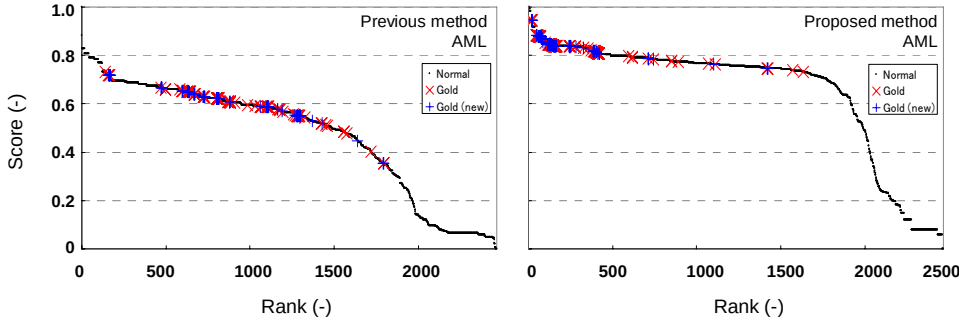


Fig. 5. Performance comparison between the previous method (left) and the proposed method (right): AML data. Cancer-specific glycans stored in the database are indicated by red crosses, and cancer-specific glycans newly identified by our procedure are indicated by blue cross joints.

Finally, we investigated the difference in prediction performance between the previous and proposed methods in more detail. Fig. 6 shows the distribution of the ranks based on candidate scores for gold standard cancer-specific glycans. The left and right panels in the figure correspond to ALL and AML data, respectively. Compared with the previous method, the proposed method seems to assign higher score ranks to gold standard cancer-specific glycans in both cases. We conducted a paired t-test to examine the statistical significance of the performance difference. It is shown that the score rank distributions are different at a statistically significant level, where the p-values in the case of ALL and AML data are 3.2e-29 and 1.6e-48, respectively. These results suggest that our proposed method significantly outperforms the previous method.

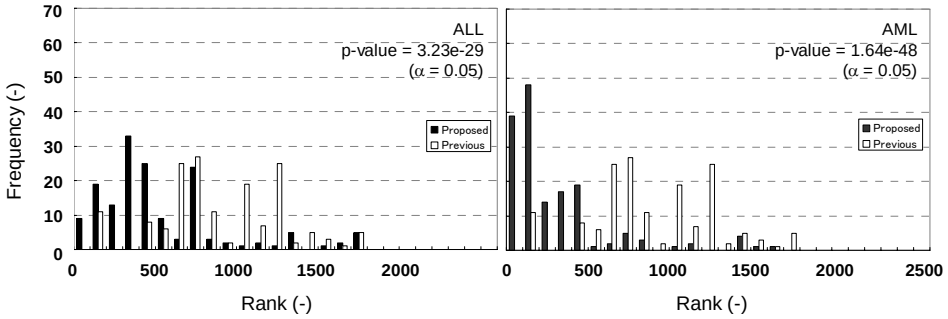


Fig. 6. Distribution of the score ranks for gold standard glycans based on the previous and proposed methods: ALL data (left) and AML data (right). Black bars and white bars correspond to the previous and proposed methods, respectively.

4. Discussion and conclusion

In this paper we developed a new method for predicting glycan structures from gene expression data. There are mainly two original themes in the proposed method. First, we proposed to increase the number of predictable glycan structure candidates by estimating missing glycans from a global glycan composite structure map, which enables us to identify new glycan structures that are not stored in the database. Secondly, we proposed a more general scoring scheme based on real-valued gene expression intensity rather than conversion into binary data. In the result we applied the proposed method to predicting cancer-specific glycan structures from gene expression profiles for ALL and AML leukemia patients. We confirmed that several predicted glycan structures successfully correspond to known cancer-specific glycan structures in the literature, and our method outperforms the previous methods at a statistically significant level.

An advantage of our method is that we can predict new glycan structures that can be synthesized theoretically according to the expression of GTs. Since the experimental determination of glycan structures is still difficult even now, computational prediction of glycan structures might contribute to obtaining new biological findings in glycobiology. In this study we focused on the prediction of cancer-specific N-glycans, but it should be pointed out that our method is applicable to any glycan structure prediction problems for other domains, for example, tissue-, organ-, organism-specific glycans. We are currently working on more comprehensive identification of such domain-specific glycans and application to not only N-glycans but also O-glycans, glycolipids, and several more.

From a technical viewpoint, our scoring scheme is based on the use of expression information about GTs. Recent biotechnology developments have also enabled us to observe the expression level of not only genes but also proteins on a large scale. It will be worth integrating both gene expression and protein expression profiles in the framework of our scoring scheme. In addition, the use of time series data for gene or protein expression might lead to further biological insights.

Acknowledgments

We would like to express our gratitude to Dr. Nelson Hayes for helpful comments and overall improvement of our manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan and the Japan Science and Technology Agency, as well as a bridging grant from the NIH/NIGMS Consortium for Functional Glycomics and a research fellowship for young scientists from the Japan Society for the Promotion of Science. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University and the Human Genome Center, Institute of Medical Science, The University of Tokyo.

References

[1] Akama, T. O., Nakagawa, H., Sugihara, K., Narisawa, S., Ohyama, C., Nishimura, S., O'brien, D. A., Moremen, K. W., Millan, J. L., and Fukuda, M. N., Germ cell survival through carbohydrate-mediated interaction with Sertoli cells, *Science*, 295(5552):124-127, 2002.

[2] Aoki, K. F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M., KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains, *Nucleic Acids Res.*, 32(Web Server issue):W267-272, 2004.

[3] Bishop, J. R. and Gagneux, P., Evolution of carbohydrate antigens--microbial forces shaping host glycomes?, *Glycobiology*, 17(5):23R-34R, 2007.

[4] Goldberg, D., Sutton-Smith, M., Paulson, J., and Dell, A., Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra, *Proteomics*, 5(4):865-875, 2005.

[5] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286(5439):531-537, 1999.

[6] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M., KEGG as a glycome informatics resource, *Glycobiology*, 16(5):63R-70R, 2006.

[7] Kawano, S., Hashimoto, K., Miyama, T., Goto, S., and Kanehisa, M., Prediction of glycan structures from gene expression data based on glycosyltransferase reactions, *Bioinformatics*, 21(21):3976-3982, 2005.

[8] Kim, Y. J. and Varki, A., Perspectives on the significance of altered glycosylation of glycoproteins in cancer, *Glycoconj J.*, 14(5):569-576, 1997.

[9] von der Lieth, C. W., Bohne-Lang, A., Lohmann, K. K., and Frank, M., Bioinformatics for glycomics: status, methods, requirements and perspectives, *Brief. Bioinform.*, 5(2):164-178, 2004.