BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 3976–3982
doi:10.1093/bioinformatics/bti666

Gene expression

Prediction of glycan structures from gene expression data based on glycosyltransferase reactions

Shin Kawano, Kosuke Hashimoto, Takashi Miyama, Susumu Goto and Minoru Kanehisa*

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

Received on June 7, 2005; revised on September 5, 2005; accepted on September 6, 2005
Advance Access publication September 13, 2005

ABSTRACT

Motivation: Glycan chains are synthesized by a combination of several kinds of glycosyltransferases (GTs). Thus, once we know the repertoire of GTs in the genome, in the transcriptome or in the proteome, it should in principle be possible to predict the repertoire of possible glycan structures in an organism or at a specific stage of the cell. Here, we show that a repertoire of glycan structures can be predicted from the set of GTs in the transcriptome. That is, using knowledge about glycan structure characteristics, we can predict glycan structures from incomplete or noisy data such as DNA microarray data.

Results: First, we constructed a reaction pattern library consisting of bond-formation patterns of GT reactions and investigated the co-occurrence frequencies of all reaction patterns in the glycan database. This was followed by the prediction of glycan structures using this library and a co-occurrence score. A penalty score was also implemented in the prediction method. Then we examined the performance of prediction by the leave-one-out cross validation method using individual reaction pattern profiles in the KEGG GLYCAN database as virtual expression profiles. The accuracy of prediction was 81%. Finally, we applied the prediction method to real expression data. Using expression profiles from the human carcinoma cell, glycan structures with sialic acid and sialyl Lewis X epitope were predicted, which corresponded well with experimental results.

Contact: kanehisa@kuicr.kyoto-u.ac.jp

Supplementary information: http://web.kuicr.kyoto-u.ac.jp/~kawano/suppl/bioinfo2005/

1 INTRODUCTION

Glycans, which attach to some lipids and Asn/Ser/Thr residues of proteins, draw attention as the third type of biological chain next to DNA and proteins, since they play key roles in many biological processes such as fertilization (Vo *et al*., 2003), embryogenesis (Schachter *et al*., 2002), immunity (Rudd *et al*., 2001) and diseases (Birkle *et al*., 2003; Brockhausen *et al*., 1998; Hakomori, 2002). Half of the proteins in nature are glycosylated, based on estimates by the analysis of the Swiss-Prot database (Apweiler *et al*., 1999). It is proposed that the glycosylated proteins in the cell membrane and glycosylated lipids form a lipid raft, and that they are involved in signal transduction (Simons and Toomre, 2000). It is well known that some pathogenic bacteria and viruses infect their hosts via glycan–receptor interactions (Cossart and Sansonetti, 2004; Sacks and Kamhawi, 2001). Since glycans may have many functions such as localization signaling, protein stabilization, degradation signaling, signal transduction and immune reaction via glycan–glycan (Bucior and Burger, 2004) and/or glycan–protein interaction (Kogelberg *et al*., 2003; Weis and Drickamer, 1996), it is important to understand glycan functions for understanding life.

To understand glycan functions, determination of their structures (sequences) like DNA and proteins is required. In spite of the improvements in purification and analytical methods for glycans such as high performance liquid chromatography, capillary electrophoresis, mass spectrometry and nuclear magnetic resonance technology (von der Lieth *et al*., 2004), the determination of the glycan structure is still difficult. Glycans have more complicated structures compared to nucleotide and amino acid sequences. 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 various monosaccharides. In addition, they are multivalent, and linkages have anomeric configurations (alpha and beta). These complexities make it difficult to determine glycan structure. Furthermore, the amplification method of glycan is not yet fully established, while DNA and proteins are easy to amplify using polymerase chain reactions and cloning-expression systems, respectively. This means that only a few samples are available for glycan structure analysis. Therefore, a reasonable prediction method for glycan structures is useful for glycomics research.

While the amino acid sequence of proteins is determined by the genetic code and the templates in the genome, the carbohydrate sequence of glycans is determined by the biosynthetic code, which is a specific set of biosynthetic reactions catalyzed by different types of glycosyltransferases (GTs). Each GT catalyzes formation of a glycosidic-bond between the glycan precursor as an accepter and the nucleotide-activated sugar as a donor (Varki *et al*., 1999). Thus, once we know the repertoire of GTs in the genome, in the transcriptome or in the proteome, it should, in principle, be possible to predict the repertoire of possible glycan structures in an organism or at a specific stage of the cell (von der Lieth *et al*., 2004). Here, we construct a reaction pattern library consisting of bond-formation patterns of GT reactions to link genome to glycome, and we predict glycan structures from gene expression profiles.

However, gene set on DNA chips is incomplete, and gene expression data is noisy. To obtain appropriate prediction result, we extract knowledge of glycan structure from the glycan database and apply it to our prediction method. In particular, a co-occurrence frequency

*To whom correspondence should be addressed.

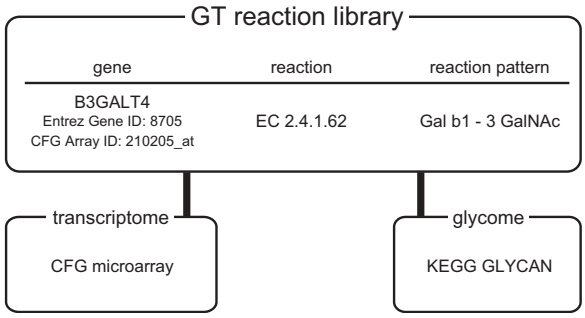

Fig. 1. A schematic diagram of the relationship of data used in this study. The GT reaction library prepared from KEGG GENES and public literature links the transcriptome and the glycome.

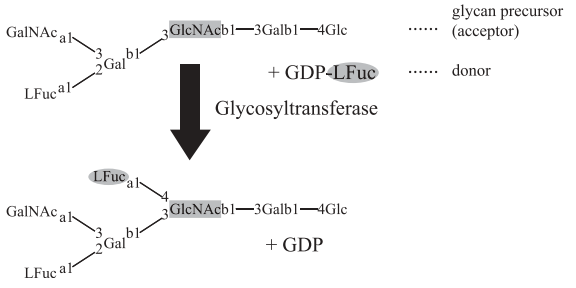

Fig. 2. An example of GT reaction pattern determination. Upper glycan and nucleotide sugar are reaction substrates, and lower glycan and nucleotide are reaction products. This reaction is catalyzed by a GT without a template. The gray square represents the acceptor monosaccharide, and the gray oval represents the donor monosaccharide. The reaction pattern of this example is thus 'LFuc a1-4 GlcNAc'. The number in edge represents a position of covalent bonding in molecule, and 'a' and 'b' in edge represent anomeric configuration, alpha and beta, respectively.

of reaction patterns is calculated from the KEGG GLYCAN database, which is a comprehensive resource encapsulating the latest knowledge of glycans and a part of the KEGG resource containing genomic information and pathways (Kanehisa *et al*., 2004; Hashimoto *et al*., 2005a), and we use it together with the reaction pattern library. First, we evaluate the prediction method using virtual expression data generated from the KEGG GLYCAN database. Then, we apply our method to publicly available DNA microarray expression data and find characteristic glycan structures in a particular cell.

2 DATASET

A relationship of our dataset is shown in Figure 1.

2.1 Glycosyltransferase reactions

In order to construct a GT reaction library, GT genes were obtained from the human genome in the KEGG GENES database (Kanehisa *et al*., 2004) based on their annotations. The reaction specificity of each GT was determined according to the published literature and was characterized by the following three features: (1) acceptor monosaccharide residue in the glycan chain, (2) the donor monosaccharide residue and (3) the linkage between them (Fig. 2). In the human genome, currently, 98 GT genes are annotated, and the reaction pattern library contains 42 reaction patterns (Supplemental data S1). The reaction pattern library consists of nine kinds of monosaccharides: glucose (Glc), galactose (Gal), mannose (Man), *N*-acetyl-glucosamine(GlcNAc),*N*-acetyl-galactosamine(GalNAc), fucose (LFuc), xylose (Xyl), glucuronic acid (GlcA) and *N*-acetyl-neuraminic acid (= sialic acid, Neu5Ac). The number of reaction patterns was less than that of GT genes because the human genome has paralogous genes encoding similar functional proteins. For example, the enzyme reactions of ST8SIA1 (Entrez GeneID: 6489), ST8SIA2 (Entrez GeneID: 8128), ST8SIA3 (Entrez GeneID: 51046), ST8SIA4 (Entrez GeneID: 7903) and ST8SIA5 (Entrez GeneID: 29906) are all the same under our definition of 'Neu5Ac a2-8 Neu5Ac' (see Supplemental data S1).

2.2 Glycan structures

The primary glycan structures were collected from the KEGG GLYCAN database (rel.32, http//www.genome.jp/kegg/glycan/), which contains 10 938 entries. To obtain glycan entries consisting of only carbohydrates, non-carbohydrate residues in the entries, such as Cer (ceramide), Asn, Ser/Thr, S (sulfate) and P (phosphate) were deleted and duplicated structures were merged. Furthermore, glycan entries including monosaccharides that are not present in the reaction library were removed. Finally, our dataset contained 4107 glycan entries.

2.3 Microarray expression data

DNA microarray expression profiles of human were obtained from the Consortium for Functional Glycomics (CFG) (http://www.functionalglycomics.org/static/consortium/organization/sciCores/coree.shtml). CFG has published statistically processed DNA microarray data with expression status, indicated 'P'resent, 'M'arginal and 'A'bsent, from five experiments using human cell such as lung, leukemia and carcinoma cell lines (Supplemental data S2). GT genes were collected from the expression profiles with their annotations and accession numbers. Corresponding to 37 reaction patterns 80 GT genes were mounted on the arrays (Supplemental data S1). When a GT gene was found to be expressed (P) in the majority of the same experimental conditions, it was determined to be positively expressed.

3 RESULTS

3.1 Co-occurrence score

Although glycan structures are diverse, many combinations of reaction patterns are conserved. For example, both 'GlcNAc b1-4 GlcNAc' and 'Man b1-4 GlcNAc' are components of the *N*-glycan core. To obtain an optimal prediction result from DNA microarray expression data, we introduced a co-occurrence score between the reaction patterns calculated from the current glycan database (KEGG GLYCAN). All glycan structures in the database were broken down into reaction pattern components consisting of two adjacent monosaccharides and their linkage, from which a 'reaction pattern matrix' was constructed. Here, the correlation coefficient [S_P, Equation (1)], the Tanimoto coefficient [T_C, Equation (2)] and

the Cosine coefficient [S_C, Equation (3)] between reaction patterns in the matrix were used as the co-occurrence score.

$$S_P(i,j) = \frac{\sum_k (x_i(k) - \bar{x}_i)(x_j(k) - \bar{x}_j)}{\sqrt{\sum_k (x_i(k) - \bar{x}_i)^2 \sum_k (x_j(k) - \bar{x}_j)^2}} \quad (1)$$

$$T_C(i,j) = \frac{\sum_k (x_i(k)x_j(k))}{\sum_k x_i^2(k) + \sum_k x_j^2(k) - \sum_k (x_i(k)x_{j(k)})} \quad (2)$$

$$S_C(i,j) = \frac{\sum_k (x_i(k)x_j(k))}{\sqrt{\sum_k (x_i(k))^2 \sum_k (x_j(k))^2}} \quad (3)$$

where x_i represents reaction pattern vector for reaction pattern i. Thus $x_i(k)$ gives the number of reaction patterns of the k-th element in the reaction vector i. $\bar{x}_i$ is the average of the values in reaction pattern vector i.

Next, we investigated the relationship between the reaction patterns using the hierarchical clustering method. The Ward clustering (Ward, 1963) dendrogram was obtained using the R version 1.9.0 statistics program (http://www.r-project.org/). The co-occurrence score of the reaction patterns appearing higher in the database (top 50 of 302 reaction patterns) was calculated, and the scores were converted into negative values for the clustering. The result of cluster analysis using the scores calculated with the Cosine coefficient is shown in Figure 3. The clustering can be divided into groups according to structural features of glycans. Cluster I corresponds to the reaction pattern in the *N*-glycan core, including 'GlcNAc b1-4 GlcNAc', 'Man b1-4 GlcNAc' and 'LFuc a1-6 GlcNAc' and their terminal/internal structures such as 'Neu5Ac a2-6 Gal' and 'Gal b1-4 GlcNAc'. Cluster II consists of the reaction patterns in the *O*-glycan and glycolipid core such as 'Gal b1-3 GalNAc' (*O*-glycan core 1), 'GlcNAc b1-6 GalNAc' (*O*-glycan core 2) and 'Gal b1-4 Glc' (lactosyl ceramide) and their terminal structures such as 'Neu5Ac a2-6 GlcNAc' and 'LFuc a1-4 GlcNAc'. Cluster III consists of proteoglycan chains, and clusters IV and V contain the components of polysaccharides including xyloglucans and galactomannans, respectively. Thus, reaction patterns conserved in core structures were assembled into the same group, indicating that the coefficient score accurately represents the co-occurrence of reaction patterns conserved in the glycan entries. In the case of using other coefficients, although the topology of the dendrograms was slightly different, reaction patterns in the same core sub-structures also fell into the same groups (see Supplemental data S3 and S4).

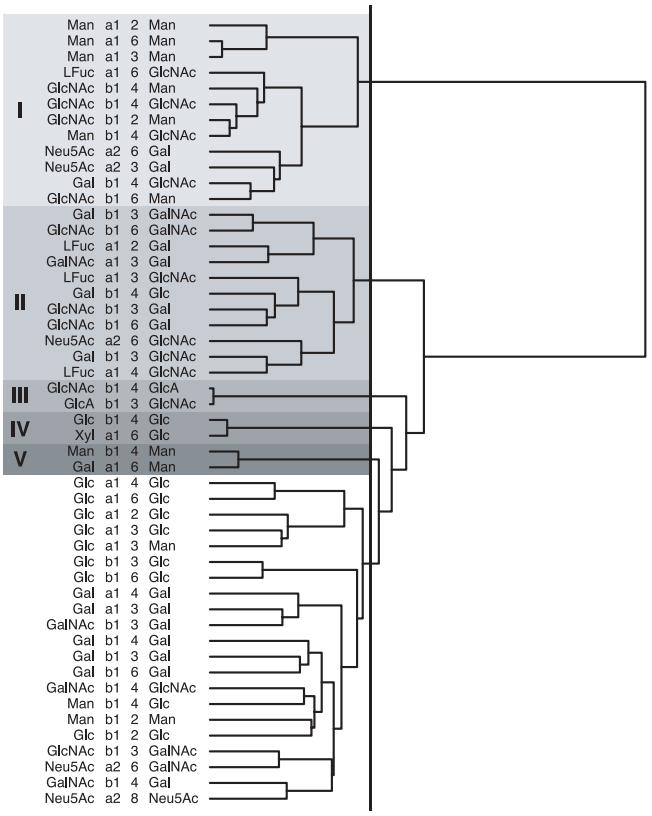

Fig. 3. The cluster dendrogram of the co-occurrence of the reaction patterns in the glycan database. This dendrogram was drawn using the Ward method using the negative value of the cosine coefficient as distance. The clusters were divided into six groups. Cluster I consists of *N*-glycan core reaction patterns, and cluster II consists of *O*-glycan and glycolipid cores. Clusters III, IV and V contain the components of proteoglycan, xyloglucan and galactomannan, respectively. Other dendrograms, drawn using the correlation coefficient and the Tanimoto coefficient, are shown in the supplemental data S3 and S4, respectively.

3.2 Prediction method and evaluation

To evaluate the usefulness of GT reaction library, we developed a prediction method and evaluated its performance. When an expression profile is given as a query, it is converted to an 'expressed reaction pattern list', $Q = \{q_1, q_2, \ldots, q_n\}$. Then, the glycan structures that contain each query reaction pattern are searched for in the database. The score for each candidate glycan, S_E, is calculated as follows:

$$S_E = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} S_B(q_i, b_j) \quad (4)$$

where S_B is the co-occurrence score between the query reaction pattern (q) and the reaction pattern in the candidate glycan structure (b, $B = \{b_1, b_2, \ldots, b_m\}$) and m represents the number of reaction patterns in the candidate glycan structure. In this scoring system, the sum of the co-occurrence score between the query reaction pattern and reaction pattern in the candidate glycan structure is calculated and normalized by the number of reaction patterns in the candidate glycan structure. This operation is repeated against all query reaction patterns.

To test the accuracy of prediction, the reaction pattern list in each glycan structure from the database was used as a virtual expression profile, and the performance of prediction was evaluated with the leave-one-out cross validation method. The evaluation method is summarized in Figure 4. One glycan profile was removed from the 'reaction pattern matrix' (Fig. 4A), and the 'co-occurrence score matrix' (Fig. 4C) was calculated from the 'training matrix' (Fig. 4B) using the three coefficients. In the case that a co-occurrence score could not be calculated, 1 was used between the same reaction patterns, and 0 was used otherwise. The removed glycan profile was converted into a binary 'virtual expression profile' indicating that a pattern exists (1) or not (0) because real expression profiles are generally qualitative rather than quantitative and it is only clear that a gene is expressed (1) or not (0). Furthermore, the virtual

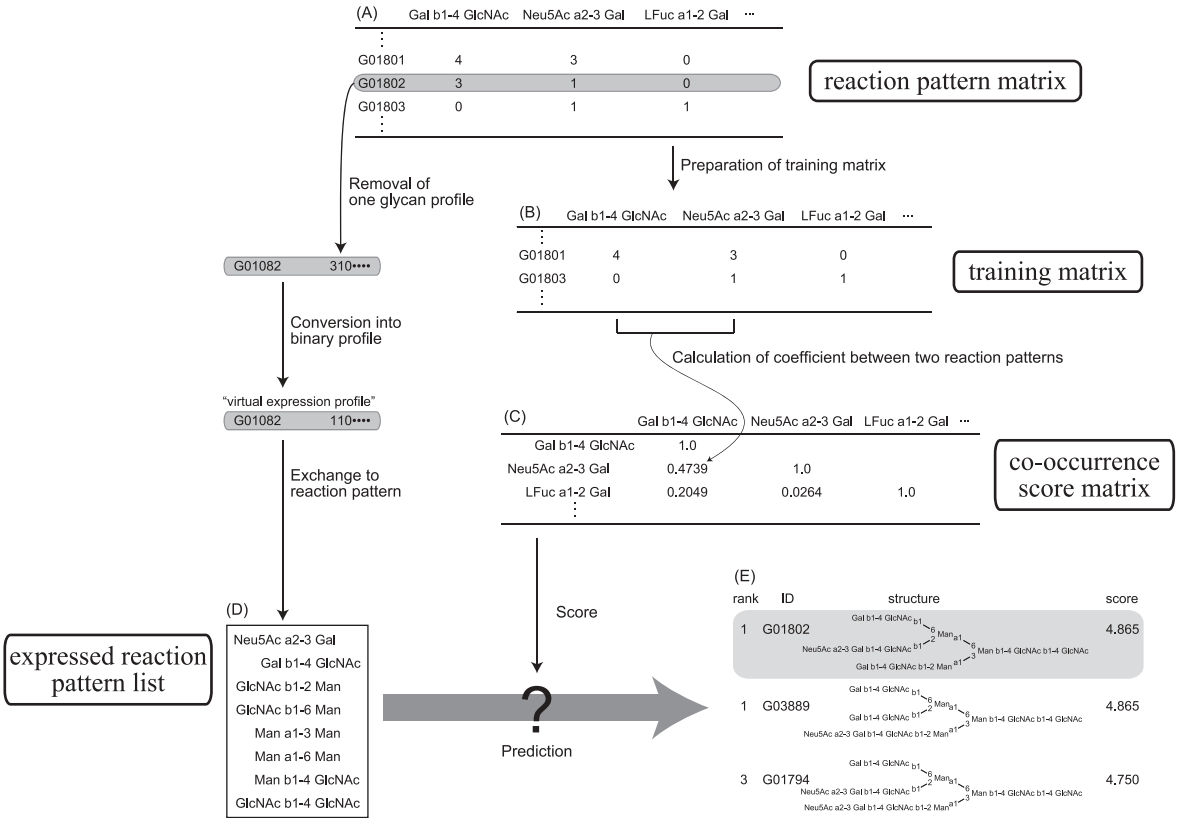

Fig. 4. The scheme of the evaluation method. In this figure, G01802 is used for a query as an example. (**A**) The 'reaction pattern matrix' of the glycan database, which consists of 4107 glycan structures and 302 reaction patterns, is shown. The number in the matrix represents the frequency of appearance of the specific reaction pattern in each glycan. (**B**) A 'training matrix' for evaluation prepared from (A) by removing the query glycan. (**C**) A 'co-occurrence score matrix', which is a diagonal 302×302 matrix, calculated from (B). In this example, the scores are calculated using the cosine coefficient. (**D**) A 'expressed reaction pattern list', which is a list of expressed reaction patterns in the virtual expression profile converted from the query reaction pattern distribution. (**E**) The prediction results from the virtual expression profile with the score matrix (C). Rank, ID number in KEGG GLYCAN, structure and score are shown, and the gray square represents the query glycan. This operation is repeated for all entries in the database.

expression profile was converted into the 'expressed reaction pattern list' (Fig. 4D). S_E was calculated using the 'co-occurrence score matrix' (Fig. 4C) and the 'expressed reaction pattern list' (Fig. 4D), and all predicted glycan structures were sorted by score (Fig. 4E). The rank of the query glycan in the predicted glycans was detected. This operation was iterated for every glycan as the query profile. The accuracy of prediction was defined as the ratio of the query glycan predicted within the given ranks. The random data was generated by shuffling the prediction results and evaluating the rank.

Figure 5 shows the accuracy of the prediction. The prediction results using all co-occurrence scores clearly performed well compared with those using random data. In particular, the cosine coefficient had the best performance; 72% of all entries could be predicted among the top 10 of the prediction results. The fact that the distribution of values in the matrix was not normal and that most of the values was 0 might have resulted in the low accuracy of prediction using the correlation coefficient and the Tanimoto coefficient.

As described above, expression data are binary, so it is possible to determine whether one enzyme is expressed or not. To improve

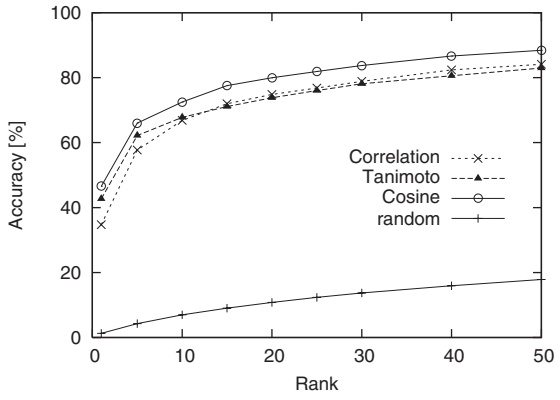

Fig. 5. The accuracy of the prediction with the various scoring systems. The accuracy was calculated as the ratio of the query glycan predicted within the given ranking. The dotted line with x, the dashed line with closed triangle and the solid line with open circle represent the accuracy of the prediction using the correlation coefficient, the Tanimoto coefficient and the cosine coefficient, respectively, as the score. Solid line with plus symbols is the accuracy from random data.

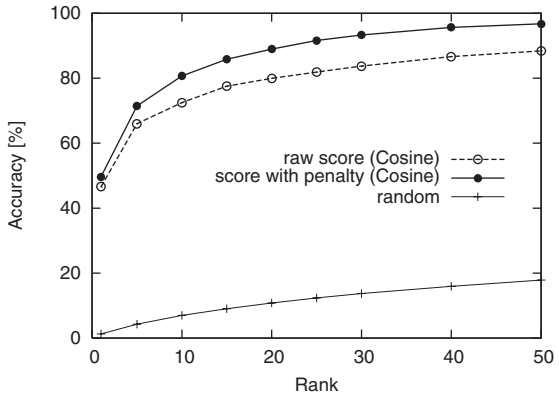

Fig. 6. The accuracy of the prediction with the penalty score. The accuracy was calculated as the ratio of the query glycan predicted within the given ranking. The solid line with closed circle and the dashed line with open circle represent the accuracy using the cosine coefficient with and without the penalty score, respectively. The solid line with plus symbols is the accuracy from random data. The figures using other coefficients are shown in the supplemental data S5.

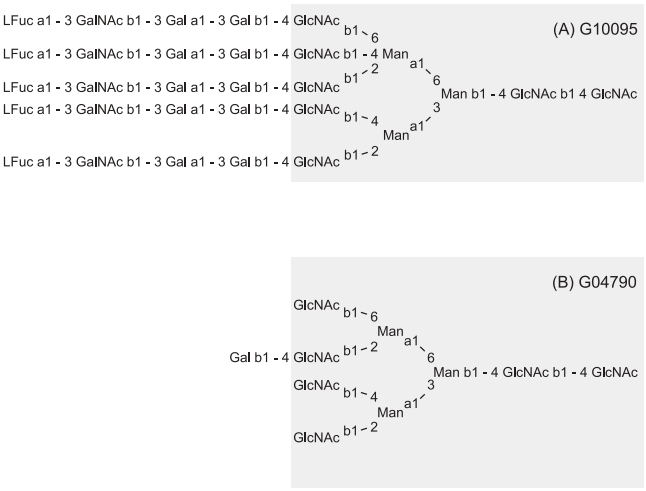

Fig. 7. An example of a falsely predicted glycan. (A) Structure of G10095, which is the query glycan. This glycan was ranked 279th among the 1750 predicted glycans. (B) Structure of G04790, which is the top predicted glycan using the reaction patterns of G10095 as the virtual expression profile. Gray area represents the shared hybrid type *N*-glycan core.

the accuracy of the prediction, we implemented a penalty score into the system.

$$S_E = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\{S_B(q_i, b_j) - p_j\}, \quad (5)$$

where

$$p_j = \begin{cases} 0 & (b_j \in Q) \\ 1 & (b_j \notin Q) \end{cases}$$

where p is the penalty score, which is applied when a reaction pattern in a candidate glycan structure does not exist in the expressed reaction pattern list.

Figure 6 shows the accuracy of the prediction with the penalty score. Implementation of the penalty score improved the accuracy of the prediction using the cosine coefficient as the co-occurrence score. For example, the ratio of the query glycan appearing among the top 10 of the prediction results was improved from 72 to 81%. The performance of the prediction using all other coefficients also improved by the implementation of the penalty score with ~4–5% improvement for every coefficient (Supplemental data S5).

To investigate the possible causes of false positive results, we looked at the individual results. An example of a falsely predicted glycan structure is shown in Figure 7. The virtual expression profile from G10095 (G number is entry ID of the KEGG GLYCAN database) as the query resulted in G04790 as the top score (5.229), while the query glycan was predicted as the 279th (3.513) among 1750 predicted glycans. Although glycans that have a large number of reaction patterns and repeated structures such as *N*-acetyllactosamine (Gal b1-4 GlcNAc) tend to give more false positive, a fundamental core structure was commonly observed. In the example of Figure 7, the hybrid type *N*-glycan core structure ($GlcNAc_n$-Man_3-$GlcNAc_2$) was shared among the top-scoring predicted entries.

While the accuracy of the prediction within the top 10 predicted glycans was >80%, the accuracy of the prediction of the top score was ~50%. In particular, glycans with high-scored repeat sequence tend to falsely predict. However, almost every glycan predicted as the top score was similar in structure to each other (e.g. Fig. 4E). Furthermore, various glycans are expressed in the same cell, regardless of the set of GTs expressed (Mortz *et al*., 1996). Thus, we claim that our prediction results were rather favorable.

3.3 Application to DNA microarray data

Finally, we applied our novel prediction method to real DNA microarray expression data. It is noted that GTs mounted on the array are limited. In the application of our prediction method to real expression profiles, the penalty score may be falsely assigned to practically expressed but unidentified GTs. In this study, however, we presumed that all GTs corresponding to reaction patterns on the microarray have been identified, since homologous enzymes may catalyze the same reaction and can be found easily from genome data using homology search. When a reaction pattern in a candidate glycan was not on the CFG microarray, we did not apply the penalty score to the prediction method. Here, we give examples of the predicted results from the human carcinoma U937 cell line (supplemental data S6). Top 10 of predicted glycans are G05467 (score: 2.627), G05814 (2.627), G04844 (2.568), G11846 (2.549), G04206 (2.442), G04056 (2.421), G00197 (2.419), G10271 (2.419), G10278 (2.419), G00193 (2.413) and G00194 (2.413). Their structures are listed in supplemental data S7, and some representative examples are shown in Figure 8. Briefly, they can be divided into two groups, the hybrid type *N*-glycans and ganglioside. Although they have different core structures, they share the non-reducing end terminal structure, sialic acid (Neu5Ac) and fucose, and the internal *N*-acetyllactosamine structure (Gal b1-4 GlcNAc). Some predicted glycans contain sialyl Lewis X epitope [Neu5Ac a2-3 Gal b1 (LFuc a1-3) 4 GlcNAc]. It has been reported that carcinoma cells overproduce sialyl Lewis X epitope and terminal sialic acid (Kim and Varki, 1997), supporting our prediction results.

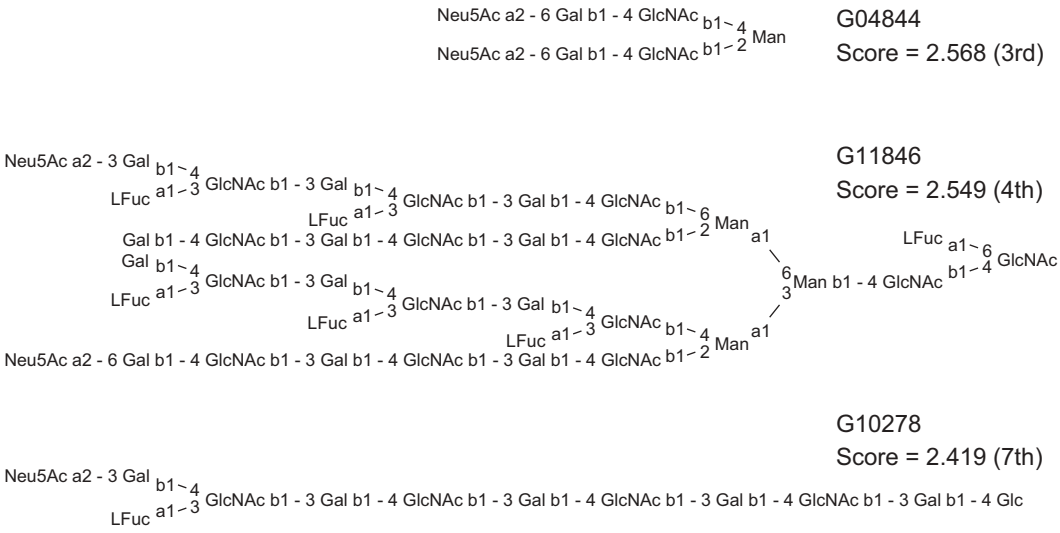

Fig. 8. Representative examples of predicted glycan structures from the gene expression profile of human carcinoma U937 cell line. The glycan structures of top 2 entries are included in G04844. The other glycans among top 10 shared their fundamental structures. All predicted glycan structures among the top 10 are listed on the supplemental data S7. While they have different core structures, they share the terminal sialic acid (Neu5Ac) and fucose, and some glycans contain sialyl Lewis X structure [Neu5Ac a2-3 Gal b1-4 (LFuc a1-3) GlcNAc].

As indicated above, not all GT have been identified and mounted on the array. The glycan database used here contained 302 kinds of reaction patterns, but only 98 GT genes corresponding to 42 reaction patterns (14%) were identified from the human genome. Furthermore, only 80 GT genes corresponding to 37 reaction patterns (12%) were mounted on the CFG microarray. More identification of GTs and update of DNA chips are required for more improvement of our prediction method to apply to real expression data.

4 DISCUSSION

We developed a new method for the prediction of glycan structures from expression profiles. First, we constructed a reaction pattern library of GTs. Based on the annotations of the genome 98 GT genes were collected from the human genome. Although it is believed that GT enzymes generally recognize two or three monosaccharide residues of an acceptor glycan (Nakayama *et al*., 1997; Narimatsu, 2004), only a few enzymes have their exact acceptor residue determined (Kakuda *et al*., 2004), and it has come to be known that some enzymes recognize different acceptor residues depending on their condition (Ramakrishnan *et al*., 2004). Regardless, at the least, the acceptor monosaccharide residue to which the donor monosaccharide is added at the non-reducing end has been elucidated in most enzymes. Thus, we defined the reaction pattern of GT genes as the combination of donor monosaccharide, acceptor monosaccharide and the linkage between them (Fig. 2).

In our reaction pattern rule, many possible reaction patterns can be determined. Although 558 reaction patterns can be determined theoretically from 9 monosaccharides included in the library, only 302 patterns (54.12%) appear in the database. Furthermore, while an enormous number of reaction pattern combinations are possible (2^{558}-1), only 2178 reaction pattern combinations actually appear in the database. This suggests that the structure of glycans is of great variety, but that the actual combination of reaction patterns is limited. Thus, we used co-occurrence between each reaction pattern as a score for prediction when an expression profile is provided. We make note that in our virtual experiment on expression data, the prediction was made for a single glycan structure for simplicity. However, in real experiments the nature of the data is complex in that it is a mixture of multiple glycan structures and responsible gene sets. Therefore, our prediction method should be considered as an upper bound.

In this study, we have challenged for the first time the prediction of glycan structures from genomic data, including GT expression data. This method has the advantage that it applies to any expression data, not only to DNA microarray expression profiles but also to protein expression profiles. Since the determination of glycan structure is still difficult in spite of the recent improvements in analytical technology, it is important to predict glycan structures using reasonable data such as DNA microarray expression profiles and protein array profiles. For further improvement of prediction, introduction of a relative position score between two reaction patterns in the glycan structure may be effective. In addition, the scoring method employed here depends on the training dataset. There may be some data biases, i.e. the glycans that bind to antibodies or lectins with high affinity are easy to purify and to determine their structures. The preparation of the training dataset also admits of improvement. Furthermore, the final glycan product is determined not only by GTs, but also by sugar metabolic enzymes, which synthesize and degrade monosaccharides, transporters, which transfer nucleotide activated sugar from cytoplasm to endoplasmic reticulum and Golgi apparatus, and glycosidases, which trim and degrade the glycan (Yarema and Bertozzi, 2001). Moreover, the conditions of the reaction field, such as pH, temperature, ion concentration, substrate concentration and product concentration, affect the activity of such enzymes (Ramakrishnan *et al*., 2004). Increasing the accuracy of the prediction requires the incorporation of such information of these enzymes and the condition of the reaction field. Furthermore, combining this method

with other data such as mass data and antibody binding profiles will improve the accuracy of prediction.

In addition, the composite structure map (Hashimoto *et al*., 2005b) is created by connecting glycan structures in KEGG GLYCAN like metabolic pathways, and it is possible to map DNA expression data. Combination of our method and the composite structure map not only improves accuracy of prediction but also predicts glycan structures that are absent in KEGG GLYCAN and reveals glycans biosynthetic pathways.

The glycan structure prediction has been improved by combining glycan structural data and knowledge of glycans. For example, our method combined DNA microarray expression data and co-occurrence of reaction pattern to improve the prediction method, and the availability of the known repertoire of glycan structures in a given cell line allowed the validation of the prediction method. Furthermore, other glycan data analyzed by chromatograph, nuclear magnetic resonance and mass spectrometry would also enable better glycan structure prediction. Combination of various type data will permit improvement of glycan structure characterization.

Recently, prognostic expectation and appropriate use of medications adapted for individuals is possible using genotype data (SNPs) and expression data (Efferth and Volm, 2005; Ferrando and Look, 2004). However, it is almost unknown what types of pathways mediate genotype/transcriptotype to phenotype. The prediction of glycan structures from expression profiles illustrates a part of the pathway from transcriptotype to phenotype, and enables higher precision treatment and medication.

ACKNOWLEDGEMENTS

We thank Prof. Hiroshi Mamitsuka, Dr Yasushi Okuno, Dr Kiyoko F. Aoki-Kinoshita and Dr Yoshihiro Yamanishi for critical reading of the manuscript and helpful suggestions. We also thank Masami Hamajima, Tomomi Kamiya, Yuriko Matsuura, Kana Matsumoto, Atsuko Yano, Ami Tanaka and Fujitsu Kyushu System Engineering Ltd for development and maintenance of the KEGG GLYCAN database. Computational time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University. This work was supported by the grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science and the Japan Science and Technology Corporation.

Conflict of Interest: none declared.

REFERENCES

Apweiler,R. *et al*. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.

Birkle,S. *et al*. (2003) Role of tumor-associated gangliosides in cancer progression. *Biochimie*, **85**, 455–463.

Brockhausen,I. *et al*. (1998) Glycoproteins and their relationship to human disease. *Acta Anat. (Basel).*, **161**, 36–78.

Bucior,I. and Burger,M.M. (2004) Carbohydrate–carbohydrate interactions in cell recognition. *Curr. Opin. Struct. Biol*., **14**, 631–637.

Cossart,P. and Sansonetti,P.J. (2004) Bacterial invasion: the paradigms of enteroinvasive pathogens. *Science*, **304**, 242–248.

Efferth,T. and Volm,M. (2005) Pharmacogenetics for individualized cancer chemotherapy. *Pharmacol. Ther*., **107**, 155–176.

Ferrando,A.A. and Look,A.T. (2004) DNA microarrays in the diagnosis and management of acute lymphoblastic leukemia. *Int. J. Hematol*., **80**, 395–400.

Hakomori,S. (2002) Glycosylation defining cancer malignancy: new wine in an old bottle. *Proc. Natl Acad. Sci. USA*, **99**, 10231–10233.

Hashimoto,K. *et al*. (2005a) KEGG as a glycome informatics resource. *Glycobiol*., in press.

Hashimoto,K. *et al*. (2005b) A global representation of the carbohydrate structures: a tool for the analysis of glycan. *Genome Informatics*, **16**, 214–222.

Kakuda,S. *et al*. (2004) Structural basis for acceptor substrate recognition of a human glucuronyltransferase, GlcAT-P, an enzyme critical in the biosynthesis of the carbohydrate epitope HNK-1. *J. Biol. Chem*., **279**, 22693–22703.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.

Kim,Y.J. and Varki,A. (1997) Perspectives on the significance of altered glycosylation of glycoproteins in cancer. *Glycoconj. J*., **14**, 569–576.

Kogelberg,H. *et al*. (2003) New structural insights into carbohydrate–protein interactions from NMR spectroscopy. *Curr. Opin. Struct. Biol*., **13**, 646–653.

Mortz,E. *et al*. (1996) Does matrix-assisted laser desorption/ionization mass spectrometry allow analysis of carbohydrate heterogeneity in glycoproteins? A study of natural human interferon-gamma *J. Mass Spectrom*., **31**, 1109–1118.

Nakayama,K. *et al*. (1997) Substrate specificity of alpha-1,6-mannosyltransferase that initiates *N*-linked mannose outer chain elongation in *Saccharomyces cerevisiae*. *FEBS Lett*., **412**, 547–550.

Narimatsu,H. (2004) Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj. J*., **21**, 17–21.

Ramakrishnan,B. *et al*. (2004) Structure and catalytic cycle of beta-1,4-galactosyltransferase. *Curr. Opin. Struct. Biol*., **14**, 593–600.

Rudd,P.M. *et al*. (2001) Glycosylation and the immune system. *Science*, **291**, 2370–2376.

Sacks,D. and Kamhawi,S. (2001) Molecular aspects of parasite-vector and vector-host interactions in leishmaniasis. *Annu. Rev. Microbiol*., **55**, 453–483.

Schachter,H. *et al*. (2002) Functional post-translational proteomics approach to study the role of *N*-glycans in the development of *Caenorhabditis elegans*. *Biochem. Soc. Symp*., **69**, 1–21.

Simons,K. and Toomre,D. (2000) Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell. Biol*., **1**, 31–39.

Varki,A., Cummings,R., Esko,J., Freeze,H., Hart,G. and Marth,J. (1999) *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, NY.

Vo,L.H. *et al*. (2003) Identification of the ZPC oligosaccharide ligand involved in sperm binding and the glycan structures of *Xenopus laevis* vitelline envelope glycoproteins. *Biol. Reprod*., **69**, 1822–1830.

von der Lieth,C.W. *et al*. (2004) Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief. Bioinform*., **5**, 164–178.

Ward,J.H. (1963) Hierarchical Grouping to optimize an objective function. *J. Am. Stat. Assoc*., **58**, 236–244.

Weis,W.I. and Drickamer,K. (1996) Structural basis of lectin-carbohydrate recognition. *Annu. Rev. Biochem*., **65**, 441–473.

Yarema,K.J. and Bertozzi,C.R. (2001) Characterizing glycosylation pathways. *Genome Biol*., **2**, REVIEWS0004.