Extraction of Species-Specific Glycan Substructures

Yoshiyuki Hizukuri yosh@scl.kyoto-u.ac.jp
Yoshihiro Yamanishi yoshi@kuicr.kyoto-u.ac.jp
Kosuke Hashimoto khashimo@scl.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

Abstract

Glycans, which are carbohydrate sugar chains attached to some lipids or proteins, have a huge variety of structures and play a key role in cell communication, protein interaction and immunity. The availability of a number of glycan structures stored in the KEGG/GLYCAN database makes it possible for us to conduct a large-scale comparative research of glycans. In this paper, we present a novel approach to compare glycan structures and extract characteristic glycan substructures of certain organisms. In the algorithm we developed a new similarity measure of glycan structures taking into account of several biological aspects of glycan synthesis and glycosyltransferases, and we confirmed the validity of our similarity measure by conducting experiments on its ability to classify glycans between organisms in the framework of a support vector machine. Finally, our method successfully extracted a set of candidates of substructrues which are characteristic to human, rat, mouse, bovine, pig, chicken, yeast, wheat and sycamore, respectively. We confirmed that the characteristic substructures extracted by our method correspond to the substructures which are known as the species-specific sugar chain of γ-glutamyltranspeptidases in the kidney.

Keywords: glycobiology, glycan structure, support vector machine, structure matching

1 Introduction

Glycobiology is the study of the glycans that are carbohydrate sugar chains attached to some lipids or proteins. Glycans have a huge variety of structures and play a key role in cell communication, protein interaction and immunity. For example, glycans can serve as intermediates in generating energy, as signaling molecules, or as structural components. The structural roles of glycans become particularly important in constructing complex multicellular organs and organism, which require interactions of cells with one another and with the surrounding matrix [1, 16]. In the post-genome era, comparative analysis of glycans is an important issue such as comparative analysis of DNA or proteins. There is therefore an incentive to develop methods for comparative analysis of glycans, which is expected to be used for organism and tissue classifications of glycans.

In the traditional sequence-comparison methods, the problem is reduced to a string matching problem of nucleotide or amino acid sequences. However, these methods can not be directly applied to glycan comparison, because the mechanisms and structure of glycans are completely different from those of DNA or protein [1]. There are mainly the following three differences: First, DNA/protein have template sequences to synthesize the primary structures and their sequences are uniquely determined by the template structures. On the other hand, glycan does not have such a template structure to determine the sequence and its structure is mainly determined by the substrate specificity of glycosyltransferases. Second, DNA/protein are synthesized in a single compartment such as the nucleus and cytoplasm, while glycans are synthesized in multiple compartments of the cell and their localization varies from cytoplasm to endoplasmic reticulum and Golgi body. Third, DNA/protein are synthesized

as linear structures, while glycans are synthesized as a tree structure with various types of branced linkages. The nodes of the tree correspond to monosaccharides. The edges contain information on the linkage position and anomality, because for glycosyl bond, there are six possible hydroxy groups and two possible anomality α or β. Such linkage patterns give glycans a huge variety of structure patterns with small sequence size. Therefore, it is necessary to take into account of these biological aspects when in comparing glycan structures.

The complex mechanism of glycan synthesis has been a factor of the difficulty in studying the glycan structures, but recent advances of NMR and mass spectrometry technologies have made it possible to determine a growing number of glycan structures experimentally [7]. The CarbBank [8, 19] is a well-known database of carbohydrates, and its contents are taken from many literature reports on glycan research, but the CarbBank has not been updated and maintained in recent years. The KEGG/GLYCAN [10, 12] database has been developed by inheriting and refining the structure data stored in the CarbBank. The availability of a number of glycan structures in the KEGG/GLYCAN database enables us to conduct a large scale comparative research of glycans [2]. In this study, we consider conducting a comparative analysis of glycans in order to understand the variability of the substructures across species and to investigate the relationship between the substructures and biological functions. To the best of our knowledge, there have been no reports on the comprehensive analysis of species-specific glycan structures from computational viewpoints.

In this paper, we present a novel approach to comparing glycan structures and extracting characteristic glycan substructures to certain species. First of all, we develop a new measure for evaluating the similarity of glycan structures in order to classify them from the species-specific point of view. The originality of our similarity measure is that we take into account several biological parameters on glycan and glycosyltransferases such as position variability of saccharides and properties of glycosyltransferase interactions. We confirm the validity of our similarity measure on its ability to classify glycans between species in the framework of a support vector machine (SVM) [11, 17]. Finally, we extract characteristic glycan substructures of certain species using the result of the classification of glycan structures. The results shows that our approach successfully extracts a set of candidates of substructrues which are characteristic to human, rat, mouse, bovine, pig, chicken, yeast, wheat and sycamore, respectively. Several reports have already been made concerning the comparative study of the sugar patterns of the glycoproteins in the same tissue across different animals [13]. Comparing our results with the previous reports, we show that characteristic substructures extracted by our method correspond to the substructures which are identified in the previous reports.

2 Materials

2.1 Glycan Structures

All carbohydrate structures were collected from the KEGG/GLYCAN database [18] and the annotations of the biological sources (BS) were collected from the CarbBank/CCSD (Complex Carbohydrate Structure Database) [19]. The CarbBank/CCSD and KEGG/GLYCAN can be linked by the CCSD ID number. We identified glycan structures with specific biological sources according to the annotation information in the BS fields of the CarbBank/CCSD. Using the CCSD ID number in the BS fields, we collected the carbohydrate structures from the KEGG/GLYCAN. Glycans form a tree structure and mainly consist of eight types of monosaccharides, glucose (Glc), galactose (Gal), mannose (Man), fucose (Fuc), Xylose (Xyl), N-Acetyl-glucosamine (GlcNac), N-acetyl-galactosamine (GalNac), and N-acetyl neuramic acid (Neu5Ac). The structure data in the KEGG/GLYCAN contains several modifications and other biomolecules such as lipids and amino acids. In this study, we deleted almost all modifications except for phosphorous (denoted by P) and sulfur (denoted by S) which do not bind to the monosaccaride at the root. Next, we constructed a data set which consists of unique structures. The total number of glycans in our dataset is 1,377 including 166 glycan structures conserved across

more than one organism. It contains the glycan structures from nine species: human, rat, mouse, bovine, pig, chicken, yeast, wheat and sycamore.

2.2 Monosaccharide Composition

The numbers of glycans from human, rat, mouse, bovine, pig, chicken, yeast, wheat, and sycamore are 453, 176, 115, 291, 232, 89, 100, 44, and 43, respectively. Table 1 shows the compositions of monosaccharides in each organism. The compositions of yeast and plants are completely different from those of animals, respectively. For example, mannose (Man) accounts for more than 80% of the monossacarides in the yeast glycans, which accounts for the fact that 14% of the dry weight of yeast is Man [14]. Wheat contains xylose (Xyl) and arabinose (Ara) at high rates. These monossacarides are the components of cell walls. Animals have a comparatively similar composition each other. To investigate the relationship between carbohydrates and species, we applied the principal component analysis (PCA) to the composition data. Figure 1 shows a scatter plot of the 1st and 2nd loadings of the principal components, which represents the relationship between species. It clearly shows that plants and yeast are located far from the group of animals. Figure 2 shows a scatter plot of the 1-st and 2-nd scores of the principal components, which represents the relationship between carbohydrates. It is found that the compositions of Man, Xyl and Ara are important characteristics that distinguish the difference between yeast, sycamore and wheat. Animals form a cluster in the PCA analysis, therefore have similar glycan structures, compared with non-animals.

Table 1: Composition of monosaccharide of each organism.

species	Glc	Gal	Man	Fuc	Xyl	GlcNAc	GalNAc	Neu	Ara	others
human	3.5 %	25.8 %	15.2 %	9.5 %	0.1 %	31.0 %	4.8 %	8.4 %	0.0 %	1.7 %
rat	6.5 %	30.2 %	11.9 %	0.0 %	0.0 %	24.0 %	8.5 %	7.9 %	0.0 %	11.0 %
mouse	2.9 %	19.2 %	27.8 %	5.0 %	1.2 %	33.7 %	2.8 %	3.6 %	0.0 %	3.8 %
bovine	2.4 %	22.1 %	19.8 %	3.6 %	0.4 %	29.1 %	3.8 %	7.3 %	0.0 %	11.5 %
pig	6.1 %	17.6 %	17.2 %	5.9 %	0.2 %	24.0 %	5.5 %	2.7 %	0.0 %	20.8 %
chicken	2.5 %	14.0 %	29.4 %	0.9 %	0.0 %	41.7 %	3.9 %	7.2 %	0.0 %	0.4 %
yeast	8.4 %	0.0 %	80.1 %	0.0 %	0.0 %	10.8 %	0.0 %	0.0 %	0.0 %	0.7 %
wheat	6.7 %	0.7 %	1.8 %	0.0 %	55.0 %	0.7 %	0.0 %	0.4 %	30.3 %	4.4 %
sycamore	32.1 %	12.3 %	6.5 %	9.6 %	27.6 %	6.5 %	0.0 %	0.0 %	2.5 %	2.9 %

3 Methods

3.1 Similarity Score of Glycans

We develop a similarity measure between glycans based on the biological knowledge about them and glycosyltransferase. Glycans have the branched tree structures and we refer to the 1st monossacharide as root and the tail as leaf. From a biological viewpoint, the glycan and associated glycosyltransferase have the following properties [1]:

1. Glycosyltransferase physically interacts with about three monosaccharides at the leaves.

2. The variability of the sugars near the leaf (variable part) is larger than those near the root (core part).

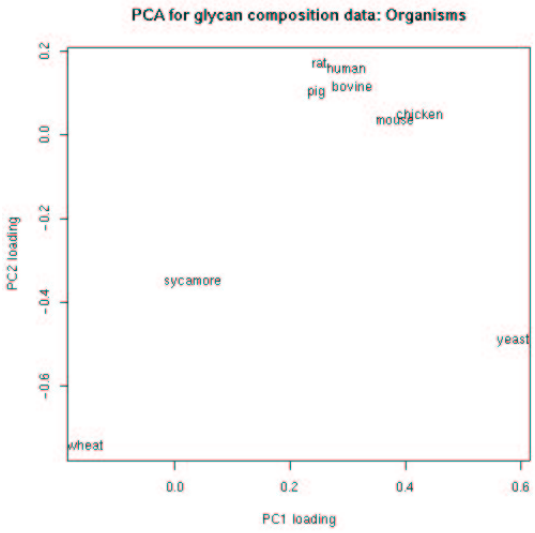

Figure 1: PCA for composition data: Relationship between species.

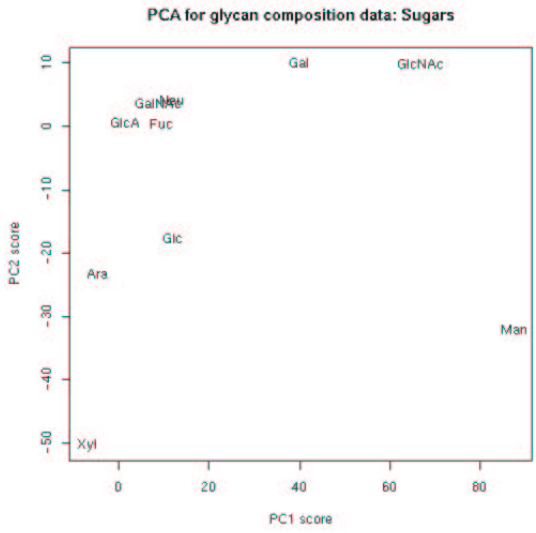

Figure 2: PCA for composition data: Relationship between monosaccharides.

3. The localization of glycosyltransferase varies in the cell, and the sugar chains are constructed in different cellular compartments.

4. Glycosyltransferase recognizes both a monosaccharide binding to the amino acids or lipids at the root, and a specific pattern in the variable part of a glycan.

It is desirable that we develop a similarity measure of glycans taking into account of the above biological properties.

Suppose that we have two glycans $\mathbf{x}$ and $\mathbf{y}$. At the first stage, the glycans $\mathbf{x}$ and $\mathbf{y}$ are decomposed into sets of 3-mers. This operation is motivated by the property 1. As a result, we obtain sets of substructures as $\{\mathbf{x} : x_1, x_2, \cdots, x_{n_x}\}$ and $\{\mathbf{y} : y_1, y_2, \cdots, y_{n_y}\}$, where n_x (resp. n_y) is the number of 3-mers of $\mathbf{x}$ (resp. $\mathbf{y}$). Figure 3 illustrates the decomposition of the glycans $\mathbf{x}$ and $\mathbf{y}$. Considering the match and mismatch of the substructures between glycans $\mathbf{x}$ and $\mathbf{y}$, a straightforward similarity is defined as

$$Sim(\mathbf{x}, \mathbf{y}) = \frac{\#\{\mathbf{x} \cap \mathbf{y}\}}{\#\{\mathbf{x} \cup \mathbf{y}\}}, \tag{1}$$

where the numerator indicates the number of common substructures between $\mathbf{x}$ and $\mathbf{y}$ and the denominator indicates the number of unique substructures of $\mathbf{x}$ or $\mathbf{y}$. Let h_x (resp. h_y) be the layer on which the monosaccharide of $\mathbf{x}$ (resp. $\mathbf{y}$) is located near the root. Note that the root is assumed to be located on the 1st layer. Also, we set $d = |h_x - h_y|$ and $h = \max(h_x, h_y)$. To take into account of the biological properties 2, 3, and 4, we conduct the following weighted summation in counting the common substructures:

$$Sim(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{p} w_k}{q}, \tag{2}$$

where p is the number of the common substructures between $\mathbf{x}$ and $\mathbf{y}$, and q is the number of the unique substructures of $\mathbf{x}$ or $\mathbf{y}$. The w_k is designed as

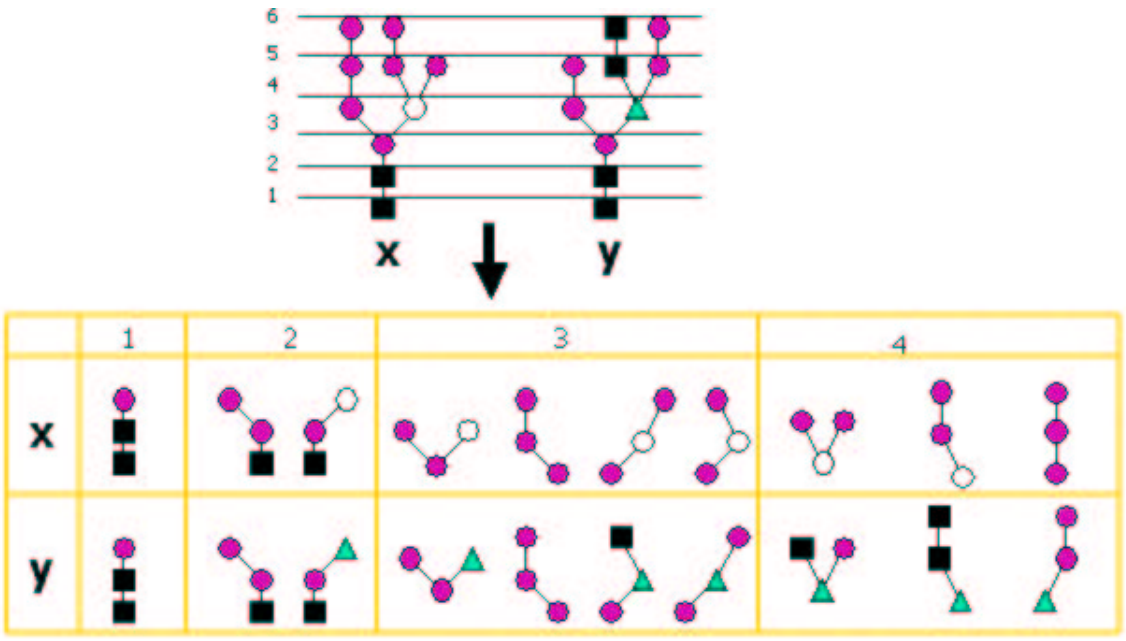

Figure 3: Decomposition of the glycan structures.

$$w_k = \begin{cases} 1 - \exp(-\alpha h + \beta d) & \text{if } h > 1, \\ 1 & \text{if } h = 1 \end{cases} \tag{3}$$

where α and β are positive constants. The weight w plays the following roles: First, the larger the distance between the substructure and the root, the larger the value of the similarity, which reflects the property 2. Second, the larger the distance between the layer h_x and h_y, the smaller the value of the similarity, which reflects the property 3. Third, if the common substructure is found on the root, the weight is set to 1, which reflects the property 4.

3.2 Support Vector Machine

To classify glycans between species based on the glycan structures, we use the support vector machine (SVM) [11, 17] in this study. The SVM is a statistical method for supervised classification which has been shown to perform better than other machine learning techniques such as Fisher's discriminant and decision trees. In recent years, the SVM is gaining popularity in the analysis of biological problems such as gene and tissue classifications from microarray expression data [6, 9], prediction of protein subcellular localization [6, 15], protein function classification [4], and prediction of protein structural classes [5].

A SVM basically learns how to classify an object $\mathbf{x}$ into two classes $\{-1, +1\}$ from a set of labelled examples $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$. The resulting classifier is formulated as

$$y = f(\mathbf{x}) = \sum_{i=1}^{m} \tau_i K(\mathbf{x}_i, \mathbf{x}), \tag{4}$$

where $\mathbf{x}$ is any new object to be classified, $K(\cdot, \cdot)$ is a kernel function, and $\{\tau_1, \cdots, \tau_m\}$ are the parameters learned. If $f(\mathbf{x})$ is positive, $\mathbf{x}$ is classified into class $+1$. On the contrary, if $f(\mathbf{x})$ is negative, $\mathbf{x}$ is classified into class -1. In this study we use this algorithm by assuming that we have a set of glycans $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$, and $+1$ corresponds to a target organism, and -1 corresponds

to the other species. We use our similarity measure as a kernel function in the SVM algorism as $K(\mathbf{x}, \mathbf{y}) = Sim(\mathbf{x}, \mathbf{y})$. If the resulting kernel matrix is not positive definite, we add an identity matrix whose diagonal elements are the absolute of the minimum engenvalue.

3.3 Characteristic Substructure Extraction

We train the SVM classifer such that the glycans are maximally separated between a target organism and the other species in feature space. That is why the discriminant score y computed by eq.(4) is related to the distance between the objects (glycans in this study) and linear boundary between a target organism and the other species. Therefore, the discriminant score y can be used as the feature which represents the difference of the affiliation of the glycans between the target organism and the other species. In this study, high scoring glycans indicate that they have substructures which are characteristic of a target organism with interest. On the contrary, low scoring glycans indicate that they have substructures which are *not* characteristic of the target organism. The SVM learning process produces a set of discriminant scores $\{y_i\}_{i=1}^m$ for a set of glycans $\{\mathbf{x}_i\}_{i=1}^m$. To evaluate the quantity of how much substructure x is characteristic of a target organism, we use a measure

$$z = \sum_{i=1}^{m} y_i \cdot I\{x \in \mathbf{x}_i\}, \tag{5}$$

where $I\{A\}$ is an indicater function, i.e., $I\{A\} = 1$ if A is true and $I\{A\} = 0$ otherwise. It means that z is the summation of the discriminant scores of the glycans which have the substructure x. Therefore, the high scoring z means that the substructure x is characteristic of a target organism, while the low scoring z means that the substructure x is *not* characteristic of a target organism.

4 Results

4.1 Classification of Species

For each organism, we applied a SVM to predict whether a glycan is assigned to it or not. We used our similarity measure as a kernel function, where parameter α and β are set to 0.5. The number of positive examples (belonging to a target organism) is comparatively fewer than that of negative examples (belonging to the other species). We handled this issue by sampling negative examples randomly in the SVM learning such that the total number of negative examples be equal to the number of the positive examples. Table 2 shows the results of the Jackknife cross-validation experiments of the classification. The results showed that, although the glycan compositions of animal species were similar, our organism classification worked well. The non-animal species (yeast, wheat, and sycamore) were classified with extremely high accuracy because of their unique compositions of carbohydrates. Therefore, we can conclude that our similarity score is valid with respect to classification performance.

4.2 Characteristic Substructures of Glycans

Our goal is to identify characteristic substructures in accord with the organism classification. We extracted a set of candidates of characteristic substructures by applying the procedure explained in section 3.3. Here we focused on the top 20 substructures with highest score for each organism. Figure 4, 5, 6, 7, 8, 9 and 10 show the examples of extracted characteristic substructures of human, rat, mouse, bovine, pig, chicken, yeast, wheat, and sycamore, respectively. In these figures, the 20 substructures of each organism are aggregated such that some characteristic sugar chains are reconstructed from the 20 substructures based on their layer information. Because of space limitations, all the substructures are not shown. All the subcharacteristic structures extracted by our method can be obtained from the author's website (`http://web.kuicr.kyoto-u.ac.jp/~hizukuri/ibsb04/`).

Table 2: Prediction accuracy of classification by cross-validation experiment.

Organism	Total rate	Sensitivity	Specificity
human	80.6 %	78.1 %	83.2 %
rat	78.6 %	84.0 %	73.2 %
mouse	79.5 %	81.7 %	77.3 %
bovine	74.0 %	73.8 %	74.2 %
pig	83.6 %	82.3 %	84.9 %
chicken	84.2 %	80.8 %	87.6 %
yeast	95.0 %	93.0 %	97.0 %
wheat	89.7 %	79.5 %	100.0 %
sycamore	91.8 %	83.7 %	100.0 %

Figure 4, 5, 6, 7, 8 and 9 show the examples of the characteristic substructures in human, rat, mouse, bovine, pig and chicken respectively. Looking at the figures, we can discern several aspects on characteristic glycan substructures across the species. Human and chicken share the same substructure of bisecting GlcNAc which links to Man in the 3rd layer. This substructure is characteristic to the hybrid or complex type of N-glycan core, but the binding type of human is different from that of chicken. The binding type of chicken is β1-4, while the binding type of human is β1-2 and it elongates to Gal in the 5th layer. Chicken has a characteristic binding type of Man in the 4th layer (Manα1-6 Man in the 3-4th layer), which has five binding types to the next monossacharide. This result suggests that the position in chicken has flexibility to bind to the next sugar. Mouse has similar substructures to human, but Fuc in the 6th layer binds to GlcNac in the 5th layer and Gal in the 7th binds to Gal in the 6th layer. Pig has a unique structure of Manα1-2Manα1-3Man in the 3-5th layer and two types of substructures of Man-Man-Man in the 3-5th layer. Bovine has a characteristic linear structure of Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAc. Besides, bovine has a characteristic substructure which has Man in the 2nd layer. Rat has Fuc in the 3rd layer, which links to GlcNAc in the 2nd layer and Galβ1-3GlcNAc in the 5-6th layer.

Figure 10 shows the examples of the characteristic substructures in yeast, wheat, and sycamore respectively. It is found that the extracted substructures of non-animal species reflect their unique carbohydrate composition. For examples, wheat has a characteristic substructures which contain Ara-Xyl at high rates. Actually, it is known that this part is one of the main components of the cell wall of poaceous. The characteristic substructures of yeast are shared by Man as anticipated.

5 Discussion and Conclusion

In this paper, we proposed an approach to classify glycan structures and extract the characteristic glycan substructure of certain species. The biological motivation of this research is similar to the traditional comparative genomic research. The key difference is that our target is glycan not protein. We confirmed that our organism classification of glycans performed well by conducting cross-validation experiments. Finally, our method successfully extracted a set of candidate substructrues which are characteristic to human, rat, mouse, bovine, pig, chicken, yeast, wheat and sycamore, respectively.

The extracted characteristic structures enables us to discern several aspects of the glycan variation in different species, for example, between animals (human, rat, mouse, bovine, pig and chicken), and other species (yeast, wheat and sycamore). The extracted substructures of animals contain GlcNAcβ1-2Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc, which is known to be the N-Glycan core structure of animals.

Several reports have already been made concerning the comparative study of the sugar patterns

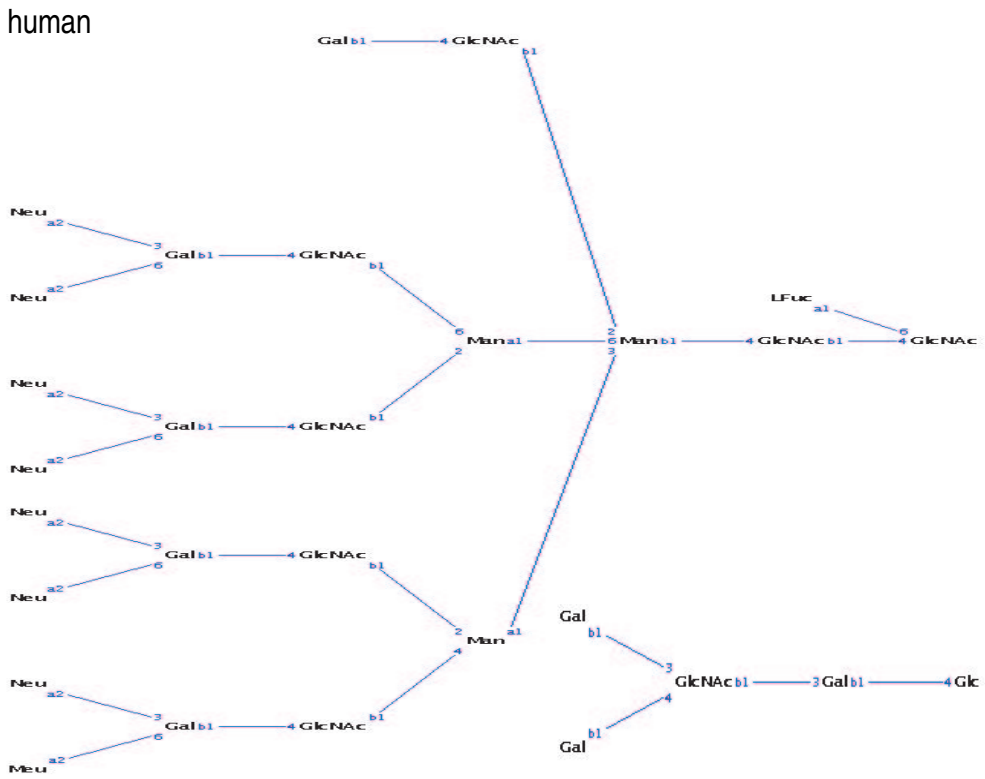

Figure 4: Characteristic substructures of human.

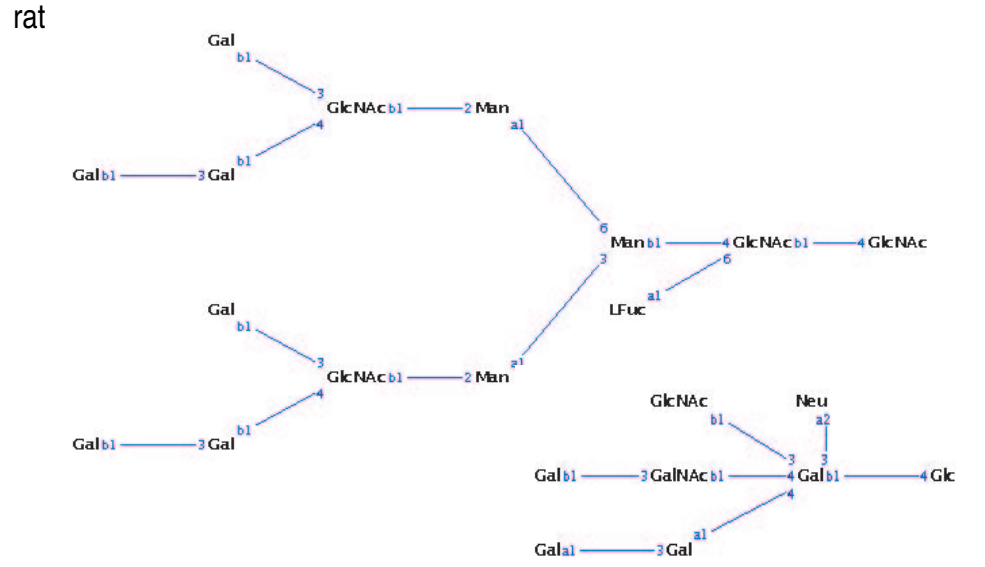

Figure 5: Characteristic substructures of rat.

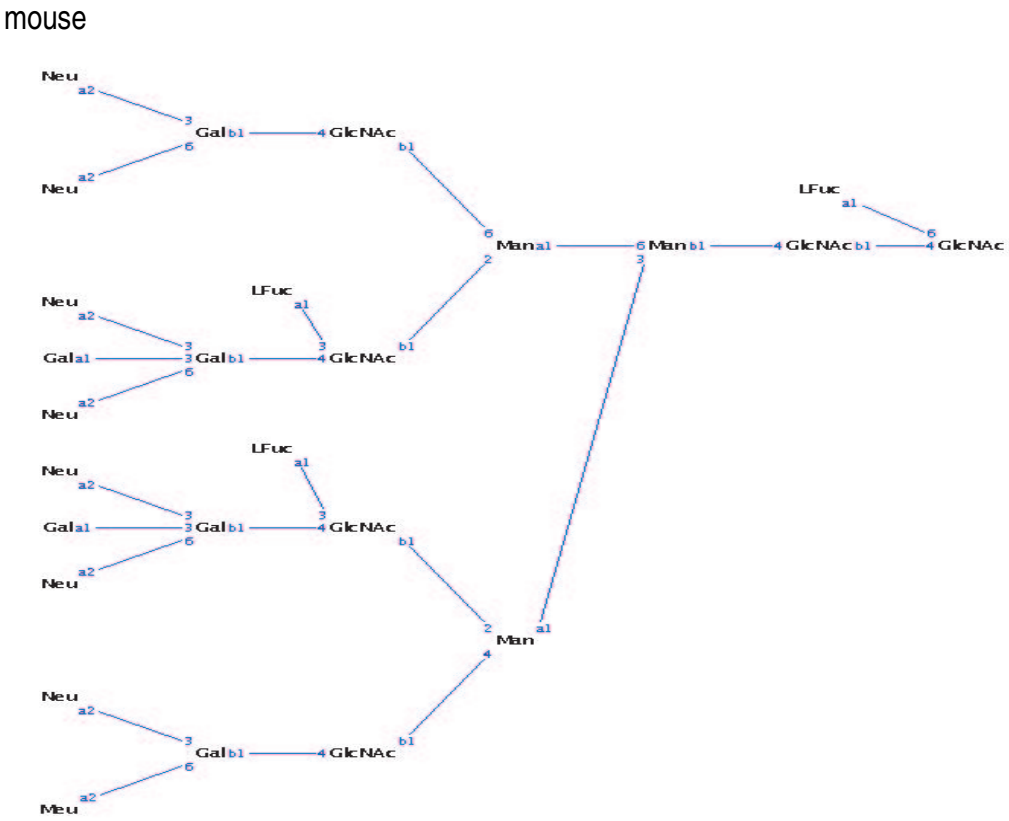

Figure 6: Characteristic substructures of mouse.

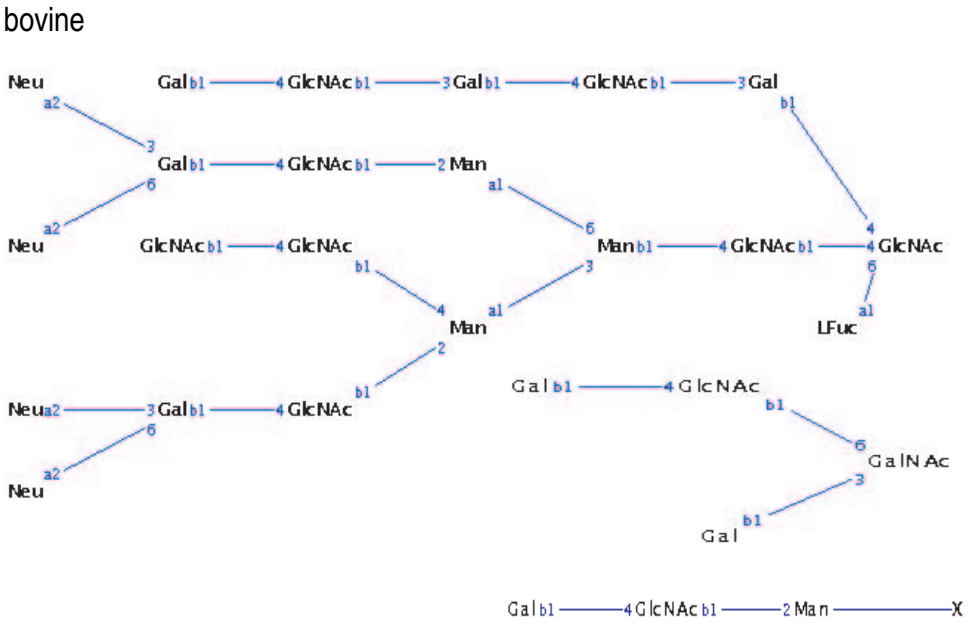

Figure 7: Characteristic substructures of bovine.

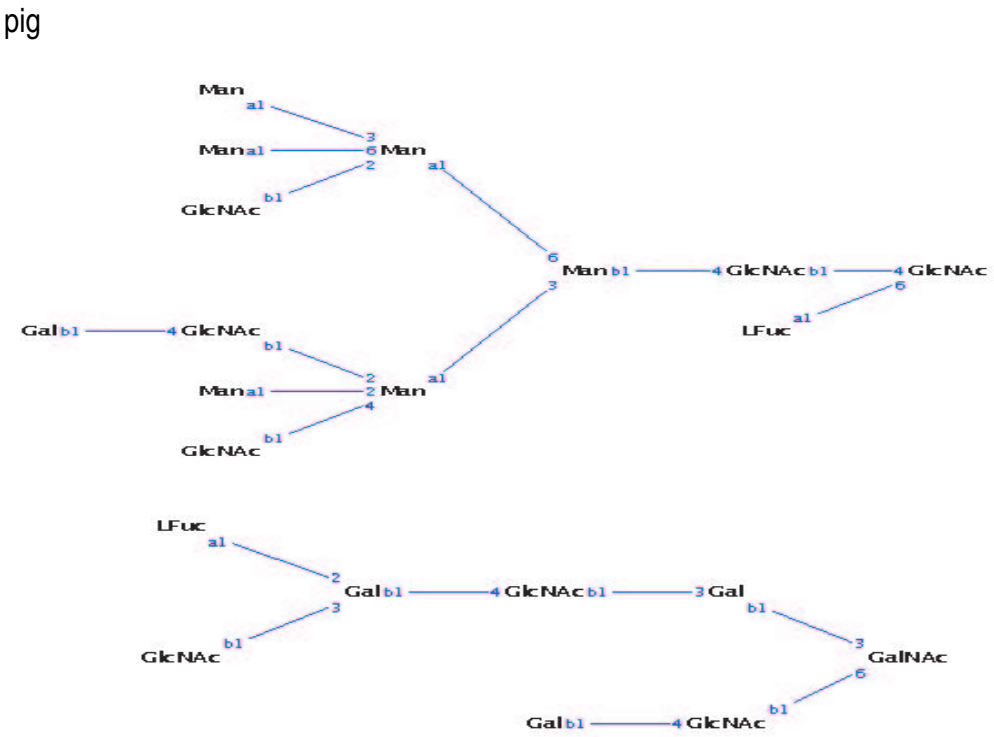

Figure 8: Characteristic substructures of pig.

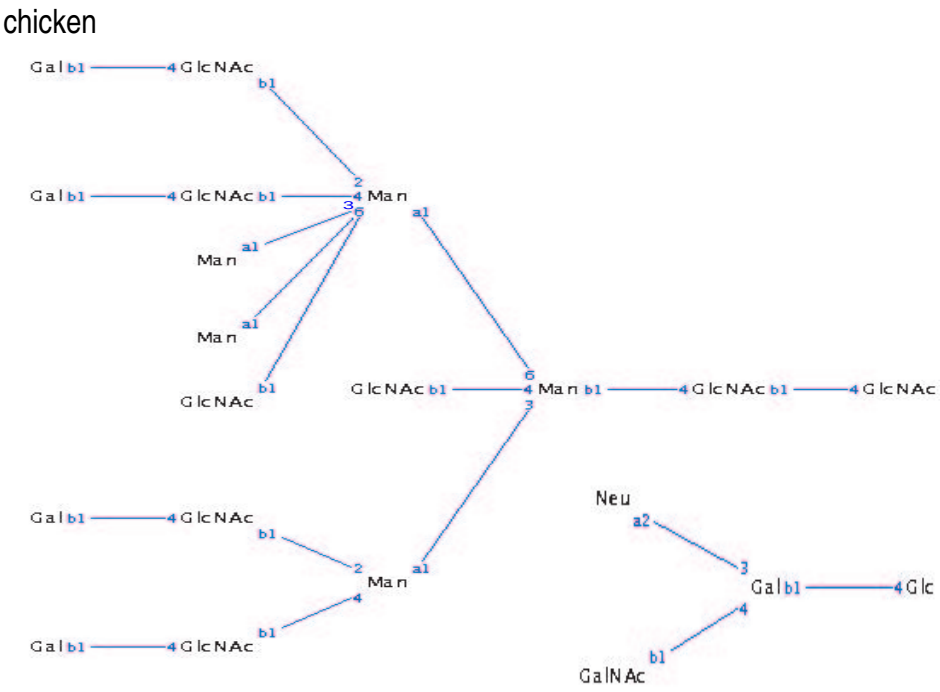

Figure 9: Characteristic substructures of chicken.

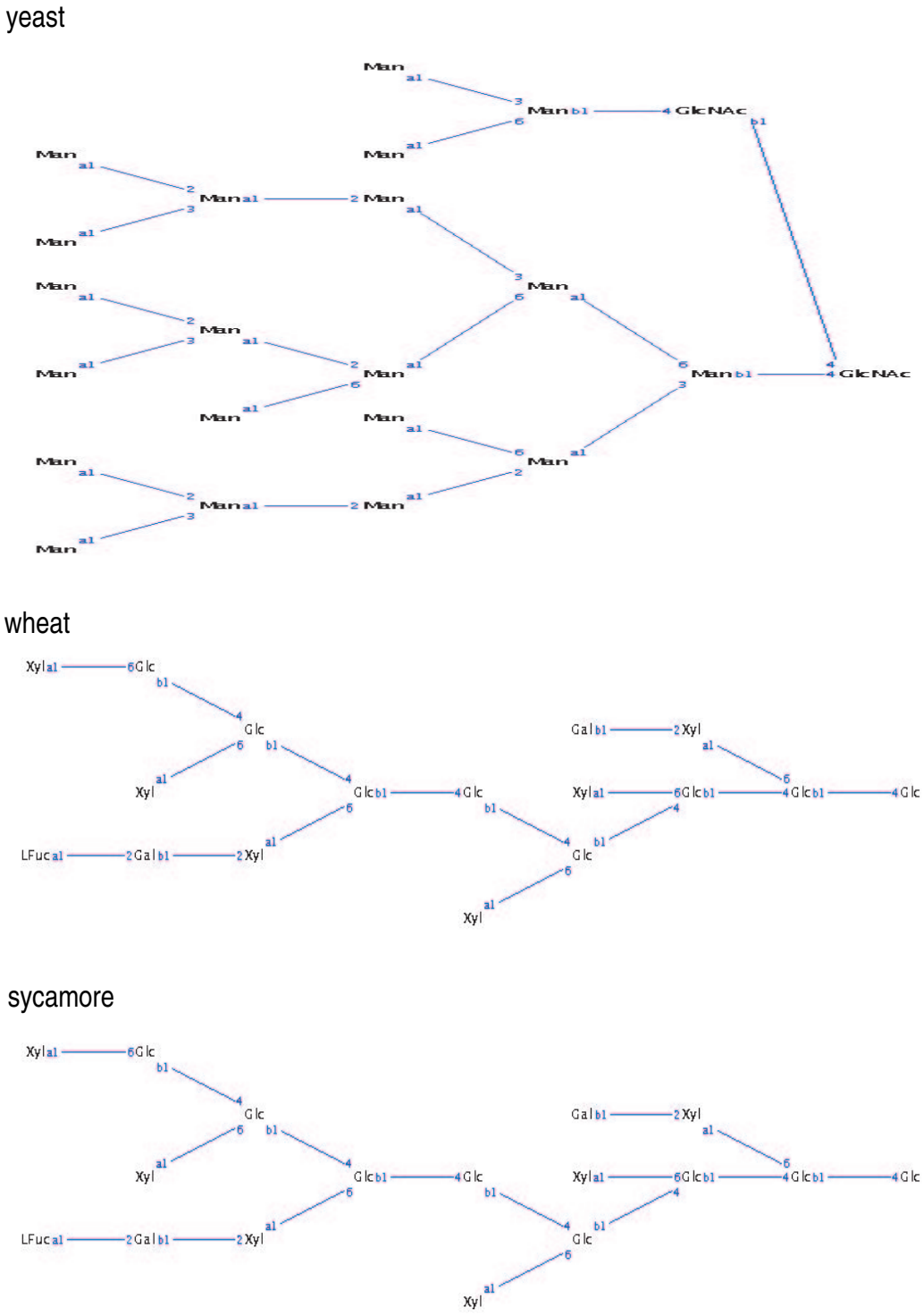

Figure 10: Characteristic substructures of yeast, wheat and sycamore.

of the glycoproteins in the same tissue across different animals from a experimental viewpoint. Examples include the studies on the sugar chains of γ-glutamyltranspeptidases (γ-GTPs) [13, 16]. The reports suggested that a species-specific glycosylation of these proteins often occurs, and confirmed the occurrence of similar phenomena in the N-glycosylation of rhodopsin [13]. Comparing our result with those from the previous reports, we confirmed that the characteristic substructures extracted by our method correspond to the substructures which are known as the species-specific sugar chain of γ-glutamyltranspeptidases in kidneys.

One problem of this study is the limitation of the number of glycan data and bias of their structures in each organism. The number and kinds of glycan structure data depend on the interests of the biologists who determine the glycan structures experimentally. If more comprehensive data of carbohydrates for all the species were available, the results and interpretation might be more clear.

It should be pointed out that our method can be used for other classification problems of glycans. We are currently working on tissue classification of glycans and extracting a set of characteristic substructures of tissue specific carbohydrates. The identification of tissue specific glycan substructures is expected to contribute to the development of drug delivery systems.

Acknowledgments

We thank Susumu Goto and Ruy Jaureguifor for useful discussions. This work was supported by 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. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Ajit, V., Richard, C., Jeffrey, E., Hudson, F., Gerald, H., and Jamey, M., *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, 1999.

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

[3] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., and Haussler, D., Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA*, 97:262–267, 2000.

[4] Cai, C.Z., Wang, W.L., Sun, L.Z., and Chen, Y.Z., Protein function classification via support vector machine approach, *Math. Biosci.*, 185(2):22–111, 2003.

[5] Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C., Prediction of protein structural classes by support vector machines, *Comput. Chem.*, 26(3):293–296, 2002.

[6] Chou, K.C. and Elrod, D.W., Protein subcellular location prediction, *Protein Eng.*, 12:107–118, 1999.

[7] Cristoni, S., and Bernardi L.R., Development of new methodologies for the mass spectrometry study of bioorganic macromolecules, *Mass Spectrom. Rev.*, 22(6):369–406, 2003.

[8] Doubet, S., and Albersheim, P., CarbBank, *Glycobiology*, 2(6):505, 1992.

[9] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16:906–914, 2000.

[10] Hashimoto, K., Hamajima, M., Goto, S., Masumoto, S., Kawashima, M., and Kanehisa, M., GLYCAN: The database of carbohydrate structures, *Genome Informatics*, 14:649–650, 2003.

[11] Hearst, M.A., Schölkopf, B., Dumais, S., Osuna, E., and Platt, J., Trends and controversies - support vector machines, *IEEE Intelligent Systems*, 13(4):18–28, 1998.

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

[13] Kobata A., A journey to the world of glycobiology, *Glycoconjugate Journal* , 17:443–464, 2000.

[14] Kukuruzinska, M.A., Bergh, M.L., and Jackson, B.J., Protein glycosylation in yeast, *Annu. Rev. Biochem*, 56:915–944, 1987.

[15] Park, K. and Kanehisa, M., Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics*, 19(13):1656–1663, 2003.

[16] Rademacher, T.W., Parekh, R.B., and Dwek, R.A., Glycobiology, *Annu. Rev. Biochem*, 57:785–838, 1988.

[17] Schölkopf, B. and Smola, A.J., *Learning with Kernels*, MIT Press, 2002.

[18] `http://glycan.genome.ad.jp/`

[19] `http://www.boc.chem.uu.nl/sugabase/carbbank.html`