Available online at www.sciencedirect.com
SCIENCE DIRECT®
Carbohydrate Research 340 (2005) 2270–2278
Carbohydrate RESEARCH

Extraction of leukemia specific glycan motifs in humans by computational glycomics

Yoshiyuki Hizukuri,[a] Yoshihiro Yamanishi,[a] Osamu Nakamura,[b] Fumio Yagi,[c] Susumu Goto[a] and Minoru Kanehisa[a,*]

[a]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan*
[b]*National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan*
[c]*Kagoshima University, 1-21-24 Korimoto, Kagoshima, Kagoshima 890-0065, Japan*

Received 14 June 2005; received in revised form 19 July 2005; accepted 22 July 2005
Available online 10 August 2005

Abstract—There have been almost no standard methods for conducting computational analyses on glycan structures in comparison to DNA and proteins. In this paper, we present a novel method for extracting functional motifs from glycan structures using the KEGG/GLYCAN database. First, we developed a new similarity measure for comparing glycan structures taking into account the characteristic mechanisms of glycan biosynthesis, and we tested its ability to classify glycans of different blood components in the framework of support vector machines (SVMs). The results show that our method can successfully classify glycans from four types of human blood components: leukemic cells, erythrocyte, serum, and plasma. Next, we extracted characteristic functional motifs of glycans considered to be specific to each blood component. We predicted the substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc as a leukemia specific glycan motif. Based on the fact that the *Agrocybe cylindracea* galectin (ACG) specifically binds to the same substructure, we conducted an experiment using cell agglutination assay and confirmed that this fungal lectin specifically recognized human leukemic cells.

Keywords: Computational glycomics; Glycan classification; Glycome informatics; Glycan motif; Support vector machine

1. Introduction

Glycans play key roles in cellular functions including cell–cell communication, protein interaction, immunity, etc. However, it has been difficult to assess the structural elements that are responsible for specific functions. One of the main obstacles to performing such assessments has been a lack of computerized resources for glycan structures and functions, as well as associated computational methods. In contrast, much computational research has been conducted on DNA and proteins. For example, much work has been done on the classification of protein sequences and 3D structures from publicly available databases, and the relationships between structural patterns and functional implications are stored in motif databases such as Pfam,[1] ProDom,[2] and PROSITE.[3] However, there have been almost no reports of similar computational studies on broad-scale glycan structure analyses for understanding function. There have been experimental reports on comparative studies on glycan structures across species and tissues,[4,5] but these were conducted on a small scale. We have developed the KEGG GLYCAN database[6] containing all known glycan structures in the literature. Here, we present a novel method utilizing this resource for comparative glycomics to classify glycan structures and extract their functional motifs.

There are mainly two differences between glycan structures and the linear structures of DNA and

Abbreviations: SVM, support vector machine; ACG, *Agrocybe cylindracea* galectin; OD, optical density; DDS, drug delivery system; DEAE, diethylaminoethyl; PBS, phosphate-buffered saline.

* Corresponding author. Tel.: +81 774 38 3270; fax: +81 774 38 3269; e-mail: kanehisa@kuicr.kyoto-u.ac.jp

0008-6215/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carres.2005.07.012

proteins. First, they are biosynthesized by completely different mechanisms. The synthesis of DNA and proteins strictly depends on the linear template structures of their corresponding DNA and RNA; such simple structures are well suited for automatic analysis. On the other hand, glycans are synthesized by glycosyltransferases that individually attach monosaccharides to a substrate structure, so glycan structures depend heavily on the localization, expression pattern, and substrate specificity of the glycosyltransferases. Second, the branched structures of glycans are much more complex, especially considering the variety of linkages in glycans compared to the single bond type of DNA and proteins (i.e., phosphodiester and peptide bonds). It is therefore desirable that we take these glycan-specific properties into consideration in our analysis.

Another unique feature of glycans is their diversity in various biological classes. For example, their structures are known to be species-specific, but their structures also differ between organs and tissues even in the same species.[4,5] This suggests that there might exist characteristic glycan substructures that contribute to the differences between organs and tissues. In other words, substructures in glycans may function as signal molecules, being recognized by lectins and enzymes to play important biological functions. There have been interesting reports that altered glycan structures occur in tumor cells[7,8] and that some plant lectins bind to certain types of leukemic cells.[9] This suggests that there may be tumor-specific substructures, or glycan motifs, that may be important to distinguish tumor cells from normal cells.[10] It is therefore necessary to take species-specific, tissue-specific, and cell-specific properties into consideration to reveal the pathogenic diversities and biological functions of glycans.

To our knowledge, no previous reports have focused on the identification of common characteristic substructures across all leukemic cells through a computational analysis. In this study, we computationally determine such common characteristic substructures by a novel method to classify blood components and to extract glycan motifs specific to leukemia. First, we developed a new similarity measure for comparing glycan structures, because standard methods for analyzing DNA or protein structures cannot be directly used. The originality of our similarity measure is that we take into account biological properties of glycan biosynthesis such as substrate specificity of glycosyltransferases and structural flexibility in the interaction region. Specifically, we assign similarity scores to trimers in a position dependent manner. Using this similarity measure we classified glycans from different blood components in the framework of support vector machines (SVMs), statistical classifiers[11] commonly used in DNA and protein sequence analysis. Our SVM was able to accurately classify glycans from different human blood components: leukocyte in a diseased condition, erythrocyte, plasma, and serum. Next, using the SVM-classified dataset we extracted the characteristic functional units (motifs), which we claim to be substructures specific to each blood component. We found the substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc as a motif specific to leukemia. Finally, we conducted an experiment based on cell agglutination assay to verify the specificity of this motif. From this experiment, we found that a fungal lectin from *Agrocybe cylindracea*, which is known to recognize the same substructure,[12] was able to discriminate leukemic cells.

2. Results

2.1. New similarity measure of glycan structures

We developed a similarity measure between glycans based on their tree structures and biological knowledge about glycan biosynthesis. From a biological viewpoint, glycans and associated glycosyltransferases have the following properties:

1. Glycosyltransferases physically interact with about three monosaccharides at the leaves.[13]
2. The variability of the sugars near the leaf (referred to as the variable part) is larger than those near the root (referred to as the core part, see Fig. 1).
3. Major functional and structural classes of glycans are determined by the chain of monosaccharides at the root (e.g., *N*-glycans, *O*-glycans, Glycolipids, etc.).

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

In our similarity measure, we propose to decompose glycan tree structures into sets of substructures (3-mers in this study), because there is an observation that many glycosyltransferases physically interact with about three linked monosaccharides, as described in Property 1 above. Suppose that we have two glycans X and Y, and we decompose glycans X and Y into sets of 3-mers. Figure 2 shows an illustration of the decomposition of glycan structures, where each node represents a monosaccharide, and the number indicates the layer defined by the distance of each substructure from the root. We decomposed all of the glycan structures in our data set and obtained 260 types of substructures. For a given glycan, we counted the occurrence of each substructure pattern. As a result, we obtained feature vectors $\vec{X}\ (x_1, x_2, \ldots, x_{260})$ and $\vec{Y}\ (y_1, y_2, \ldots, y_{260})$ for glycans X and Y, respectively. The straightforward similarity score for X and Y is thus defined as

$$\mathrm{sim}(X, Y) = \vec{X} \cdot \vec{Y}. \quad (1)$$

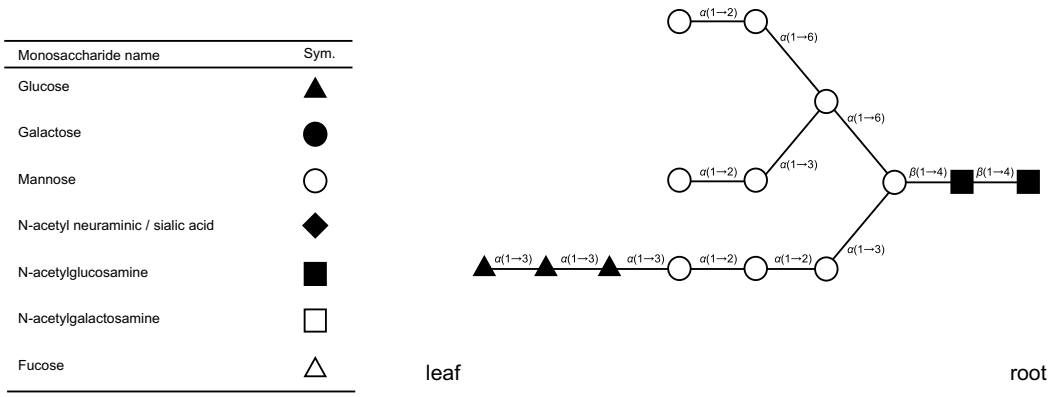

Figure 1. An illustration of glycan structure showing the symbol nomenclature for common monosaccharides in humans.

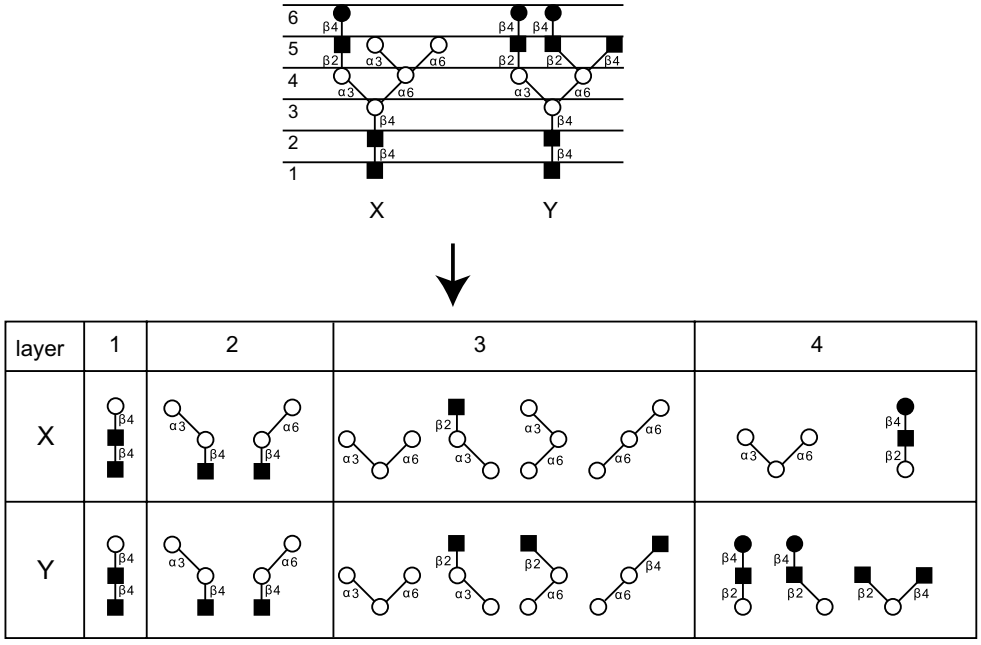

Figure 2. An illustration of the decomposition of glycan structures into sets of 3-mers. Each node represents a monosaccharide. The numbers on the left (layers) indicate the distance of each substructure from the root.

Most glycan structures can be divided into two regions: the core region and the variable region, as described in Property 2 above. The core region corresponds to the substructure near the root, and the variable region corresponds to the rest of the structure. The glycan structure in the variable region is known to be more diverse than the core regions, so the variation in the variable region is considered as a factor that determines the characteristics and specific functions of glycans. Therefore, it is natural to either put more weight on the matching of substructures in the variable region, or put less weight on the matching of substructures in the core region. To do so, we incorporate the following weighting parameter in calculating the inner product of $\vec{X}$ and $\vec{Y}$.

$$\mathrm{sim}(X, Y) = \sum_{k=1}^{260} w_k x_k y_k, \tag{2}$$

where w_k is defined as

$$\begin{aligned} w_k &= 1 - \exp(-\alpha h) \quad \text{if } h > 1, \\ w_k &= 1 \qquad\qquad\quad\ \ \text{if } h = 1, \end{aligned} \tag{3}$$

where h is the layer of the matching substructures and α is a positive constant (in this work, the parameter is set to 0.5). When the matching substructure is found at the root, the weight is set to 1, which reflects distinction of glycan types, such as *N*-glycans and *O*-glycans (Property 3). The weight w plays the following roles: the larger the distance between the matching substructure and the

root, the larger is the value of the similarity, which reflects Property 2.

2.2. Classification of glycans between blood components

By using our similarity measure, we performed a supervised classification of glycan structures with respect to blood components for automatic classification. For each blood component, we applied a support vector machine (SVM), or a statistical classifier, to predict whether a glycan is assigned to the target blood component (positive class) or not (negative class). We tested its classification ability with Jack-knife (leave-one-out) cross-validation. In each cross-validation run, we adjusted the sizes of the training dataset and the test dataset so that the numbers of positive and negative structures are equal. Initially, we compared several sizes of substructures to classify the glycans. After that, we assessed the accuracy of classification by evaluating sensitivity, specificity, and overall accuracies (details in Materials and methods). Figure 3 shows the classification accuracies (Q) of each substructure and we found that 3-mers are the optimum size. Table 1 shows the results of computing sensitivity and specificity for each blood classification using 3-mers. In each case, our method seems to capture the information to classify glycans into different blood components. This result suggests that all glycan families from each blood component might have their own specific substructures, or glycan motifs.

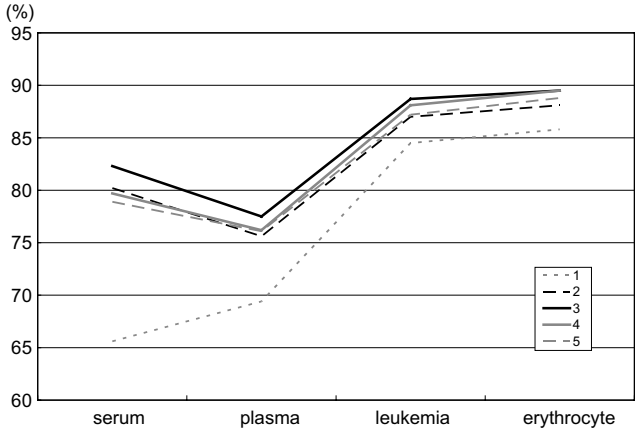

Figure 3. The classification accuracy (Q) of blood components. Each line represents the accuracy of different sizes of glycans.

Table 1. Prediction accuracies in the classification of four different blood components with SVM

	Number of glycan	Sensitivity (%)	Specificity (%)
Leukemic cells	162	87.0	95.7
Erythrocyte	112	91.1	88.4
Serum	85	83.5	87.1
Plasma	73	84.9	74.0

2.3. Extraction of leukemia specific glycan motifs

We extracted glycan motifs specific to each blood cell. The main objective of this work is to detect the characteristic glycan substructures for the successful discrimination between blood cells. Because we train the SVM classifier such that the glycans are maximally separated between a target blood component and other blood components in a feature space, the discriminant score is related to the distance between the glycan and the linear boundary (Fig. 4; details in Materials and methods). Therefore, the discriminant score can be used as a quantity that represents the difference between the glycans of the target component and the other components. In this study, high-scoring glycans indicate that they might have substructures that are characteristic to the target blood component. On the contrary, low-scoring glycans indicate that they do not contain substructures that are characteristic to the target component. The learning process of the SVM produces a set of discriminant scores $\{y_i\}_{i=1}^{m}$ for a set of target glycans $\{X_i\}_{i=1}^{m}$. To evaluate how characteristic substructure x is, we use the following measure:

$$z(x) = \sum_{i=1}^{m} y_i \cdot I\{x \in X_i\}, \tag{4}$$

where $I\{\mathrm{A}\}$ is an indicator function, that is, $I\{\mathrm{A}\} = 1$ if A is true and $I\{\mathrm{A}\} = 0$ otherwise. z is the summation of the discriminant scores of the glycans that contain substructure x. Therefore, a high score of z means that the substructure x is characteristic of the target component, while the low score of z means that the substructure x is not characteristic of the target component. We define the score of z as our 'characteristic score'.

Table 2 shows the top five high-scoring motif candidates considered to be the substructures specific to leukemic cells, erythrocyte, serum, and plasma. The

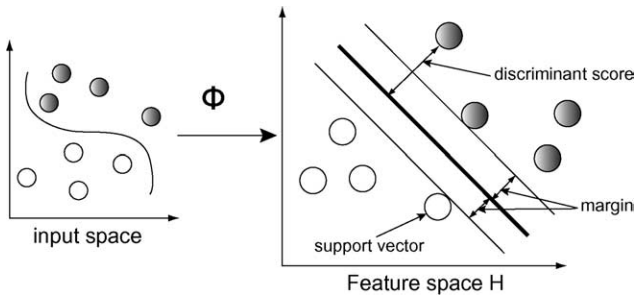

Figure 4. An illustration of the SVM training method. SVM defines a mapping Φ and constructs the optimal separating hyperplane in the high-dimensional feature space H. Black and white circles indicate positive and negative samples, respectively.

Table 2. The top five high-scoring motif candidates suspected to be the substructures specific to leukemic cells, erythrocyte, serum, and plasma

Substructures	Layer	Characteristic scores
Leukemic cells		
α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc	5	161.2
β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→2)-D-Man*p*	4	159.6
α-D-Neu*p*5Ac-(2→6)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc	5	148.8
β-D-Glc*p*NAc-(1→2)-α-D-Man*p*-(1→3)-D-Man*p*	3	78.7
β-D-Glc*p*NAc-(1→2)-α-D-Man*p*-(1→6)-D-Man*p*	3	77.6
Erythrocyte		
β-D-Glc*p*NAc-(1→3)-β-D-Gal*p*-(1→4)-D-Glc*p*	1	41.4
β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→3)-D-Gal*p*	2	37.9
β-D-Glc*p*NAc-(1→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc	3	19.9
β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→3)-D-Gal*p*	4	19.0
α-L-Fuc*p*-(1→2)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc	5	11.2
Serum		
β-D-Glc*p*NAc-(1→4)-β-D-Man*p*-(1→4)-D-Glc*p*NAc	2	13.0
α-D-Man*p*-(1→3)-[β-D-Glc*p*NAc-(1→4)]-D-Man*p*	3	13.0
α-D-Man*p*-(1→6)-[β-D-Glc*p*NAc-(1→4)]-D-Man*p*	3	13.0
α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*	1	10.7
β-D-Man*p*-(1→4)-β-D-Glc*p*NAc-(1→4)-D-Glc*p*NAc	1	8.6
Plasma		
α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*	1	10.1
β-D-Glc*p*NAc-(1→4)-α-D-Man*p*-(1→3)-D-Man*p*	3	9.4
α-L-Fuc*p*-(1→3)-β-D-Glc*p*NAc-(1→4)-D-Man*p*	4	8.8
β-D-Gal*p*NAc-(1→4)-β-D-Gal*p*-(1→4)-D-Glc*p*	1	7.7
β-D-Gal*p*-(1→3)-β-D-Glc*p*NAc-(1→3)-D-Gal*p*	2	7.3

characteristic score is the sum of the discriminant scores of each substructure at each layer, and a score over 100 is considered to be meaningful. Our result shows that leukemic cells have many high-scoring characteristic substructures. The substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc at the fifth layer is the highest in leukemic cells and it frequently appeared in leukemic cells over 20 times more than in other components. The characteristic scores of the substructures of serum and plasma are relatively low. The substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p* at the first layer is highest in plasma and it appeared about 2.6 times more than other components. This result reflects the fact that many components of serum and plasma are common and that they share many frequently appearing glycan substructures.

We focused on the characteristic glycan motifs derived from leukemic cells. Among the high-scoring motif candidates, we predicted the substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc as a leukemia specific glycan motif. This substructure received the highest characteristic score and the experimental studies indicate that sialic acid tends to appear in many tumor cells.[10]

2.4. Experimental verification

We conducted an experiment to verify that α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc is a key substructure to distinguish leukemic cells from normal cells. Based upon the fact that *A. cylindracea* galectin (ACG) specifically binds to the substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc,[12] we examined whether or not ACG binds specifically to leukemic cells using a cell agglutination assay. In this assay, we used normal cells and five types of leukemic cells from B-cell and T-cell lines. The cytoagglutinating activity of ACG toward normal lymphocytes and cultured human leukemic cells was determined at 50 nM. ACG showed different cytoagglutinating activities depending on the cell origin. As shown in Figure 5 cytoagglutination of normal lymphocytes by ACG was clearly lower than any leukemic cell. Furthermore, we tested whether or not this cytoagglutination is inhibited by the α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc structure. We confirmed that while the addition of mannose had no effect, the addition of α-D-Neu*p*5Ac-(2→3)-*N*-acetyl-lactosamine indeed inhibited the cytoagglutination of leukemic cells from both T cell lines (Jurkat, MOLT-4, and CCRF-HSB-2) and B cell line (NALM-6) (data not shown). These results suggest that several types of leukemic cells contain the same characteristic glycan motif and that our method successfully extracted an informative glycan motif.

3. Discussion

The objectives of this paper are to classify glycans by comparing their structures and extracting the character-

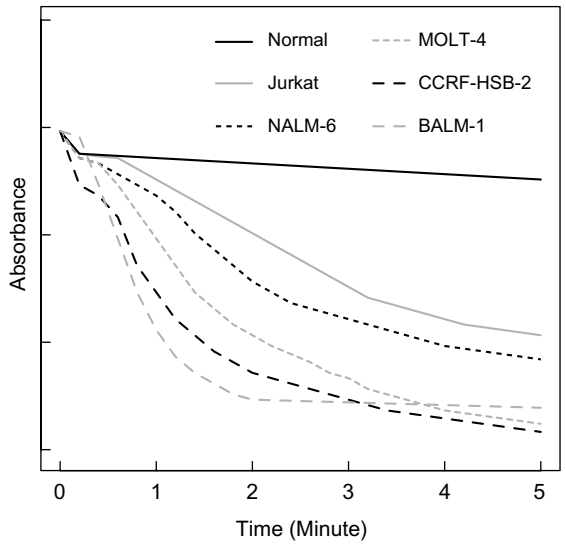

Figure 5. The agglutination reaction of ACG with leukemic cells. Human leukemic T cell lines (Jurkat, MOLT-4, and CCRF-HSB-2), B cell lines (NALM-6 and BALM-1), or normal leukocytes in PBS (−) (1.5×10^6 cells/mL), and PBS (−) were placed into the sample and reference cuvettes, respectively. ACG dissolved in PBS (−) (final concentration, 50 nM) was added into the sample cuvette containing cell suspension and the turbidity of cell suspension was monitored at 600 nm with continuous stirring. The cytoagglutination of normal lymphocytes by ACG was clearly lower than any leukemic cell.

istic glycan motifs from each classification group. Toward these goals, we developed a new similarity measure of glycan structures taking several important biological properties of glycan biosynthesis into consideration. Our similarity measures were adapted as the kernel function of an SVM, which utilized the discriminant scores for glycan motif detection. The SVM classifier is a popular method used in bioinformatics for classifying DNA and proteins. One advantage of SVMs is that they can be applied to any data structure including sequences, graphs, and trees. Because of this, SVMs are well suited for the classification of complicated structures such as chemical compounds and glycans.

In our algorithm, we incorporate biological properties of glycan synthesis into our similarity measure: 'trimers' and 'layers.' We confirmed the effectiveness of 'trimers' in glycan–protein interactions. Figure 3 shows that 3-mers work best, which fits with the fact that glycosyltransferase interact with about three monosaccharides (Section 2.1, Property 1) the accuracy of classification peaks using trimers. In the case of monomers and dimers, SVM cannot learn the variety of linkage and branch types as accurately. In the case of 4- and 5-mers, the number of substructure patterns becomes large and the performance of the substructure matching becomes inefficient. The concept of 'layers' was used because it is well known that the substructure position in the glycan is important for glycan–protein interaction both in the biosynthesis and protein recognition. It has been shown that the layer parameter has advantageous effects in finding glycan motifs.[14]

The glycan substructure α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc scored the highest in Table 2, but the scores of the second β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→2)-D-Man*p* and the third α-D-Neu*p*5Ac-(2→6)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc were almost equally high. In fact, these three top-ranking structures consistently scored highly regardless of the choice of the parameters or any changes in the negative sample set. Based on the layer information, these structures can be reconstructed as α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→2)-D-Man*p* and α-D-Neu*p*5Ac-(2→6)-β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→2)-D-Man*p*. Our experimental verification shows that ACG, which specifically recognizes the α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→4)-D-Glc*p*NAc structure, may discriminate leukemic cells from the normal cells. It has also been reported that β-galactoside-α-(2→6)-sialyltransferase is expressed more highly in chronic lymphocytic leukemia B cells than in normal B cells,[15] indicating that the α-D-Neu*p*5Ac-(2→6)-β-D-Gal*p* structure may accumulate in leukemic cells. These results suggest that the α-(2→3), and α-(2→6) sialylation of β-D-Gal*p*-(1→4)-β-D-Glc*p*NAc-(1→2)-D-Man*p* in the fourth layer is characteristic in leukemic cells, confirming our computational results. Further experiments quantitating lectin binding would provide stronger experimental results, especially considering that ACG is also known to bind to α-D-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→3)-D-Glc*p*NAc. For this case, we make note that our leukemia data set only contained one such structure, indicating that this structure has a very low possibility of being a characteristic motif of leukemic cells. As such, we believe that these current results sufficiently confirmed the accuracy of our method.

In this study, we performed blood classification of glycans, and we focused on the extraction of leukemia specific glycan motifs. However, our work can easily be applied to any type of target such as species, tissues, or pathogenic cells, and with emerging glycomics techniques quantitating relative amounts of glycans in different biological sources, our method will be even more efficient in mapping structural features to specific classes of these biological sources. Recently, the classification of tissues and the discrimination of tumor cells have become important issues in pharmaceutical applications such as targeted drug delivery systems (DDS). Targeted DDS are a developing technology for making an effective drug, but there have been few successful reports. One of the main problems in developing these DDS stems from difficulties in distinguishing pathogenic cells from normal cells. Glycans are strong candidates as targets in DDS development, because of their wide-ranging diversity in tissues and between normal and pathogenic cells. We expect that our method might contribute to

DDS development through the classification of glycan structures and the extraction of glycan functional motifs.

The next stage of this work is the integration of glycosyltransferase gene data. Glycan structures are indirectly synthesized by the dynamic functioning of a variety of glycosyltransferase genes. That is, they are dependent on the behavior of the spatial and temporal expression patterns of these genes. Thus, there is a relationship between glycans and genes, but this has not been explored in depth. There are comprehensive analyses of gene expression patterns between tumor and normal cells,[16] but no comprehensive analysis of the connections between gene expression and glycan structure. By comparing the glycan motifs with the glycosyltransferase expression patterns of the tumor and normal cells, we can clarify the pathogenic pathway of glycan synthesis.

From the viewpoint of classification and motif extraction, the global research direction of our work corresponds to that of comparative genomics and proteomics. Motifs in protein sequences are essential functional units conserved across many species, because it is observed that the presence of such motifs determine biological functions of proteins. Therefore, classification and motif extraction have been important issues in analyzing the protein sequences.

Similarly, the classification and motif extraction of glycans are important to understand biological diversity. However, because of the structural and biological complexity of glycans, there have been no standard methods to study glycomics computationally. Compared to computational genomics or proteomics, the difficulty in computational glycomics is in the biological diversity of the structures in question. Glycan structures are diverse not only within species but also between different tissues and cells. Insight into biological properties of glycans will open the door to a new research field in comparative computational biology.

4. Materials and methods

4.1. Collection and annotation of glycan structures

All glycan structures used in this study were obtained from the KEGG/GLYCAN database[6] and the corresponding annotations of biological source were obtained from the CarbBank/CCSD database.[17,18] The CarbBank/CCSD and KEGG/GLYCAN structures can be linked by CCSD ID number. We identified glycan structures with specific biological sources according to the annotation information in the BS fields of CarbBank/CCSD. Using the CCSD ID number in the BS fields, we collected the carbohydrate structures from KEGG/GLYCAN. We used the glycan structures of four human blood components: leukemic cells, erythrocyte, serum, and plasma. In the case of the human dataset, glycan structures mainly consist of seven monosaccharides: Glc*p*, Gal*p*, Man*p*, Fuc*p*, Glc*p*NAc, Gal*p*NAc, and Neu*p*5Ac, making up around 98% of the glycan structures in this dataset. In this work, any modified monosaccharides are relabeled as its corresponding base monosaccharide, and we removed any non-carbohydrate moieties such as phosphate and sulfate. We selected the blood components that contain over 50 glycan structures in the database. Finally, we obtained a set of 356 glycan structures derived from human blood components, and the numbers of glycans of leukemic cells, erythrocyte, serum, and plasma are 162, 112, 85, and 73, respectively. We called each set of glycan structures from these blood components a 'glycan family.'

4.2. Classification of glycans in blood by support vector machines

To classify glycans between cell types based on glycan structure, we used the support vector machine (SVM) in this study. SVM is a supervised classifier based on statistical learning theory. In recent years, SVM has gained popularity in the analysis of biological problems such as gene, tissue, and protein function classification, remote homology detection, and protein structural prediction.[19–23] For example, by learning a set of positively and negatively labeled training samples, SVM classifies new unlabeled test samples. From a given set of labeled m samples $\{X_1, X_2, X_3, \ldots, X_m\}$, SVM determines whether a new object X can be classified or not $\{-1$ or $1\}$.

In Figure 4 the SVM method defines a mapping Φ, and builds a linear SVM in the high-dimensional feature space H. Instead of explicitly mapping the objects to H, SVM usually works implicitly in the feature space by only computing the corresponding kernel function $K(X, Y)$. The resulting classifier is formulated as

$$y = f(X) = \sum_{i=1}^{m} \tau_i K(X_i, X), \tag{5}$$

where X is any new object to be classified, K is a kernel function, and $\{\tau_1, \ldots, \tau_m\}$ are the parameters learned. If $f(X)$ is positive, X is classified into class positive. On the contrary, if $f(X)$ is negative, X is classified into class negative. In this study, we use this algorithm by assuming that we have a set of glycans $\{X_1, X_2, \ldots, X_m\}$, and positive corresponds to a certain target blood component, and negative corresponds to the other components. We used our similarity measure as a kernel function in the SVM algorithm as $K(X, Y) = \mathrm{sim}(X, Y)$. We used the Gist publicly available SVM software implementation (http://microarray.cpmc.columbia.edu/gist/).[24]

The accuracy of the classification ability is assessed by evaluating the sensitivity, specificity, and overall

accuracy (Q) from the Jack-knife cross-validation experiment, which are defined as

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \quad (6)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}), \quad (7)$$

$$Q = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}), \quad (8)$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives, respectively. A true positive is a result that accurately predicts a positive example, and a false positive is the one that inaccurately predicts a negative example as being positive. A true negative is a result that accurately predicts a negative example, and a false negative is the one that inaccurately predicts a positive example as being negative.

4.3. Experimental verification of extracted glycan substructure

4.3.1. Purification of lectin (galectin) from *A. cylindracea*. *A. cylindracea* galectin (ACG) was purified according to the procedure by Yagi et al.[12] Briefly, ACG was purified by ammonium sulfate precipitation, and ion exchange column chromatography on DEAE-cellulofine A-200 (Seikagaku Co., Tokyo, Japan) and then DEAE-Toyopearl 650M, followed by gel filtration chromatography on Toyopearl HW55F (Toso Ltd., Tokyo, Japan).

4.3.2. Preparation of cells. The human leukemic T cell lines (Jurkat, MOLT-4, and CCRF-HSB-2) and B cell Lines (NALM-6 and BALM-1) (Hayashibara Biochemical Laboratories, Inc., Okayama, Japan) were cultured in RPMI-1640 medium supplemented with 10% heat-inactivated fetal bovine serum, 100 μg/mL streptomycin, and 100 units/mL of penicillin under a humidified atmosphere with 5% CO_2 at 37 °C. The cells used for assay were in a logarithmic phase. Normal human lymphocytes were purified from heparinized peripheral blood obtained from normal adults by Lymphosepar I (IBL Co., Ltd., Takasaki, Japan). They were collected by centrifugation and washed by phosphate-buffered saline (Ca^{2+} and Mg^{2+} free) (PBS (−)) before assay.

4.3.3. Cytoagglutination assay. ACG-induced cytoagglutination of human leukemic cells was observed as described before.[25] Briefly, the cells for assay were washed with PBS (−) three times and re-suspended into PBS (−) to a concentration of 1.5×10^6 cells/mL. The cell suspension of leukemic cells or normal leukocytes and PBS (−) were placed into the sample and reference cuvettes, respectively, and left until the baseline at 600 nm became constant by stirring. ACG dissolved in PBS (−) (final concentration, 50 nM) was added into the sample cuvette containing cell suspension and the turbidity of cell suspension was monitored at 600 nm with continuous stirring. The experiment was stopped when the decrease of OD_{600} reached plateau. A spectrophotometer equipped with a magnetic stirrer, Jasco V-550, was used for this experiment.

4.3.4. Reagents. All reagents of analytical grade were purchased from Wako Pure Chemical Industries, Ltd. (Osaka, Japan) and Invitrogen Corp. (Carlsbad, CA, USA).

Acknowledgments

We thank Dr. Kiyoko F. Aoki-Kinoshita for helpful discussions and critical reading of the manuscript. 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. Sonnhammer, E. L.; Eddy, S. R.; Durbin, R. *Proteins: Struct. Funct. Bioinf.* **1997**, *28*, 405–420.
2. Corpet, F.; Gouzy, J.; Kahn, D. *Nucleic Acids Res.* **1998**, *26*, 323–326.
3. Bairoch, A. *Nucleic Acids Res.* **1991**, *19*, 2241–2245.
4. Kobata, A. *Glycoconjugate J.* **2000**, *17*, 443–464.
5. Rademacher, T. W.; Parekh, R. B.; Dwek, R. A. *Annu. Rev. Biochem.* **1988**, *57*, 785–838.
6. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. *Nucleic Acids Res.* **2004**, *32*, D277–D280.
7. Kannagi, R.; Levery, S. B.; Hakomori, S. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 2844–2848.
8. Kannagi, R.; Cochran, N. A.; Ishigami, F.; Hakomori, S.; Andrews, P. W.; Knowles, B. B.; Solter, D. *EMBO J.* **1983**, *2*, 2355–2361.
9. Ohba, H.; Bakalova, R.; Moriwaki, S.; Nakamura, O. *Cancer Lett.* **2002**, *184*, 207–214.
10. Kannagi, R.; Fukushi, Y.; Tachikawa, T.; Noda, A.; Shin, S.; Shigeta, K.; Hiraiwa, N.; Fukuda, Y.; Inamoto, T.; Hakomori, S.; Imura, H. *Cancer Res.* **1986**, *46*, 2619–2626.
11. Hearst, M. A.; Schoelkopf, B.; Dumais, S.; Osuna, E.; Platt, J. *IEEE Intell. Syst.* **1998**, *13*, 18–28.
12. Yagi, F.; Miyamoto, M.; Abe, T.; Minami, Y.; Tadera, K.; Goldstein, I. J. *Glycoconjugate J.* **1997**, *14*, 281–288.
13. Hindsgaul, O.; Cummings, R. D. In *Essentials of Glycobiology*; Varki, A., Cummings, R., Esko, J., Freeze, H., Hart, G., Marth, J., Eds.; Cold Spring Harbor Laboratory Press, 1999; pp 41–56.
14. Hizukuri, Y.; Yamanishi, Y.; Hashimoto, K.; Kanehisa, M. *Genome Informatics* **2004**, *15*, 69–81.
15. Zheng, Z.; Venkatapathy, S.; Rao, G.; Harrington, C. A. *Leukemia* **2002**, *16*, 2429–2437.

16. 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.; Lander, E. S. *Science* **1999**, *286*, 531–537.
17. Doubet, S.; Albersheim, P. *Glycobiology* **1992**, *2*, 505.
18. Doubet, S.; Bock, K.; Smith, D.; Darvill, A.; Albersheim, P. *Trends Biochem. Sci.* **1989**, *14*, 475–477.
19. Park, K. J.; Kanehisa, M. *Bioinformatics* **2003**, *19*, 1656–1663.
20. Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. *Bioinformatics* **2000**, *16*, 906–914.
21. Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. *Comput. Chem.* **2002**, *26*, 293–296.
22. Cai, C. Z.; Wang, W. L.; Sun, L. Z.; Chen, Y. Z. *Math. Biosci.* **2003**, *185*, 111–122.
23. Brown, M. P.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M., Jr.; Haussler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 262–267.
24. Pavlidis, P.; Wapinski, I.; Noble, W. S. *Bioinformatics* **2004**, *20*, 586–587.
25. Moriwaki, S.; Ohba, H.; Nakamura, O.; Sallay, I.; Suzuki, M.; Tsubouchi, H.; Yamasaki, N.; Itoh, K. *J. Hematother. Stem. Cell Res.* **2000**, *9*, 47–53.