Classification and Motif Extraction of Glycans in Bloods

Yoshiyuki Hizukuri[1] yosh@scl.kyoto-u.ac.jp
Yoshihiro Yamanishi[1] yoshi@kuicr.kyoto-u.ac.jp
Osamu Nakamura[2] osamu.nakamura@aist.go.jp
Fumio Yagi[3] fyagi@chem.agri.kagoshima-u.ac.jp
Susumu Goto[1] goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

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

Keywords: glycan structure comparison, support vector machine, glycan classificaiton, glycan motif

1 Introduction

Glycans are an important class of biological macromolecules in addition to DNAs and proteins. Actually, glycans play key roles in cellular functions including cell-cell communications, protein interactions and immunity. In this paper, we conducted a comprehensive analysis on comparative glycomics using glycan structures stored in the KEGG/GLYCAN database [2]. First, we developed a new similarity measure for comparing glycan structures and tested its ability to classify glycans of different blood components in the framework of Support Vector Machine (SVM). The result shows that our method successfully classified glycans from different human blood cells. Next, we extracted characteristic functional units (motifs) of glycans suspected to be substructures specific to each blood component. Finally, we conducted an experiment based on agglutination assay in order to verify the result of our prediction, and we confirmed that the fungal lectin specifically recognized the glycan motif predicted by our method.

2 Materials and Methods

All glycan structures used in this study were obtained from the KEGG/GLYCAN database and the corresponding annotations of biological sources were obtained from the CarbBank/CCSD database [1]. We used glycan structures of four human blood components: leukemia, erythrocyte, serum and plasma, and the numbers of corresponding glycan structures were 162, 112, 85 and 73, respectively.

In the algorithm we propose to decompose the tree structures of glycans into sets of substructures (3-mers in this study), because there is an observation that many glycosyltransferases physically interact with about three linked monosaccharides. Suppose that we have two glycans $\mathbf{x}$ and $\mathbf{y}$, and decompose the glycans $\mathbf{x}$ and $\mathbf{y}$ into sets of 3-mers. 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 substructures of $\mathbf{x}$ (resp. $\mathbf{y}$). Considering the match and mismatch of the substructures between glycans $\mathbf{x}$ and $\mathbf{y}$, we define the similarity as $Sim(\mathbf{x}, \mathbf{y}) = \frac{\#\{\mathbf{x} \cap \mathbf{y}\}}{\#\{\mathbf{x} \cup \mathbf{y}\}}$, where the numerator is the number of the common substructures, and denominator is the number of the unique substructures between $\mathbf{x}$ and $\mathbf{y}$. More details of the technical points can be found in our previous work [3]. We use this similarity measure as a kernel function in the SVM classifier as $\mathrm{K}(\mathbf{x},\mathbf{y}) = \mathrm{Sim}(\mathbf{x},\mathbf{y})$.

In order to detect characteristic substructures that are specific to each blood component, we use the discriminant score of the SVM, because the score reflects the difference between the target group

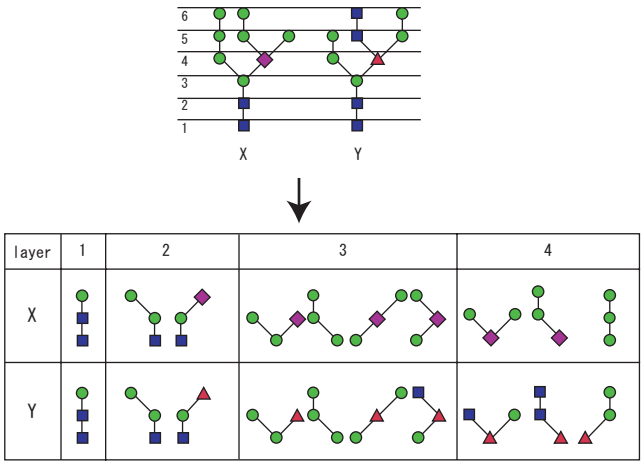

Figure 1: An illustration of the decomposition of glycan structures. Each node represents a monosaccharide. The number indicates the layer defined by the distance of each substructure from the root.

and the others. For each substructure, we take a summation of all the discriminant scores over the whole glycan set, and use it as the characteristic score.

3 Results and Discussions

For each blood component, we applied the SVM to predict whether a glycan is assigned to the component or not. The result of the Jackknife cross-validation tests showed high accuracy (around 80

Next, we extracted glycan motifs specific to each blood cell by selecting high scoring substructures based on the approach explained in the previous section. For a certain blood component, we predicted the substracture of a characteristic glycan motif. To verify the prediction, we finally conducted an experiment based on agglutination assay using *Agrocybe cylindracea* galectin (ACG) [4]. In the experiment, it was observed that the ACG distinguished the target blood cell from the other cells in vitro. This result suggests that our method successfully extracted an informative glycan motif.

4 Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, 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] Doubet, S., Albersheim, P. CarbBank, *Glycobiology.*, 2(6), 1989.

[2] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 277–280, 2004.

[3] Hizukuri, Y., Yamanishi, Y., Hashimoto, K., and Kanehisa, M. Extraction of species-specific glycan substructures. *Genome Informatics*, 15:93–104, 2004.

[4] Yagi F, Miyamoto M, Abe T, Minami Y, Tadera K, Goldstein IJ. Purification and carbohydrate-binding specificity of Agrocybe cylindracea lectin, *Glycoconj J.*, 14(2):281–288, 1997.