Global Tree of Known Carbohydrate Structures to Analyze Biosynthetic Pathways

Kosuke Hashimoto[1]
khashimo@kuicr.kyoto-u.ac.jp

Shin Kawano[1]
kawano@kuicr.kyoto-u.ac.jp

Yasushi Okuno[2]
okuno@pharm.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] Graduate School of Pharmaceutical Sciences Department of Genomic Drug Discovery Science, Kyoto University 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Keywords: carbohydrate structure, global tree, glycoinformatics, KEGG GLYCAN

1   Introduction

Carbohydrates, following genes and proteins, play important roles in many biological processes. Recently, much knowledge about carbohydrate structures and functions has become clear because of the improvement of experimental techniques such as MS spectrometry and knockout mice analysis. Some databases of carbohydrate structures for bioinformatics have been developed, prompting the start of the "Glycoinformatics" field. There has been some research on the analysis of carbohydrates using protein data. For example, the positions of the proteins attached to a carbohydrate [2,4] has been investigated. However, a comprehensive analysis of carbohydrate structures has not been investigated. In this study, we constructed a global tree of all carbohydrate structures in the KEGG GLYCAN database [3]. Our first step involved the definition of "one link" as the relationship between two carbohydrate structures; one carbohydrate structure becomes another carbohydrate structure after the transfer of a monosaccharide by a glycosyltransferase reaction. Next, the carbohydrate structures of "one link" relations are connected to each other. In addition, we mapped the publicly known core structures as well as the well-known biosynthetic pathway of N-Glycan onto the tree. In this way, the global tree is able to clarify the overall picture of all the carbohydrate structures, to contribute to the future of glycoinformatics research.

2   Method

The global tree was constructed from the carbohydrate structures derived from the KEGG GLYCAN database. First, the entries that have a bond between the reducing ends of two monosaccharides were removed. Next, non-carbohydrate molecules, such as amino acids and lipids, were removed from the data set. Modification molecules, which are attached to the monosaccharide after carbohydrate biosynthesis (ex. methyl group and O-acetyl group) were also deleted from the data set. Finally, the redundant carbohydrate structures were pruned to result in our final non-redundant data set of carbohydrate structures. Using this data set, we calculated the "one link" structures. The structures and their relationships were described as nodes and edges, respectively. The global tree was drawn with the network-drawing tool Pajek [1].

3   Result and Discussion

3.1   Construction of the global tree

The non-redundant data set contains 7696 carbohydrate structures obtained from the KEGG GLYCAN database (10082 entries). After the connection of "one link" structures, 4713 carbohydrate structures related to each other could be represented in one tree (Fig. 1). This tree covers 61.2% of the carbohydrate structures in the non-redundant data set. Fig. 1 also shows groups of carbohydrate structures that contain the

well-known core structures of N-glycan (GlcNAc b1-4 GlcNAc), O-glycan (Gal b1-3 GalNAc) and glycolipid (Gal b1-4 Glc). Half of the carbohydrate structures in the global tree could be characterized by core structures.

3.2   The carbohydrate biosynthetic pathways

The global tree also reflects the biosynthetic pathways of carbohydrate sugar chains. In the tree, each "one link" carbohydrate structure, which corresponds to a glycosyltransferase reaction, is connected. In other words, a path from the root node to any node represents a biosynthetic pathway of the carbohydrate structures. Fig. 2 shows the biosynthetic pathway of N-glycan, which is synthesized in the following process of three steps: synthesis of a core structure (line 1), trimming of it (line 2) and synthesis of a leaf structure connected to the core structure (line 3). The synthetic pathway of the core structure (line 1) and the degradation pathway (line 2) is strongly conserved among many eukaryotes, and the global tree has comparatively few branches. In contrast, the biosynthetic pathway of leaf structures (line 3) is more complex. This suggests that N-glycan provides various leaf structures corresponding to their functions and/or the variety of species.
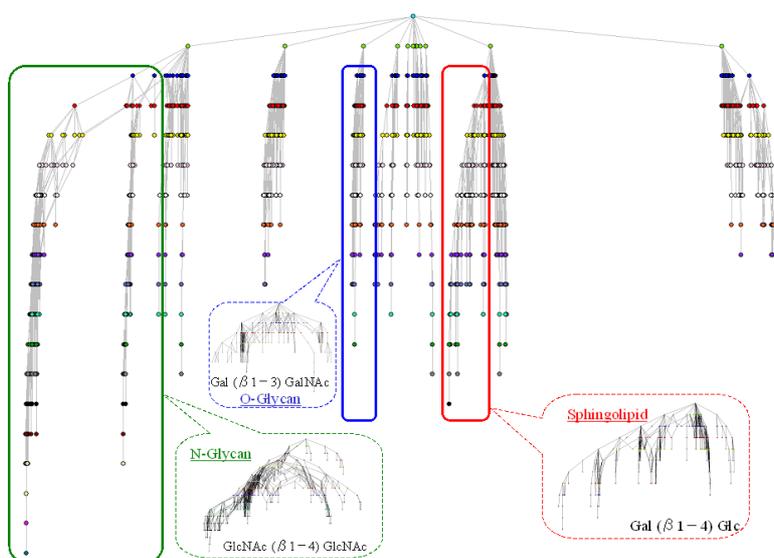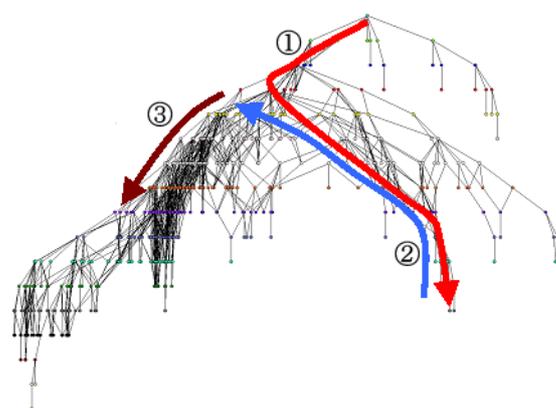

Figure 1: The global tree

Figure 2: The Pathway of N-glycan

Acknowledgments

References

[1] Batagelj V. and Mrvar, A., Pajek - Program for Large Network Analysis, *connections*, 21(2):47-57, 1998.

[2] Ben-Dor, S., Esterman, N., Rubin, E. and Sharon, N., Biases and complex patterns in the residues flanking protein N-glycosylation sites, *Glycobiology*, 14(2):95-101, 2004.

[3] Hashimoto, K., Hamajima, M., Goto, S., Masumoto, S., Kawasima, M. and Kanehisa, M., GLYCAN: The Database of Carbohydrate Structures, *Genome Informatics,* 14: 649-650, 2003.

[4] Julenius, K., Mølgaard, A., Gupta, R. and Brunak, S., Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites, *Glycobiology*, 14(2):103-114, 2004.