A Graphic User Interface for the Glyco-MGrid Portal System for Collaboration among Scientists

YAN-JI ZHAO[1], SUNG-RYUL KIM[2]
[1]Department of Advanced Fusion Technology
Konkuk University
Seoul
KOREA

[2]Department of Internet and Multimedia Engineering
Konkuk University
Seoul
KOREA

Abstract: - Recently there are several reports that carbohydrate molecules take important biological roles in intercellular signal transduction, cell growth and differentiation. As a consequence, the analysis of carbohydrate sugar chains, or glycans are becoming one of the major problems in computational molecular biology. Collaboration among scientists are difficult because of the tree-like structure of glycans, which makes searching much more difficult than proteins or DNA. This paper presents a graphical user interface for MGrid-based Glyco-MGrid portal system, which is currently still in development, that supports simulation database for glycoconjugates. We also discuss a linear notation for glycan structure and an intuitive method to perform glycan structure matching.

Key-Words: - Carbohydrate sugar chain, Glycan, User interface, Glyco-MGrid system, Linear notation, Glycan structure matching

1   Introduction

Recently there are several reports that carbohydrate molecules take important biological roles in intercellular signal transduction, cell growth and differentiation. Thus the analysis of carbohydrate sugar chains, or glycans are becoming one of the major problems in computational molecular biology. Glycans or sugar chains have branched tree structures rather than a linear chain, and the connection between carbohydrate residues shows significant structural heterogeneity, varying, for example, in position and anomeric (α or β)configuration[3]. Thus, the modeling and structured matching of primary structural features is considerably more difficult for glycans than for genes and proteins.

In this paper, we propose a graphical user interface for a portal system called Glyco-MGrid, which is currently in development by Konkuk University. The Glyco-MGrid environment is an MGrid-based grid portal which supports simulation and data management for glycoconjugates in a collaborative way. As a visual structure search part of Glyco-MGrid system, we have designed graphical user interface which enables users to enter the glycan structure easily. We also discuss an algorithm which performs tree-structure matching of glycan on database.

Users can perform full or partial search for glycans via the Glyco-MGrid system.
(1) Using the interface, user can input glycan structure directly and visually instead of textually,
(2) sends this structure in a canonical format which will discuss at section 4 to the algorithm and
(3) visualizes the resulting alignment.

2   Related works

There are several carbohydrate database systems which have been built in the past, such as CarbBank/CCSD, GlycoSuiteDB and KEGG Glycan.
The CCSD is a database containing complex carbohydrate structure and associated text. The database has a flat file format, i.e., one record contains

a single structure with its associate text and citation[4,6,9].

GlycoSuiteDB is a relational database that curates information from the scientific literature on glyco-protein derived glycan structures, their biological sources, the references in which the glycan was described and the methods used to determine the glycan structures [10].

However, the two databases mentioned above, their query for structure search are not exactly intuitive in that the glycan tree structure often have to be entered textually.

The KEGG GLYCAN, can retrieve information about glycans and analyze the glycan structure through KCaM. However, because of the elementary methods used for scoring, the biological retrieval of the alignments as presented is not necessarily clear[7].

3 Glyco-MGrid System

Most of the scientists perform simulation tasks alone and share only the final analysis results with others, because there are few integrated and collaborative molecular database systems. The MGrid system is a grid computing environment to support computation and data management in a single platform which aims at developing a grid computing system for collaborative molecular simulation.

e-Glycomics is an advanced computer technology based research approach to glycomics with uses molecular simulation, molecular sequencing and bioinformatics. The Glyco-MGrid environment is the MGrid-based cyber-infrastructure for e-Glycomics whose design is focused on molecular simulation. Glyco-MGrid is intended to provide a grid portal-based shared and integrated environments for e-Glycomics which support simulation, databases, and analysis in a collaborative way [6].

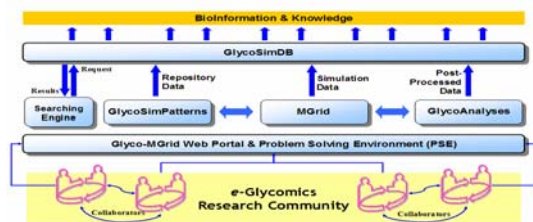The Glyco-MGrid is composed of GlycoSimPatterns, GlycoAnalyses, GlycoSimDB, and other MGrid components.

Fig. 1. System architecture of Glyco-MGrid

GlycoSimPatterns is one of the template repositories in the Glyco-MGrid components to deposit typical simulation input scripts for molecular simulation of glycoconjugates. Clients are able to use typical input template in the GlycoSimPatterns and edit the input script suitable for their own simulation conditions.

GlycoSimDB is database components storing overall input template, simulation output, trajectory files, and post-processed data for each glycol-component. Research collaborators are able to access to GlycoSimDB components and then execute resimulation or re-processing of simulation data with trajectory files done by other researchers under e-Glycomics approach.

Glyco-MGrid system can execute the simulation jobs automatically and then molecular dynamics (MD) trajectory files of many glycol-compounds are accumulated at a Data Grid. This trajectory files can be shared and transferred to local client machine upon user's request. GlycoAnalyses component is specially designed analysis facility for the post-processing of carbohydrate MD trajectory files. Users are able to generate both text-type processed data and graph-type results with GlycoAnalyses component.

The Glyco-MGrid system currently provides job-submission, scheduling, computation facility in the Glyco-MGrid web portal. The system is undergoing extensive tests and is being used by molecular biology researchers.

4 Structure search for glycans

4.1 The function of the glycan structure search
There are many jobs for scientists who want to share the result and wish to do comparative study via the system. When perform their jobs, they may input the structure query by fully or partially, and it is inconvenient to users use textual form. Thus, the visual search takes important role in Glyco-MGrid system. The functions of glycan structure search are listed below.
(1) A researcher may tag their results or jobs with the glycan structure of interest through the visual structure search editor and can submit the job to the computational Grid in the MGrid.
(2) They may search for tags that already exist in the system by partial or similar matching.

(3) Any collaborator can also perform partial search or structure similarity search of glycans and manipulate the data for further analysis.

4.2 The structure of glycan

Unlike the linear structures of DNA/protein, glycans are branched tree structure with various types of linkages[2]. The nodes and edges of the tree correspond to monosaccharides and glycosidic-bones [1]. The monosaccharide structures of glycans are generally smaller, branched trees of 10~15 monosacchairdes. Furthermore, the edge contains information on the linkage position and anomality, because for glycosyl bond, there are six possible hydroxyl groups and two possible anomality α or β [5, 9].

We use a set of monosaccharides which showed in Table 1 for glycan structure search which commonly can found in multicellular organisms. We also use a set of linkage information which showed in Table 2 such as anomeric types and to which hydroxyl of which nodes they are linked.

Monosaccharide	Abbreviation	Monosaccharide	Abbreviation
Glucose	Glc	Fructose	Fru
N-acetylglucosamine	GlcNAc	L-Iduronic acid	LIdoA
		3-Deoxy-D-manno-oct-2-ulosonic acid	Kdo
Glucuronic acid	GlcA		
Galactose	Gal		
N-acetylgalactosamine	GalNAc	L-Arabinose	LAra
		N-Glycoloylneuraminic acid	Neu5Gc
Mannose	Man		
L-Fucose	LFuc		
N-acetylneuraminic	Neu5Ac	Glucosamine	GlcN
Xylose	Xyl	Galacturonic acid	GalA
L-rhamnose	LRha		

Table 1. monosaccharides and abbreviations

Linkage type	Linkage type
a1-2	1-b1
a1-3	1-a2
a1-4	2-a1
a1-6	2-b1
b1-1	3-a1
b1-2	3-a2
b1-3	3-b1
b1-4	4-a1
b1-6	4-b1
a2-1	6-a1
a2-3	6-a2
a2-6	6-b1

Table 2. Linkage type

4.3 User Interface for glycan structure edit

The user interface we have developed is a Java application for drawing glycan structures in a way similar to KCaM. It allows users to input a glycan structure directly into the applet with their mouse. A user can enter each carbohydrate structure as a graph and the graphical nodes and edges represent mono sugars and glycosidic-bonds respectively.

A user can insert glycan structure easily via the user-friendly Glyco-MGrid web portal. Figure 2 and Figure 3 shows screen shots of the Glyco-MGrid web portal and glycan structure editor respectively. At the searching part of the portal, users can choose the Glyco-structure item and push the "Visual search" button to display the structure editor. This editor consists of a panel that contains three button icons and a canvas of drawing area. Icons from first to third represent node, edge and eraser respectively. Each icon corresponds to edit command of glycan structure editor. The node icon is for entering nodes, the edge icon is for entering glycosidic linkages and the eraser icon is for erasing an object.
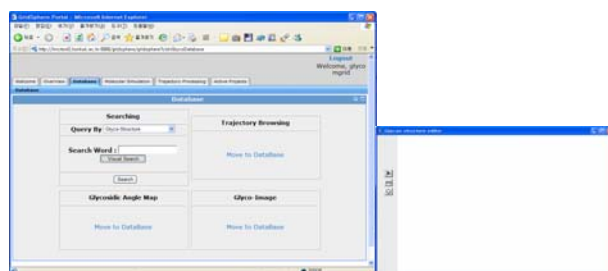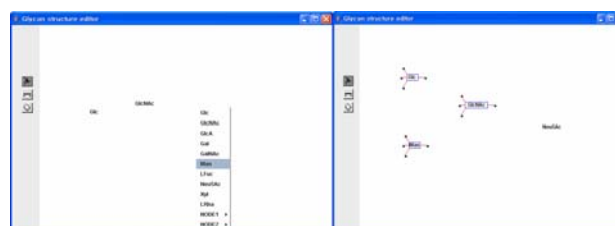

Fig. 2. Glyco-Mgrid web portal   **Fig.3.** structure search editor

Figure 4 shows how to use the glycan structure editor.

a. enter monosaccharides   b. selection of monosaccharide

c. entering edge   d. selection of linkage information

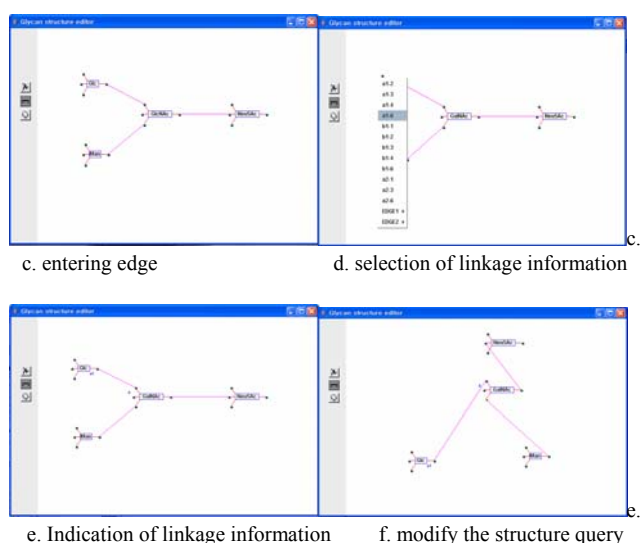e. Indication of linkage information   f. modify the structure query

Fig.4. Insertion of glycan structure query

(1) Enter a node: A user can press the right button of mouse after clicking down the node icon to popup a popup menu for mono sugars, and select an item among them (a on Figure 4).

(2) Select the node: Every time the user pressing mono sugars with the mouse, nodes can be selected within a specified rectangular boundary with four different directions of ports (b on Figure 4).

(3) Enter an edge: For entering the edge between two mono sugars, first, you have to click the edge button. Second, you press a port in each two nodes, which you wish to connect (c on Figure 4).

(4) Indicate glycosidic linkage information: The edges also contain information such as to which hydroxyl (carbon number) of which nodes they are linked and their linkage type, or anomer (i.e., α or β) [4]. Once you click the port which connected with the line, a popup menu that contains linkage information will popup, and then you can select an item among them. The information of the selected item will show up on the bottom-left side of the port (e on Figure 4).

(5) Erase the object: choose the erase icon, and select the node you want to erase.

(6) Modify the structure: you can drag the node for modification (f on Figure 4).

4.4 Efficient full and partial search

It is very inefficient for users to use the textual form to input structure. The complicate structure makes it

difficult to input full structure in textual as well as perform partial match.

While the user uses graphical interface, internally we define a linear notation of glycan structure that will be used for full and partial search for glycan as follows.

(1) In each monosaccharide molecule we consider, there is one unique reducing end. We label it with 'R'. Also, there can be at most three non-reducing ends, we label them by '1', '2', '3'. That is, each monosaccharide will have unique labelings of their non-reducing ends.

(2) In a glycan structure, the unique free reducing end is designated as the root, and a molecule linked with non-reducing end '1', is enclosed with(…), linked with non-reducing end '2' is enclosed with {…}, and linked with non-reducing end '3' is enclosed with […] respectively. Substructures of these molecules are represented inside the parenthesis, respectively.

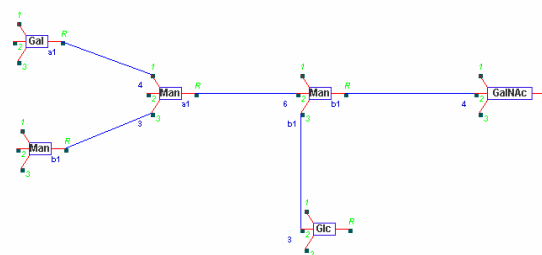Figure 5 shows the representation for a N-glycan fragment.

Fig.5. representation for a N-glycan fragment.

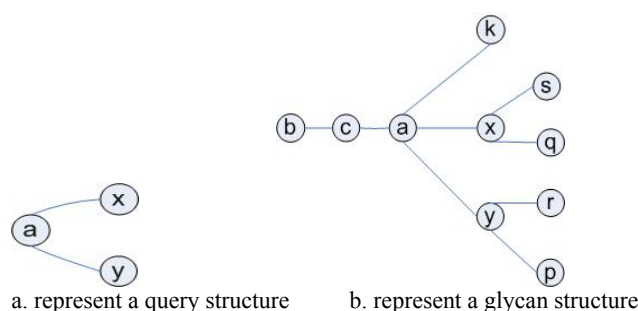Result: *GalNAc (1->4, Man (1->6, Man(1->4,Gal) {1->3,Man} []) {1->3,Glc} []) {} []*

a. represent a query structure   b. represent a glycan structure

Fig.6. a query and a glycan structure

Query: a () {x () {} []} [y () {} []]
glycan: b () {c () {a (k) {x (s) {q} []} [y (r) {p} []]}
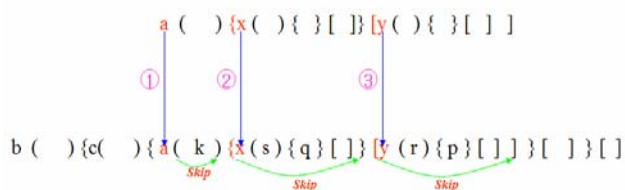   []} []

Fig.7. searching process

See Figure 7 for a description of our algorithm.
(1) We first look for 'a' in the big molecule. This will define where we have to look for x and y exactly, by matching the parenthesis. For efficiency each matching parenthesis is connected by a pointer in the representation.
(2) Find 'a' and skip over (k), using the pointer.
(3) Then find 'x' and skip (s), {q}, []. Finally find 'y'.
(4) We will repeat the above for all 'a' appearing in the large molecule. Efficiency is guaranteed because we skip over irrelevant portions in each search.
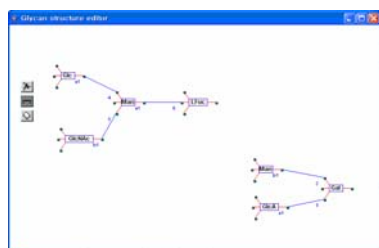

Fig. 8. entering AND query

Because the glycan structure is complicate tree structure, sometimes the user need to search the structure with more than one partial structure. Figure 8 shows entering a kind of AND query. Our algorithm can be easily modified to search for molecules including both parts as shown in Figure 8.

5 Conclusion and Future work

In this paper, we proposed graphical user interface and discussed algorithms for glycan structure search for the Glyco-MGrid system. The user interface we designed provides convenience to users with glycan structure entry and visualization. The canonical forms are very intuitive to conserve the glycan structure to standardized format, and the algorithm which is using recursive function can be performed in linear time.

The glycan structure search with Glyco-MGrid system is still in an early stage of development. Project is in a progress to implement the algorithm and the simulation result database for glycan chains has being constructed by the research group.

References:
[1] Hashimoto, K., Hamajima, M., GLYCAN: The Database of Carbohydrate structure, *Genome Informatics,* Vol.14, 2003, pp. 649-650.
[2] Aoki, K.F., Yamaguchi, A., KCaM(KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains, *Nucleic Acids Res*, Vol.32, 2004, pp. 267–272.
[3] Aoki, K.F., Yamaguchi, A., Efficient Tree-matching Methods for Accurate carbohydrate Database Queries, *Genome Informatics*, Vol.14, 2003, pp. 134-143.
[4] Hizujuri, Y., Yamanishi, Y., Extraction of species-specific glycan substructures, *Genome Informatics,* Vol.15, 2004, PP.69-81
[5] http://www.genome.jp/ligand/kcam/
[6] http://biol.lancs.ac.uk/gig/pages/gag/carbbank.htm
[7] Aoki, K.F., Mamitsuka, H., Akutsu, T., A score matrix to reveal the hidden links in glycans, *Bioinformatics,* Vol.21, 2005, pp.1457-1463.
[8] http://biol.lancs.ac.uk/gig/pages/gag/carbbank.htm
[9] C.A., joshi, H.J., Harrison, GlycoSuiteDB: a Curated relational database of glycoprotein glycan structures and their biological sources, *Nucleic Acids Res*, 2003, Vol.31, pp.511-513
[10] Bohne-Lang, A., E., Forster, LINUCS: Linear Notation for Unique description of Carbohydrate Sequences, *Carbohydr Res*, Vol.336, 2001, pp.1-11