Journal of Computer Science & Systems Biology - Open Access
JCSB/Vol.2 September-October 2009
Review Article OPEN ACCESS Freely available online doi:10.4172/jcsb.1000040

Glycomics Data Mining

V Srinivasa Rao[1]*, S K Das[2], E Kusuma Umari[3]

[1]Professor in Computer Science and Engineeing, Pvp Siddhrtha Institute of Technology, Vijayawada-520 007, India, E-mail: akrgvsr@gmail.com

[2]Deparment of Computer Science, Berhampur University, Berhampur, Orissa, India

[3]Associate Profesor in Electronics and Communication Engg, Nova College of Engineering, Jangareddygudem, India

Abstract

The amount of glycomics data being generated is rapidly increasing as a result of improvements in analytical and computational methods. Correlation and analysis of this large, distributed data set requires an extensible and flexible representational standard that is also 'understood' by a wide range of software applications. An XML-based data representation standard that faithfully captures essential structural details of a glycan moiety along with additional information (such as data provenance) to aid the interpretation and usage of glycan data, will facilitate the exchange of glycomics data across the scientific community. We reviewed importance of data warehouse, showing a method of applying data mining techniques using XML, and some of the data issues, analysis problems, and results.

Introduction

Carbohydrates are considered as the third class of information encoding biological macro molecules.Glycomics is the scientific attempt to characterize and carbohydrates,for which informatics is just beginning. Glycomics requires sophisticated algorethemic approaches.several algorithms and models heve been developed in this field.The four essential molecular building blocks of cells are proteins,nucleic acids,lipids and carbohydrates are often referred to as glycans.Nucleotide and protein sequences are very important for all bioinformatics applications and research ,whereas glycan and lipid structures have been widely neglected in bioinformatics.

Glycans are the most abundant and structurally diverse biopolymers found in nature.These glycans bound to proteins as glycoproteins and are known to affect the function of proteins.More than half of the protein sequences deposited in the swiss-prot database include potential glycosylation sites and thus may be glycoproteins.Based on an analysis of well annotated and characterized glycoproteins in swiss-prot ,it was known that half of the proteins are glycosylated.development in this field is still in the early stages.many algoriyhms are being developed.major glcosylated projects such as consortium for functional glycomics, KEGG glycan, GLYCOSCIENCES.de are being developed and well structured glycol-related data that are awaiting for data mining and analysis.Complex carbohydrates are often called as glycans and are often found attached ally to proteins and lipids.glycoproteins are usually on the cell surface ,where they are recognized by bacteria,viruses,and other proteins,such as lectins,in order to facilitate various functions.it is also known that glycans are involvd in various biological functions such as protein folding and signaling events.It is known that glycan specific pathways are responsible for diseases such as congential disorders of glycosylation are caused due to the defects in these pathways.There are reports on glycan biomarkers related to human diseases such as cancer and autoimmune diseases.

Carbohydrates are composed of monosacharides that are covalently linked by glycocidic bonds, either in alpha or beta form.in order to formulate a standardized notation to glycans ,the Consortium for Functional Glycomics(CFG) proposed a standard symbolic representation for those monosacharides which are found in nature.carbohydrates are most classically drawn as a tree in a two-dimensional plane,with the root monosaccharide placed at the right most position children branching towards the left.each node represents a monosaccharide,and each edge represents a glycocidic linkage,which includes the carbon numbers that are bond and the conformation.

The IUPAC-IUBMB has specified the nomenclature of carbohydrates to uniquely describe complex oligosaccharides based on a three letter code to represent monosacharides.each monosaccharide code is preceded by the anomeric descriptor and the configuration symbol.the ring size is indicated by an italic *f* for furanose and *p* for pyranose.There are other formats also like KEGG Chemical Function(KCF) format,which represents glycans using a connected graph LINCUS,which provides a unique and linear notation for glycans ,and linear codes by glycominds,which provides a commercial carbohydrate database. Recently Herget et al., (2008) carefully analyzed the encoding capabilities of all existing carbohydrate sequence formats and the content of publically available structure databases as part of the EUROCarbDB project (www.eurocarbdb.org) .

Recently Tamaki et al., (2009) applied a balancing technique to obtain more appropriate and robust models, and compared

***Corresponding author:** Dr V Srinivasa Rao, Professor in Computer Science and Engineeing, Pvp Siddhrtha Institute of Technology, Vijayawada-520 007, India, E-mail: akrgvsr@gmail.com

Received September 17, 2009; **Accepted** October 21, 2009; **Published** October 25, 2009

Citation: Rao VS, Das SK, Umari EK (2009) Glycomics Data Mining. J Comput Sci Syst Biol 2: 262-265. doi:10.4172/jcsb.1000040

their accuracy with that of the conventional model. To construct new models, they randomly sampled the same number of subjects with and without new dental caries. Hashimoto et al., (2008) developed an efficient method for mining motifs or significant subtrees from glycans.

There are three major databases for the complex carbohydrates, KEGG GLYCAN, and the database developed by the consortium for functional glycomics. All three databases are based on the carb bank database developed in 1990s. The major issue that was facing by the glycol-inforamatics community was that each of these databases represented their glycan structures in different formats, a workshop was held at the National Institutes of Health(NIH), united states ,at this meeting, the GLYDE-2 XML format for glycans and glycoconjugates, developed by the CCRC, was agreed upon as the standard format for exchanging carbohydrate .

Apart from the development of these various databases, several methods for analyzing the structures have been developed. They are classifies into 6 types.

1.Glycosylation analysis
2.Glycomics
3.Glycan biomarker prediction
4.Glycan structure analysis
5.Glyco-gene structure analysis
6.Glycomics data mining

Glycoproteomics

Glycosylation is a post translational modification.Half of the proteins are glycosylated. Role of the glycans depends on N- or O-linked forms. Glycan is required for biological function of proteins, such as the FC-efector function of IgG, the regulated clearance of glycochrome lutropin. Glycans also serve as recognition targets for their complimentary binding proteins,the lectins.lectins involves in fertilization, development, differentiation and lymphocyte circulation and also in the development of pathological conditions. Glycans are also used for the attachment of microbes,toxins and constitute the first point of interaction between microbes and glycans. The function of glycan depends on which the protein is attached to so, it is very difficult to document about glycoproteins. The integrated structure of compiling structure and function of glycans is called glycoproteins.Biosynthesis of N-Glycans and O-Glycans involves many glycotransferases. In the early steps of glycotansfer many steps are obligatory but there will be competition of steps in the later stages.some well known changes in glycan structures occurs during the neoplastic transformation result from changes in the balances of transferases. Various aspects of structure and function relationship of glycoproteins were discussed, [refer Glycoproteomics paper].

HIV virus contains gp120/gp40 glycan protein complexes on the outer surface. The surface of g120 contains several clusters of mannose. Scientists discovered an antibody G12 which can neutralize the infection. This antibody narrows the distance between binding pockets so that the virus cannot get attached to the host cd4. Another antiviral protein, actinohivin which is a lectin that contains 3-mannose binding sites and it is active against simian immunodeficiency virus and HIV virus. N-glycans of gP120 can also be exploited as targets for therapeutics. Many glycoprotein therapeutics have been expressed using various production vehicles such as cell lines, transgenic animals and plants.

Mass spectroscopy strategies are suitable to asses the type of glycan present and their position of attachment on a particular protein.Multi-ms strategy is used for the analysis of simple or complex mixtures from biological extract. For a detailed analysis several fluroscence-based methods as well as direct analysis by dionex system is available.

Glycome db

It is the integration of all available carbohydrate sequences into one central database called Glycome db. Approximately 100000 database records with 73341 sequences are accessible in the public domain. The biggest obstacle for data integration was the use of various sequence encoding formats by different initiatives.so, they found it necessary to develop a new sequence format, called GlycoCT, which is a superset of the structural encoding capabilities inherent in all other formats developed so far. They have implemented a translation library which can read all of the carbohydrate sequence formats and translate them to GlycoCT. GlycoUpdateDB is the application program which have to be designed to carry out the integration of the interpretable data obtained from the resources. It is a JAVA application, Depending on a PostgreSQL database which can be configured by an XML file. The configuration file contains settings for the local database and instructions for the download and data integration process. GlycomeDB consists of several database schemata with tables that store all downloaded and generated datasets. A software tool-GlycanBuilder for building and displaying glycan structures using a chosen symbolic notation was reported by Ceroni et al., (2007). The tool uses an automatic rendering algorithm to draw the saccharide symbols and to place them on the drawing board. The information about the symbolic notation is derived from a configurable graphical model as a set of rules governing the aspect and placement of residues and linkages. The algorithm is able to represent a structure using only few traversals of the tree and is inherently fast. The tool uses an XML format for import and export of encoded structures.

Carbohydrates are involved in a variety of fundamentalbiological processes (cellular differentiation, embryonic development, fertilization) and pathological situations (bacterial and viral infections, inflammatory diseases,cancer). They therefore have a large pharmaceutical and diagnostic potential. Protein-carbohydrate interactions are intensively investigated using a variety of experimental methods. Among these, X-ray and NMR measurements provide a detailed 3D picture of the spatial location of the ligand as well as the protein. It is estimated that more than half of all the proteins in the human body have carbohydrate molecules attached. Structural data taken from X-ray crystallography does not necessarily mean that a potential glycosylation site is unoccupied; but the presence of carbohydrate residues in the 3D structures provides direct and unambiguous evidence for the occupancy of a glycosylation site.

Therefore, these data have been intensively used to gain deeper insights into factors that regulate glycan attachment to a N-glycosylation site. Among the 25667 PDB entries structures were detected which contain a total of 6714 carbohydrate chains.

About half of them are covalently attached, the other half belongs to non-covalently bound ligands. It was found that about 30% of all PDB entries containing carbohydrates comprise one or several errors in glycan description, which are mainly due to wrong assignment of saccharide units. The reason for this unacceptable high rate of errors is obvious. Sequences for complex carbohydrates differ significantly from the simple linear one-letter code used to describe genes and proteins:

a) the number of naturally occurring residues is much larger for glycans

b) each pair of monosaccharide residues can be linked in severalways,

c) a residue can be connected to three or four others (branching).

Sugar resides present in the PDB are defined in the so-called HET Group Dictionary. Since a three-letter code is used to uniquely assign carbohydrates, a new residue name is required for each stereochemically different sugar unit and each substitution. This procedure makes correct assignment of sugar units tedious, complicated and obviously error-prone. Unfortunately, no software is currently available which automatically aligns the 3D information contained in PDB and the given assignments. It is the focus of pdb-care to fill this gap. The pdb-care program is based on the carbohydrate detection software pdb2linucs that is able to identify and assign carbohydrate structures using only the reported atom types and their 3D atom coordinates. To be able to compare the detected carbohydrate structures in LINUCS notation to the residue assignments as reported in the PDB HET Group Dictionary a translation table in XML format was created, where both descriptions are confronted. Three types of residues are included: monosaccharides, oligosaccharides and combine residues consisting of a carbohydrate moiety and a non-carbohydrate part. The translation table actually contains 141 monosaccharides, 31 oligosaccharides and 77 combined residues and is still growing. The pdb-care protocol reports the type of problems, inconsistencies and errors detected. pdb-care is written in C. Interaction with the user is done through a web-interface, which is implemented in PHP. The pdb-care service is hosted at the central spectroscopic department of the German Cancer Research Centre in Heidelberg, Germany. The pdb-care web interface allows either to analyse a file obtained directly from PDB using the PDB-ID, or to provide a pdb-file located on the local computer by upload or by copy paste into the provided input window. Since carbohydrate structures are described as so called hetero atoms within the HETATM records, all data assigned to the ATOM records (amino acids, nucleotides) are neglected. Sahoo et al., (2005) introduced GLYcan Data Exchange (GLYDE) standard as an XML-based representation format to enable interoperability and exchange of glycomics data. An online tool for the conversion of other representations to GLYDE format has been developed.

KEGG

The KEGG database was initiated in 1995. After 10 years of development the KEGG project has entered a new phase in accordance with the chemical genomics activity. KEGG database is a resource for understanding the higher order functions and the utility of the biological systems such as cell or the organism from genomics and molecular information.KEGG is considered as the computer representation of the biological system consists of building blocks and wiring diagrams which can be used as simulation and modelling as well as for browsing and retrieval.originally, the wiring diagrams involved endogenous molecules, both those that are directly encoded in the genome and those that are indirectly encoded through biosynthetic/biodegradation pathways. These wiring diagrams also include exogenous molecules.this will help to understand the interactions between environment and the biological system.there are four types of databases in KEGG. They are gene database,brite database, ligand database and pathway database. KEGG system can be represented in two types of graphs. Nest graph in which nodes can themselves be graphs line graph which is derived by interchanging nodes and edges of other graphs.

KEGG is the major component of the Japanese GenomeNet, which is served by the Kyoto University Bioinformatics Center. The other GenomeNet services including DBGET and BLAST/FASTA searches are now primarily developed and used to support KEGG. The KEGG API service has become an increasingly popular mode of access. It is the SOAP/WSDL interface to KEGG, enabling users to write their own programs to access, customize and utilize KEGG. KegArray and KegDraw are standalone Java applications that make use of the KEGG resources. KegArray is for microarray data analysis in conjunction with KEGG pathways and genomes. KegDraw is for drawing glycanstructures and chemical compound structures, which can then be used to query against KEGG and PubChem databases.

Mass Spectrometry

High resolution mass spectrometry is one of those biotechnologies that are highly promising to improve health outcome. Proteomics biomarkers that can distinguish that can distinguish healthy patients from cancer patients have been identified using MS data. Ms study is demonstrated which uses glycomics to identify ovarian cancer.mass spectrometry is used for protein profiling in cancer on the peptide or protein abundance from the MS data. Recent literatures on cancer classification using MS have identified some potential protein biomarkers in serum to distinguish cancer from normal samples.since glycans play important roles in cell communication and signaling events they may be implicated in cancer.Clinical glycomics is used to identify potential biomarkers for the early learning algorithms to classify high dimensional MS cancer data. Artificial neural networks are used to discriminate different tumor states. Decision tree based esembled methods were proposed to identify biomarkers for inflammatory diseases.All these studies aim to discover the potential MS biomarkers that can distinguish from one group to another.stastical method is used to combine the high dimensional MS measurements into a single score to classify cancer status jointly with suitable preprocessing of the data. There are several studies on combining biomarkers. Su and Liu studied the case where markers follow a multivariate normal distribution.they gave a closed form of optimal solution to the linear parameters. Normality is not suitable for mass spectrometry data because measurements of relative abundance are always positive. Pepe and Thompson considered linearly combining two biomarkers by optimizing the area under the ROC

curve. The method was developed only for low dimensional situation. Ma and Huang applied Thompsons idea of optimizing AUC to microarray experiment. They used multivariate normal distribution in the simulation study and assumed independence between biomarkers which is not true for mass spectrometry data. The implementation is for real mass spectrometry ovarian cancer data analysis. TGDR-AUC method is applied to low dimensional and high dimensional ovarian cancer data. It is concluded that TGDR-AUC algorithm is appropriate in the analysis of mass spectrometry glycomic data.

Conclusion

Data mining is an appropriate and sufficiently sensitive method to analyze avalable glycomics data. Diversity in conceptualization of tools and diversity of common format used for outcome measurements could have hampered actual discovery of significant output.

References

1. Ceroni A, Dell A, Haslam SM (2007) The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. Source Code Biol Med 7: 2-3. » CrossRef » Pubmed » Google Scholar
2. Hashimoto K, Takigawa I, Shiga M, Kanehisa M, Mamitsuka H (2008) Mining significant tree patterns in carbohydrate sugar chains. Bioinformatics 24: i167-73. » CrossRef » Pubmed » Google Scholar
3. Herget S, Ranzinger R, Maass K, Lieth CW (2008) GlycoCT-a unifying sequence format for carbohydrates. Carbohydr Res 343: 2162-71. » CrossRef » Pubmed » Google Scholar
4. Sahoo SS, Thomas C, Sheth A, Henson C, York WS (2005) GLYDE-an expressive XML standard for the representation of glycan structure. Carbohydr Res 340: 2802-7. » CrossRef » Pubmed » Google Scholar
5. Tamaki Y, Nomura Y, Katsumura S, Okada A, Yamada H, et al. (2009) Construction of a dental caries prediction model by data mining. J Oral Sci 51: 61-8. » CrossRef » Pubmed » Google Scholar