Journal of Glycobiology

Bennun et al., J Glycobiol 2013, S1
DOI: 10.4172/2168-958X.S1-004

Research Article Open Access

Towards Integrative Glycoinformatics for Glycan Based Biomarker Cancer Research and Discovery

Sandra V Bennun[1]*, Deniz Baycin Hizal[2], René Ranzinger[3] and Michael J Betenbaugh[2]

[1]*Nanobiology Department, Sandia National Laboratories, Albuquerque, NM 87185, USA*
[2]*Department of Chemical and Biomolecular Engineering, Johns Hopkins University 3400 North Charles Street- Baltimore, Maryland 21218, USA*
[3]*Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA*

Abstract

Despite some recent successes in deciphering new cancer molecular makers, there is still a clear and continual need to develop new technologies that help characterizing existing biomarkers or facilitate discovery of new biomarkers. An important systems biology opportunity on this respect is provided by understanding the glycosylation changes associated with cancer. Indeed, interest in cancer glycosylation has expanded over the past decade and large amount of data relevant to cancer glycosylation has been accumulating rapidly. Furthermore, new and improved sophisticated glycoinformatics tools, methods and databases for glycan analysis now offer the opportunity to investigate this data for understanding the role that glycans play in cancer glycosylation. Here we summarize developments of glycoinformatics tools to support analysis of cancer glycosylation and experimental glycoproteomics approaches. In addition, we discuss challenges faced by glycoinformatics for the integration and interrogation of disparate high-throughput glycan data sets in order to assimilate technologies and better address cancer glycosylation. We also provide examples of integrative glycoinformatics approaches that lead to a better understanding of cancer glycosylation as a complex cellular process.

Keywords: Cancer; Biomarkers; Glycobiology; Glycans; Integrative glycoinformatics; Glycoproteomics; Automatic glycan annotation; High-throughput; Technologies; Mass spectrum spectrum; Gene expression; Databases; Mathematical modeling; Monocytic leukemia; Prostate cancer; Lncap cells

Introduction

Glycans are structurally complex carbohydrate chains found on proteins, to form glycoproteins, and on lipids, to form glycolipids. Abnormalities in glycosylation are linked to cancer and many other diseases [1-4]. The glycosylation of healthy and diseased cells often diverges resulting in glycan changes that may contribute to pathological progression [5,6]. Thus changes in glycosylation provide promise for the diagnosis and therapeutics of a wide variety of cancers [7-10]. To give but one of many examples, antibodies that recognize the sialyl Lewis$^{a/x}$ carbohydrate determinants provide statistically significant correlations between the postoperative prognosis of the patients and the degree of expression of these determinants in cancers of the colon, lung, breast, stomach, prostate, and urinary bladder [2].

Given the nature of cancer as a complex and heterogeneous disease, there is a clear need to develop new technologies to improve existing or discover new biomarkers. Current biomarkers include, among others, proteins, lipids, metabolites and DNA from which the most widely utilized are proteins [11]. Considering that glycosylation is one of the most common posttranslational modifications and that more than half of the proteins undergo glycosylation [12,13], glycans offer opportunities to improve or develop new biomarkers of cancer [7,14-17]. Furthermore, since these glycan patterns are exposed on the surface of the cells, they are readily amenable to profiling using new high-throughput technologies [18,19].

Advances in high-throughput technologies are benefiting cancer glycobiology and allowing fast screening of new glycomics data at large scales. That, together with the later sophistication of analytical techniques [18-24] and glycoinformatics data analysis tools [25-33] has rendered increased opportunities for high-throughput screening of glycan cancer biomarkers similarly to those used in proteomics and genomics. However, there are significant challenges to better understanding the cellular glycosylation transformations that accompany cancer and the use of this information to develop improved diagnostics of clinical utility. Improved integrative glycoinformatics tools will likely play an important role in addressing these challenges.

Experimental Glycoproteomic Approaches

Aberrant glycosylation has been associated with cancer, and several proteins that are biomarkers and targets for cancer have been found to be glycosylated. There are many US Food and Drug Administration (FDA) approved biomarkers and targets such as epidermal growth factor receptor, which is used for therapy selection and α-fetoprotein and human chorionic gonadotropin-β which are currently used for diagnosis [34]. Rather than just studying the glycan structures, glycoproteomics evaluates the glycosylated proteins and their glycosylation sites [35]. Glycoprotein enrichment can be coupled with comparative proteomics methods and advanced mass spectrometry techniques to find the differentially expressed proteins between the aggressive and non-aggressive cancer samples which can potentially represent biomarkers for early prediction of cancer, and targets for cancer drugs. Stable isotope labelling (SILAC) [36], isobaric tag for relative and absolute quantitation (iTRAQ) [37] and tandem mass tag (TMT) [38], are some of the methods used to interpret the differentially expressed proteins between samples for biomarker and target discovery.

Tian et al. [37] used the solid phase extraction of glycosylated peptides (SPEG) technique to enrich the glycoproteins from ovarian tumors and adjacent normal ovary tissues. The enriched glycopeptides from the normal ovary, clear cell carcinoma, high grade endometrioid carcinoma, high grade serous carcinoma, low-grade endometrioid

***Corresponding author:** Sandra V Bennun, Nanobiology Department, Sandia National Laboratories, Albuquerque, NM 87185, USA, Tel: 505-284-9150; Fax: 505-284-1323; E-mail: svbennun@gmail.com

Received May 13, 2013; **Accepted** June 21, 2013; **Published** June 24, 2013

Citation: Bennun SV, Hizal DB, Ranzinger R, Betenbaugh MJ (2013) Towards Integrative Glycoinformatics for Glycan Based Biomarker Cancer Research and Discovery. J Glycobiol S1: 004. doi:10.4172/2168-958X.S1-004

carcinoma, low-grade serous carcinoma, mucinous carcinoma, and transitional carcinoma samples were labeled with iTRAQ reagents and relative protein quantitation depending on iTRAQ labelling and MS/MS spectral counting were performed. They were able to identify both the proteins showing differential expression in ovarian tumors versus normal tissues as well as uniquely over expressed proteins specific to each ovarian tumor. Further western blot analysis have also supported the proteomics results and showed the elevated levels of carcinoembryonic antigen-related cell adhesion molecule 5 (CEA5) and CEA6 in ovarian mucinous carcinoma.

In addition to the analysis of glycoproteins, hydrazide chemistry can also provide enrichment and identification of sialoglycoproteins. Using sialoglycoproteome enrichment and isotope labeling methods, the proteins showing differential expression in breast cancer have been found. Further western blot and lectin analyses confirmed that versican (versatile extracellular matrix proteoglycan, VCAN) is one of the most highly differentially expressed sialoglycoproteins in breast cancer [15]. In addition, coupling sialoglycoprotein enrichment methods with selective reaction monitoring (SRM) techniques showed the up regulation of sialylated prostate specific antigen (PSA) in the prostate cancer tissues [39]. In order to increase the accuracy of prognosis and diagnosis of cancer types, the organ specific glycosylated and sialylated proteins, such as PSA, can be used. Hydrazide chemistry, lectins, multilectin affinity chromatography and metabolic incorporation of sugar analogs for glycoprotein isolation are the commonly used methods for biomarker and target discovery aimed at early detection and therapy of different cancer types [17].

Glycoinformatics Resources for Analysis and Data Interpretation

The availability of experimental data for glycobiology research has been increasing in recent years, driven by several initiatives in the United States, Europe and Japan. Examples are the Consortium of Functional Glycomics [30,40], the Japan Consortium for Glycobiology and Glycotechnology database (JCGGDB) [41], UniCarb-DB [42] and EUROCarbDB [33] (Table 1). Unfortunately, much of this growing data has not been analyzed yet, although the numbers of publications involving both “glycosylation” and “cancer & glycosylation” have constantly increased in recent years (Figure 1). This is due to the complexity of glycans and the glycosylation processes, which have been a bottleneck for the development of high-throughput analysis workflows and integrative glycoinformatics tools. In addition the heterogeneity of cancer as disease, posses many challenges and many aspects of cancer glycosylation remain uncharacterized. We have little or no understanding of how glycosyltransferases or the pathways they control are affected by cancer and a limited knowledge of the resulting deviations in glycan structural characteristics of cancer. Therefore, it is evident the need to integrate current glycoinformatics resources, that provide databases and tools to support glycobiologists in their research, as is being accomplished for proteomics and genomics resources [43]. In that respect high-throughput processing of glycomics data offers opportunities, however it can only be handled properly with some sort of automated pipeline that requires extensive glycoinformatics support to organize, analyze, and integrate experimental data and obtain valuable insights. Towards that end many databases, tools and computational methods are being developed for glycosylation studies and several glycoinformatics platforms are also publicly available. Some of these platforms serve as repositories of glycan structures, glycogenes, enzymes, and experimental glycan data. Other platforms permit analysis of glycans from diverse perspectives and interpret different types of experimental data. Considering the complexity of cancer, a deep and integrative analytical approach is required to understand and ultimately obtain knew knowledge about this disease. In this section we will briefly describe many of the databases and tools available for glycan analysis based on the type of information to be processed and the particular software functionality. Table 1, shows some of the main databases that are publicly available with glycan structural resources, a brief description, references and web address for access is provided.

Even by the time of this writing there is no complete agreement on a common structural code representation for monosaccharide residues and glycan sequences [52]. Each database initiative (e.g. KEGG, CFG) developed their own codes for representing glycans and created their independent databases to store glycan structures, which made it almost impossible to combine the information from more than one database. Although there has been an initiative towards an agreement that the sequence format GLYDE-II should be used as a general exchange format for glycan structures [53], most database and glycomics tools still use their own sequence encoding. With the implementation of GlycomeDB [31,44], a single resource database for glycan structures was established. GlycomeDB integrates entries from the major databases including BCSDB, GLYCOSCIENCES.de, and the Consortium for Functional Glycomics (CFG), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and CarbBank as well as the glycans extracted from the Protein DataBank (PDB). GlycomeDB stores glycan structures in a sequence format GlycoCT [54] that is capable of storing all structural information of carbohydrate sequences including missing structural information or incomplete structures. In addition to the structural data GlycomeDB maintains and updates the biological source information retrieved from the integrated databases and references (IDs) of the original databases entries. The GlycomeDB web interface allows storing carbohydrate structures in different sequence formats and graphical representations that can be in other programs supporting these formats. Table 2 shows an updated number of glycan structures as of May 2013.

In Table 3 glycoinformatics resources and tools for glycan analysis are shown including web applications, standalone applications and web-based resource sites. For example, the Glycomics Portal platform is a web-based search engine for glycoinformatic tools, which is timely updated; currently it stores 34 Databases, 30 Web services, 32 Software and 1 workflow. Most of the tools mentioned in this paper and others are registered in a single place, expediting the search of glycobiology resources. Another website for resources is RING, which provides algorithmic and data mining tools. Additional descriptions with their corresponding literature references for other resources tools for glycan analysis and interpretation are defined in Table 3. Further resources and applications for molecular modeling of glycan structures are described in detail in DeMarco and Woods [55]. Given the vast collection of data and resources that many of these websites have it is not possible to cover all, in addition a classification of them based on one trait it is sometimes not trivial [20,33,56].

Challenges Faced by Integrative Glycoinformatics for Cancer Glycosylation

Integrative glycoinformatics implements methods from mathematics, statistics, biology, computer science and engineering to interrogate, analyze, interpret, and integrate diverse glycosylation data sets. The implementation of integrative glycoinformatics approaches to help understand cancer glycosylation will require the development of methods to integrate disparate datasets, visualization applications,

Publicly available Glycan Databases	Description/ Content	Web address
GlycomeDB [31,44]	Portal for glycan structures that have been integrated from several of the major glycan-related databases in this table. The GlycomeDB provides cross-references to the integrated databases.	http://www.glycome-db.org
GLYCOSCIENCES[29]	Provides glycan structure data, mainly extracted from CarbBank, thePDB (Protein DataBank) and literature research. It includes biological source information, NMR data and literature references. A glycan 3D structure generation tool (Bohne, et al. 1999)is integrated with the structure database.	http://www.glycosciences.de
CFG [30,40]	The CFG glycan database includes structures from CFG's mammalian glycan array, reagent bank, glycomics profiling analyses of different cells and tissues and structures extracted from other databases (CarbBank and GlycoMinds Ltd). Glycan structures are linked with the CFG glycan binding protein database, which includes gene and protein sequences, biological functions, and binding specificities, based on the Glycan Array data. CFG also providesa glycotransferases database that allows to identify the glycosidic linkage they form, knock-out mouse phenotype data, glycan profile data from MALDI-TOF mass spectra, data for glycan binding specificity for mammalian and pathogen CFG's glycan arrays as also enzyme transcriptomic data for various tissues and cells.	http://www.functionalglycomics.org/
KEGG Kyoto Encyclopedia of Genes [45,46]	The KEGG glycan structure database contains entrees from CarbBank, from publications and structures linked with the genes and pathways provided byother KEGG resources. These resources contain the glycan structures glycogene information, glycosyltransferase and glycosylhydrolase reactions, glycan binding protein data, and metabolic/glycosylation pathways can be obtained.	http://www.genome.jp/kegg/glycan/
JCGGDB Japan Consortium for Glycobiology and Glycotechnology database [41]	A portal allowing to search across the major bioinformatic databases in Japan. Integrated resources include tumor marker reference, Glycoepitope databases, and a glycodisease gene databases. It also providesexperiment support and resources, including glycan profile data from mass spectral, lectin arraydata, glycoprotein data, glycogene data and tools for glycan structure analysis.	http://jcggdb.jp/
EUROCarbDB [33]	An infrastructure framework for glycoinformatics. Provides glycoinformatic tools and databases for interpretation and storing of glycan structures and glycan experimental data,including experimental techniques such as tandem MS, HPLC and NMR. As this platform was discontinued, other efforts and spin off can be accessed through the first URL here. These are Glycoworkbench, MonosacharideDB,Glycoscience.de, Eurocarb mirror, Casper and GlycomicsPortal. The framework development has been continued and utilized by UniCarb-DB and UnicarbKB. A mirror of the orginail EuroCarbDB database is available in Sweden (second URL)	http://www.ebi.ac.uk/eurocarb http://relax.organ.su.se/eurocarb/home.action
CarbBank (CCSD) [47] The complex carbohydrate structure database	CarbBank was the first publicly available glycan structure database with approximately 50000 entries extracted from literature. . The database contains glycan structure, publications, biological source information and information about the experimental technique used as also data about the attached aglyca.	http://www.genome.jp/dbget-bin/www_bfind?carbbank
SUGABASE [48]	SUGABASE is a carbohydrate-NMR database that combines CarbBank Complex Carbohydrate Structure Data (CCSD) with proton and carbon chemical shift values.	http://glycoscience.wetpaint.com/page/SugaBase
UniCarbKB [49]	Database with experimental data, linked with lycol-related UniProtKB data. Besides the glycan structures the database provides biological source information and publications.	http://www.unicarbkb.org/
UniCarb-DB [42]	Database with LC-MS/MS data of glycan structures. The database provides for each structure the biological source, LC-MS/MS data and publications.	http://unicarb-db.org/unicarbdb/show_mucin.action
GlycoSuiteDB [50]	Contains literature based glycan data that has been experimentally verified. In addition to the glycan the biological source, publication information, experimental technique and the attached glycan is provided by the database. The current database is hosted by Swiss Institute of BioInformatics andwill be integrated into UniCarbKB database (www.unicarbkb.org).	http://glycosuitedb.expasy.org/glycosuite/glycodb /
BCSDB – Bacterial carbohydrate structure database [51]	Contains literature extracted bacterial glycan data. The database provides the glycan structure, publications, biological source information and NMR data.	http://csdb.glycoscience.ru/bacterial/

Table 1: Directory of links to some of the major glycan structuralresources for glycoinformatics.

network analysis, and unified computational platforms. In this section, we briefly discuss challenges of integrative glycoinformatics in cancer glycobiology.

Glycan Complexity

Integrative glycoinformatics developments based on a wholistic understanding of the complex glycosylation process and the relations among its components can provide a more complete analysis that is not only based on glycan annotation but other aspects such as enzymatic levels, glycans abundances and biosynthetic pathways. However, most of the available glycoinformatics tools do not consider the integration and the complex relationships among the different components of the glycosylation process (e.g. enzymes, glycans, sugar nucleotides, transporters), in which the glycan structures are defined as a result of the action of many enzymes. Instead, these tools are based primarily on standard bioinformatics approaches mostly developed for proteomics and genomics, which have limitations in their application to glycomics. These tools may not work properly for glycans, given that glycans are not directly encoded in the genome and differ from proteins in that they are assembled from the interconnected action of several enzymes.

Another example of the level of complexity of glycans is their structure, which is affected by the linkage information and branching resulting in numerous isomeric glycan structures. Furthermore, the glycan constituent monosacharides are characterized by enantiomeric configuration (D/L), the anomeric configuration (α/β), the type of ring formed by the monodaccharide (p/f), the modification of the monosaccharide (e.g. deoxygenation) and number and position of substituents (e.g. methyl or n-acetyl groups) resulting in a large set of monosaccharides utilized by different organisms [66,67]. This level of detail has offered challenges in the field of glycomics and resulted in slow development of methods for structure determination and availability of glycan structures in databases. However, that level of detail in the structure elucidation is important for the understanding of the function of glycans and their interaction with other molecules.

For that reason and because of the adoption of traditional

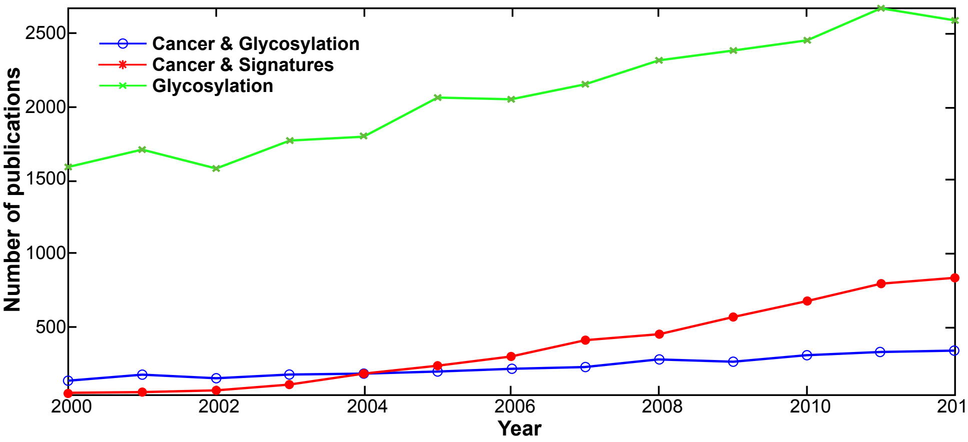

Figure 1: Growth of Publications in cancer glycosylation. The number of publication from 2000-2012 retrieved from PUBMED, by searching the terms “Cancer & Glycosylation”, “Glycosylation” and “Cancer & Signature”.

Resource	# Structures	# Entries in GlycomeDB
BCSDB	9.121	6.046
CarbBank	23.402	14.887
CFG	9.201	6.374
GlyAffinity	579	579
GlycO	1.325	1.325
GlycoBase (Lille)	252	199
GLYCOSCIENCES.de	23.475	15.936
KEGG	10.979	10.167

Table 2: Number of structures in major databases and the corresponding entrees in GlycomeDB as May 2013. The number of structures in the sequence format utilized by the integrated resource (# Structures) and the number of GlycoCT structures in GlycomeDB after integration (# Entries in GlycomeDB) is shown.

bioinformatics approaches that do not consider the complexity of glycosylation, methods and tools for analysis of glycans have lagged and most glycoinformatics tools available are specialized for the analysis of one type of data [59,60,68,69], For example, a common approach for mass spectrometry-based glycoprofiling involves a one-to-one database matching of particular mass spectrometry measurements to specific glycans from a known glycan library in order to annotate the mass spectra in terms of individual peaks that are analyzed separately [53,70]. Methods that consider the complexity of glycosylation, would require that all glycan structures used for the annotation of the spectrum can be generated by the enzymatic machinery of the studied organism. Assuring consistency among enzyme activities and those structures assigned to each peak in the same spectrum.

Biomarkers

The discovery of glycan-based biomarkers requires efficient extraction and analysis of glycan structures. Indeed, the identification of appropriate glycan based biomarkers of different cancer types is challenging and has been hindered by a number of factors, including: specific glycans may not be present in current databases, cancer associated glycans may be at low levels, not a specific glycan but a collective pattern of glycans structures may be more representative of the cancer state, absence of a unified and standard format for glycans notation, and the requirement of novel algorithms to handle glycan complexity. Most importantly, differences in glycan profiles between cancer and normal cells may involve subtle differences in amounts rather than on and off changes. Fortunately, these subtle differences can be addressed, at least potentially through systems biology methods including those described in the next section. Differences in glycan profiles and also likely different enzyme expression profiles offer the potential to be used as valuable biomarkers indicating a cancer transformation from the normal state [28], and to more malignant states [26]. One goal involves obtaining new epitopes based on differential glycan patterns observed between normal and diseased cells and tissues. Algorithms based on machine learning methods [71-74], frequent subtree mining [74-76] and mathematical modeling [26,28] have been developed to predict glycan biomarkers or glycan biomarker patterns. Some of these algorithms can compare two glycan profiles directly, each consisting of thousands of structures, to find those substructures or combinations of substructures that most sensitively characterize the differences between the two profiles.

Data Integration

Glycosylation is a highly complex multienzyme process and current bioinformatics techniques that attempt to integrate diverse data are still in early development [26,45,77,78] despite the progress on several ‘omics’ fronts. Statistical database-driven approaches to relate gene expression levels to the abundance of specific glycan linkages did not provide quantitative predictions of detailed glycan distributions [45,77]. This reflects the need for integrative glycoinformatics tools to identify glycan structural data and also to link these with gene expression data of glycosylation enzymes that produce these glycan structures. Mathematical modeling of glycosylation may represent a promising approach to start understanding how mRNA levels relate to the actual amount and distribution of glycans found within healthy or diseased cells [26,28,78].

An approach that considers data integration could be highly effective to reduce variability (false positives and negatives) in the analytical high-throughput experimental platform. Results obtained

Resource	Content	Web address
GlycomicsPortal	A web-based portal for registration and discovery of GlycoInformatics tools. It host webpages for glycoinformatic resources. The portal contains updated URL's for web services, workflows, databases and software, together with information necessary to evaluate and use these tools.Resource providers can also upload addition material, such as screenshots tutorials, source code or related publications.	http://glycomics.ccrc.uga.edu/GlycomicsPortal/welcome.action
RINGS [25]	Web resources site with algorithmic and data mining glycan structure tools.	http://rings.t.soka.ac.jp
GlycoBase [57]	Glycobase 3.2contains elution positions for N-linked and O-linked glycan structures determined by a combination of HPLC, UPLC, exoglycosidase sequencing and mass spectrometry (MALDI-MS, ESI-MS, ESI-MS/MS, LC-MS, LC-ESI-MS/MS). Provides glycoinformatics and isualization functionalities for the data mining of glycans of biological interest. A web application for the identification of glycan structures bases on retention time (AutoGU) is integrated in the database.	http://glycobase.nibrt.ie/glycobase/show_nibrt.action
GGDB Glycogene data base [41]	Database containinginformation for analysis of lycogens, includes genes associated with glycan synthesis, transport, and nucleotide synthesis. More than 180 human lycogens were identified, cloned and characterized. Genes are stored in XML format.	http://riodb.ibase.aist.go.jp/rcmg/ggdb/
Elution Coordinate Database [58]	Contains more than 400 pyridylaminated N-glycan structures, code numbers, the elution positions on Shimpack CLC-ODS and Amide-80 columns, sources and references for the 2-D/3-D sugar mapping	http://www.gak.co.jp/ECD/Hpg_eng/hpg_eng.htm
GlycoWorkbench [59]	GlycoWorkbench is a standalone suite of software tools designed for rapid drawing of glycan structures and for assisting the process of structure determination from mass spectrometry data. The program also allows for the annotation of MS data using structures from databases (e.g. CarbBank, CFG).	http://www.eurocarbdb.org/applications/ms-tools http://code.google.com/p/glycoworkbench/
Glyco-Peakfinder [60]	Determination of glycan compositions from their mass signals for rapid annotation of Msand MSn data with different types of ions.	http://www.glyco-peakfinder.org/
GlycoMod [61]	Web application for the composition analysis of glycan and glycol-peptide MS data.	http://web.expasy.org/glycomod/
GlycoPep ID	Identifying the peptide moiety of glycopeptides generated using a nonspecific enzyme	http://hexose.chem.ku.edu/predictiontable.php
CancerLectinDB [62]	This database provides diverse lectin data integrated into a common framework together with analytical tools and extensive links. Data for each lectin pertains to taxonomic, biochemical, domain architecture, molecular sequence and structural details as well as carbohydrate and hence blood group specificities.	http://proline.physics.iisc.ernet.in/cancerdb/
CAZY [63]	Specialist database dedicated to the display and analysis of genomic, structural and biochemical information on Carbohydrate-Active Enzymes	http://www.cazy.org/
BRENDA [64]	Comprehensive Enzyme Information System	http://www.brenda-enzymes.info/
Enzyme [65]	Enzyme nomenclature database	http://enzyme.expasy.org/

Table 3: Web-based glycoinformatic resources for glycan analysis.

with integrative glycoinformatics tools that are confirmed by different experimental data (e.g. glycogenes expression and MS profile) will increase confidence in predictions and recommendations for biomarkers. Moreover, integrated glycoinformatics tools enable the analysis and comparisons of multiple studies with multiple platforms, which can reveal limitations in analytical sensitivities. For example, our work on analyzing prostate cancer glycosylation resulted in some predictions for undetectable protein levels from mRNA data when compared to enzymes levels predicted based on MS structural glycan profile data [26].

Examples of Integrative Glycoinformatics for Cancer Glycosylation

The accumulation of large amount of glycomics data and the improvement of analytical methods require the development of data standards and new integrative glycoinformatics tools for optimal analysis and effective validation. Here, we briefly discuss one approach from our laboratory to develop and implement some of these tools together with examples applying this integrative glycoinformatics approach to understand cancer glycosylation.

A comprehensive simulation framework can be highly beneficial in developing integrative analytical tools [26,28]. In one application, the method was applied to integrate information across a broad spectrum of mass spectral data involving leukemic cell types as compared to normal cell types [28]. In another example, the method was used to integrate mass spectral data and mRNA datasets for identifying glycan patterns associated with prostate cancer types [26]. These analyses provide a deeper understanding of the interrelation among the complex cellular processes leading to changes in cancer glycosylation.

To assist this analyses our group has developed a glycosylation model for mammalian cells that uses MALDI TOF mass spectra [28]. The result of applying this model is a complete characterization of a measured glycan mass spectrum in terms of a relatively small number of enzyme activities. In addition automatic annotation of the mass spectrum in terms of glycan structures is produced, with every peak being assigned a full range of alternative glycan structures and abundances (Figure 2).

The method was applied to mass spectra data of normal-human monocytes and monocytic leukemia cells, and it provided insights into the relevant glycosylation pathways that differentiate normal and diseased cells and the implications of the glycosyl transferases in the resulting diseased and normal glycan structures [28] (Figure 2). This is an important advance towards integrative glycoinformatics that connected the enzymatic activities of the glycosyltransferases, the complete bioprocessing pathways, and the resulting glycan structures.

A more advanced implementation of this method involved the development of an integrative glycoinformatics tool that processed and integrated mass spectra (MS) and gene expression data. The power of this novel method for interpreting and integrating mass spectra and gene expression microarray data was demonstrated and applied to low and high passage Lymph Node Carcinoma of the Prostate (LNCaP) cancer cells, which correspond to androgen dependent and the more metastatic androgen independent cell stages [26]. The novel method identified and quantified glycan structural details not typically derived from single-stage mass spectral or gene expression data (Figure 3).

Differences between the cell types uncovered include increases in the more metastatic androgen-independent cells of H type II and Lewis-y glycan structures characteristics of blood groups and the

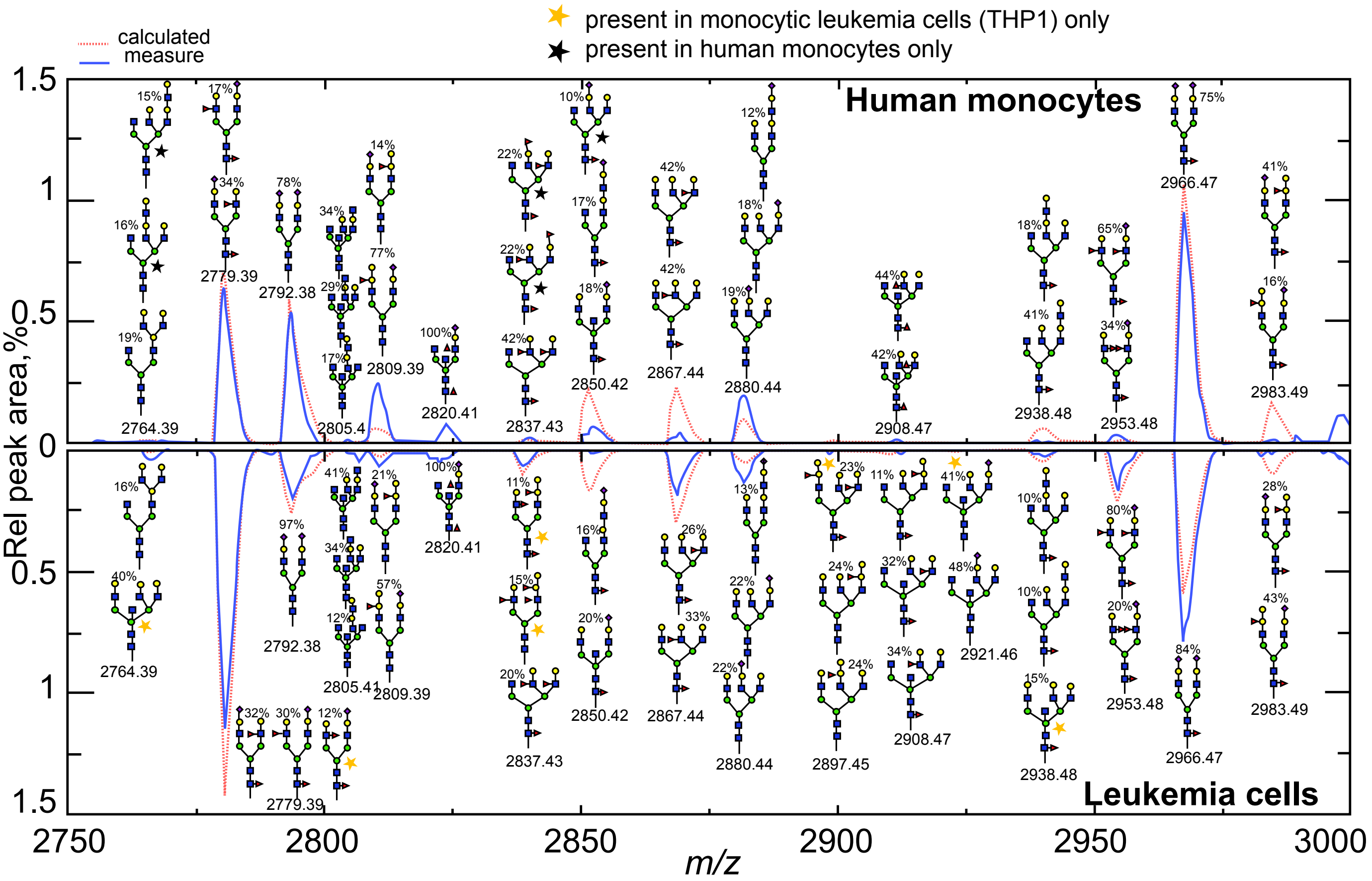

Figure 2: Comparison of model calculated synthetic mass spectra with respect to measured mass spectra for human monocites (normal) and leukemia cancer cells for selected section (2750-3000) of the mass spectra.

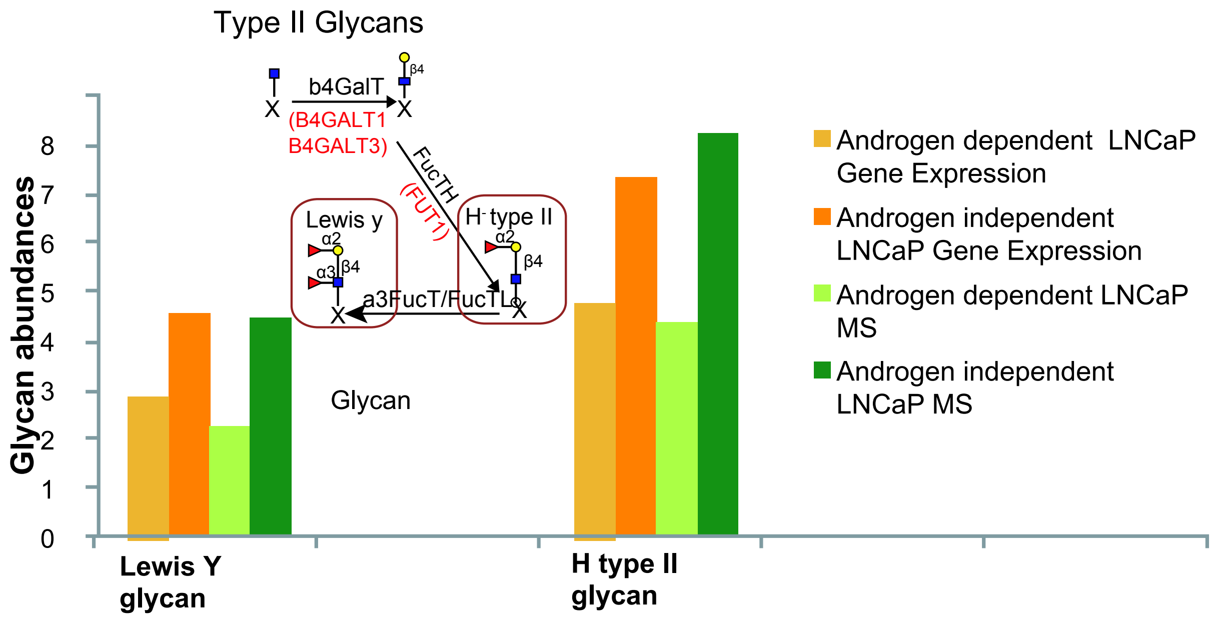

Figure 3: Integration of the glycomics and transcriptomics for elucidating biomarkers of prostate cancer LNCaP cells. This integrative glycoinformatics approach allows the simultaneous analysis of gene expression (orange) and mass spectra (green) data sets to elucidate glycan substructures and enzyme levels that differentiate prostate cancer types.Increases of H type II and Lewis Y glycans abundances in more malignant androgen independent prostate cancer tumor cells vs. less malignant cells were derived and show agreement for both MS (dark green vs. light green) and gene expression data (dark orange vs. light orange).

correlation of a correspondingly greater activity of α2-Fuc-transferase (FUT1). The model further elucidated limitations in the two analytical platforms including a defect in the microarray for detecting the GnTV (MGAT5) enzyme. The results demonstrate the potential of integrative glycoinformatics tools for elucidating key glycan biomarkers and potential therapeutic targets along with specifying limitations in the analytical platforms. The integration of multiple data sets demonstrates how a systems biology approach can provide a better understanding of complex cellular processes such as cancer glycosylation and lead to the elucidation of glycan signatures representative of potential cancer biomarkers.

Funding

This work was supported by the National Cancer Institute: Awards 5R41CA127885-02 and R01CA112314, the Consortium of Functional Glycomics: Bridging Grant Number U54-GM062116-10, and NIH/NIGMS funding the National Center for Glycomics and Glycoproteomics (8P41GM103490). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute and the Consortium of Functional Glycomics.

References

1. Brockhausen I (2006) Mucin-type O-glycans in human colon and breast cancer: glycodynamics and functions. EMBO Rep 7: 599-604.
2. Brockhausen I, Schutzbach J, Kuhns W (1998) Glycoproteins and their relationship to human disease. Acta Anat (Basel) 161: 36-78.
3. Hakomori S (2002) Glycosylation defining cancer malignancy: new wine in an old bottle. Proc Natl Acad Sci U S A 99: 10231-10233.
4. Kim YJ, Varki A (1997) Perspectives on the significance of altered glycosylation of glycoproteins in cancer. Glycoconj J 14: 569-576.
5. Dennis JW, Granovsky M, Warren CE (1999) Glycoprotein glycosylation and cancer progression. Biochim Biophys Acta 1473: 21-34.
6. Hakomori S (2001) Tumor-associated carbohydrate antigens defining tumor malignancy: basis for development of anti-cancer vaccines. Adv Exp Med Biol 491: 369-402.
7. Adamczyk B, Tharmalingam T, Rudd PM (2012) Glycans as cancer biomarkers. Biochim Biophys Acta 1820: 1347-1353.
8. Buskas T, Thompson P, Boons GJ (2009) Immunotherapy for cancer: synthetic carbohydrate-based vaccines. Chem Commun (Camb) : 5335-5349.
9. Fuster MM, Esko JD (2005) The sweet and sour of cancer: glycans as novel therapeutic targets. Nat Rev Cancer 5: 526-542.
10. Tong L, Baskaran G, Jones MB, Rhee JK, Yarema KJ (2003) Glycosylation changes as markers for the diagnosis and treatment of human disease. Biotechnol Genet Eng Rev 20: 199-244.
11. Hanash SM, Pitteri SJ, Faca VM (2008) Mining the plasma proteome for cancer biomarkers. Nature 452: 571-579.
12. Apweiler R, Hermjakob H, Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 1473: 4-8.
13. Furukawa K, Kobata A (1992) Protein glycosylation. Curr Opin Biotechnol 3: 554-559.
14. Arnold JN, Saldova R, Galligan MC, Murphy TB, Mimura-Kimura Y, et al. (2011) Novel glycan biomarkers for the detection of lung cancer. J Proteome Res 10: 1755-1764.
15. Tian Y, Esteva FJ, Song J, Zhang H (2012) Altered expression of sialylated glycoproteins in breast cancer using hydrazide chemistry and mass spectrometry. Mol Cell Proteomics 11: M111.
16. Tian Y, Bova, GS, Zhang H (2011) Quantitative glycoproteomic analysis of optimal cutting temperature-embedded frozen tissues identifying glycoproteins associated with aggressive prostate cancer. Anal Chem 83: 7013-7019.
17. Tian Y, Zhang H (2013) Characterization of disease-associated N-linked glycoproteins. Proteomics 13: 504-511.
18. Ito H, Kuno A, Sawaki H, Sogabe M, Ozaki H, et al. (2009) Strategy for glycoproteomics: identification of glyco-alteration using multiple glycan profiling tools. J Proteome Res 8: 1358-1367.
19. Rakus JF, Mahal LK (2011) New technologies for glycomic analysis: toward a systematic understanding of the glycome. Annu Rev Anal Chem (Palo Alto Calif) 4: 367-392.
20. Furukawa J, Fujitani N, Shinohara Y (2013) Recent Advances in Cellular Glycomic Analyses. Biomolecules 3: 198-225.
21. Hua S, An HJ (2012) Glycoscience aids in biomarker discovery. BMB Rep 45: 323-330.
22. Ito S, Hayama K, Hirabayashi J (2009) Enrichment strategies for glycopeptides. Methods Mol Biol 534: 195-203.
23. Tousi F, Hancock WS, Hincapie M (2011) Technologies and strategies for glycoproteomics and glycomics and their application to clinical biomarker research. Analytical Methods 3: 20-32.
24. Zhang Y, Yin H, Lu H (2012) Recent progress in quantitative glycoproteomics. Glycoconj J 29: 249-258.
25. Akune Y, Hosoda M, Kaiya S, Shinmachi D, Aoki-Kinoshita KF (2010) The RINGS resource for glycome informatics analysis and data mining on the Web. OMICS 14: 475-486.
26. Bennun SV, Yarema KJ, Betenbaugh MJ, Krambeck FJ (2013) Integration of the Transcriptome and Glycome for Identification of Glycan Cell Signatures, PLoS Comput Biol 9: e1002813.
27. GlycomicsPortal.
28. Krambeck FJ, Bennun SV, Narang S, Choi S, Yarema KJ, et al. (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. Glycobiology 19: 1163-1175.
29. Lütteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, et al. (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. Glycobiology 16: 71R-81R.
30. Raman R, Raguram S, Venkataraman G, Paulson JC, Sasisekharan R (2005) Glycomics: an integrated systems approach to structure-function relationships of glycans. Nat Methods 2: 817-824.
31. Ranzinger R, Herget S, von der Lieth CW, Frank M (2011) GlycomeDB-a unified database for carbohydrate structures. Nucleic Acids Res 39: D373-D376.
32. Taniguchi N, Paulson JC (2007) Frontiers in glycomics; bioinformatics and biomarkers in disease. September 11-13, 2006 Natcher Conference Center, NIH Campus, Bethesda, MD, USA. Proteomics 7: 1360-1363.
33. von der Lieth CW, Freire AA, Blank D, Campbell MP, Ceroni A, et al. (2011) EUROCarbDB: An open-access platform for glycoinformatics. Glycobiology 21: 493-502.
34. Pan S, Chen R, Aebersold R, Brentnall TA (2011) Mass spectrometry based glycoproteomics--from a proteomics perspective. Mol Cell Proteomics 10: R110.
35. Tian Y, Zhang H (2010) Glycoproteomics and clinical applications. Proteomics Clin Appl 4: 124-132.
36. Kashyap MK, Harsha HC, Renuse S, Pawar H, Sahasrabuddhe NA, et al. (2010) SILAC-based quantitative proteomic approach to identify potential biomarkers from the esophageal squamous cell carcinoma secretome. Cancer Biol Ther 10: 796-810.
37. Tian Y, Yao Z, Roden RB, Zhang H (2011) Identification of glycoproteins associated with different histological subtypes of ovarian tumors using quantitative glycoproteomics. Proteomics 11: 4677-4687.
38. Raso C, Cosentino C, Gaspari M, Malara N, Han X, et al. (2012) Characterization of Breast Cancer Interstitial Fluids by TmT Labeling, LTQ-Orbitrap Velos Mass Spectrometry, and Pathway Analysis. J Proteome Res Epub ahead of print.
39. Li Y, Tian Y, Rezai T, Prakash A, Lopez MF (2011) Simultaneous analysis of glycosylated and sialylated prostate-specific antigen revealing differential distribution of glycosylated prostate-specific antigen isoforms in prostate cancer tissues. Anal Chem 83: 240-245.
40. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, et al. (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. Glycobiology 16: 82R-90R.
41. Yoshida, K., Suzuki, A. and Taniguchi, N. (2004) [Japan consortium for

glycobiology and glycotechnology; toward establishment of international network and systems glycobiology]. Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme, 49: 2313-2318.

42. Hayes CA, Karlsson NG, Struwe WB, Lisacek F, Rudd PM, et al. (2011) UniCarb-DB: a database resource for glycomic discovery. Bioinformatics 27: 1343-1344.

43. Goodman N (2002) Biological data becomes computer literate: new advances in bioinformatics. Curr Opin Biotechnol 13: 68-71.

44. Ranzinger R, Herget S, Wetter T, von der Lieth CW (2008) GlycomeDB - integration of open-access carbohydrate structure databases. BMC Bioinformatics 9: 384.

45. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, et al. (2006) KEGG as a glycome informatics resource. Glycobiology 16: 63R-70R.

46. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277-280.

47. Doubet S, Bock K, Smith D, Darvill A, Albersheim P (1989) The Complex Carbohydrate Structure Database. Trends Biochem Sci 14: 475-477.

48. van Kuik JA, Hård K, Vliegenthart JF (1992) A 1H NMR database computer program for the analysis of the primary structure of complex carbohydrates. Carbohydr Res 235: 53-68.

49. Campbell MP, Hayes CA, Struwe WB, Wilkins MR, Aoki-Kinoshita KF, et al. (2011) UniCarbKB: putting the pieces together for glycomics research. Proteomics 11: 4117-4121.

50. Cooper CA, Harrison MJ, Wilkins MR, Packer NH (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. Nucleic Acids Res 29: 332-335.

51. Toukach PV (2011) Bacterial carbohydrate structure database 3: principles and realization. J Chem Inf Model 51: 159-170.

52. Ranzinger R, York WS (2011) Glyco-Bioinformatics today (August 2011) – Solutions and Problems.

53. Packer NH, von der Lieth CW, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, et al. (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). Proteomics 8: 8-20.

54. Herget S, Toukach PV, Ranzinger R, Hull WE, Knirel YA, et al. (2008) Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. BMC Struct Biol 8: 35.

55. DeMarco ML, Woods RJ (2008) Structural glycobiology: a game of snakes and ladders. Glycobiology 18: 426-440.

56. Frank M, Schloissnig S (2010) Bioinformatics and molecular modeling in glycobiology. Cell Mol Life Sci 67: 2749-2772.

57. Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. Bioinformatics 24: 1214-1216.

58. Tomiya N, Awaya J, Kurono M, Endo S, Arata Y, et al. (1988) Analyses of N-Linked Oligosaccharides Using a Two-Dimensional Mapping Technique. Analy Biochem 171: 73-90.

59. Ceroni A, Maass K, Geyer H, Geyer R, Dell A, et al. (2008) GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. J Proteome Res 7: 1650-1659.

60. Maass K, Ranzinger R, Geyer H, von der Lieth CW, Geyer R (2007) "Glyco-peakfinder"--de novo composition analysis of glycoconjugates. Proteomics 7: 4435-4444.

61. Cooper CA, Gasteiger E, Packer NH (2001) GlycoMod--a software tool for determining glycosylation compositions from mass spectrometric data. Proteomics 1: 340-349.

62. Damodaran D, Jeyakani J, Chauhan A, Kumar N, Chandra NR, et al. (2008) CancerLectinDB: a database of lectins relevant to cancer. Glycoconj J 25: 191-198.

63. Coutinho PM, Deleury E, Davies GJ, Henrissat B (2003) An evolving hierarchical family classification for glycosyltransferases. J Mol Biol 328: 307-317.

64. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, et al. (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32: D431-433.

65. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28: 304-305.

66. Herget S, Ranzinger R, Maass K, Lieth CW (2008) GlycoCT-a unifying sequence format for carbohydrates. Carbohydr Res 343: 2162-2171.

67. Werz DB, Ranzinger R, Herget S, Adibekian A, von der Lieth CW, et al. (2007) Exploring the structural diversity of mammalian carbohydrates ("glycospace") by statistical databank analysis. ACS Chem Biol 2: 685-691.

68. Goldberg D, Sutton-Smith M, Paulson J, Dell A (2005) Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. Proteomics 5: 865-875.

69. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M (2005) Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. Bioinformatics 21: 3976-3982.

70. Joshi HJ, Harrison MJ, Schulz BL, Cooper CA, Packer NH, et al. (2004) Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics 4: 1650-1664.

71. Hizukuri Y, Yamanishi Y, Nakamura O, Yagi F, Goto S, et al. (2005) Extraction of leukemia specific glycan motifs in humans by computational glycomics. Carbohydr Res 340: 2270-2278.

72. Kuboyama T, Hirata K, Aoki-Kinoshita KF, Kashima H, Yasuda H (2006) A gram distribution kernel applied to glycan classification and motif extraction. Genome Inform 17: 25-34.

73. Li L, Ching WK, Yamaguchi T, Aoki-Kinoshita KF (2010) A weighted q-gram method for glycan structure classification. BMC Bioinformatics 11: S1-S33.

74. Yamanishi Y, Bach F, Vert JP (2007) Glycan classification with tree kernels. Bioinformatics 23: 1211-1216.

75. Aoki-Kinoshita KF (2013) Mining frequent subtrees in glycan data using the RINGS glycan miner tool. Methods Mol Biol 939: 87-95.

76. Hashimoto K, Takigawa I, Shiga M, Kanehisa M, Mamitsuka H (2008) Mining significant tree patterns in carbohydrate sugar chains. Bioinformatics 24: i167-173.

77. Suga A, Yamanishi Y, Hashimoto K, Goto S, Kanehisa M (2007) An improved scoring scheme for predicting glycan structures from gene expression data. Genome Inform 18: 237-246.

78. Krambeck FJ, Betenbaugh MJ (2005) A mathematical model of N-linked glycosylation. Biotechnol Bioeng 92: 711-728.

This article was originally published in a special issue, **Cancer Glycobioogy** handled by Editor(s). Prof. Giuseppina Simone, Health Care University of Napoli, Italy; Shashikala R.Inamdar, Karnatak University Dharwad, India