Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw341
Advance Access Publication Date: 10 June 2016
Original Paper

OXFORD

Databases and ontologies

The Glycome Analytics Platform: an integrative framework for glycobioinformatics

Christopher B. Barnett[1,*], Kiyoko F. Aoki-Kinoshita[2] and Kevin J. Naidoo[1,*]

[1]Scientific Computing Research Unit and Department of Chemistry, University of Cape Town, Rondebosch 7701, South Africa and [2]Department of Bioinformatics, Faculty of Engineering, Soka University, Hachioji, Tokyo 192-8577, Japan

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 21, 2015; revised on May 6, 2016; accepted on May 26, 2016

Abstract

Motivation: Complex carbohydrates play a central role in cellular communication and in disease development. O- and N-glycans, which are post-translationally attached to proteins and lipids, are sugar chains that are rooted, tree structures. Independent efforts to develop computational tools for analyzing complex carbohydrate structures have been designed to exploit specific databases requiring unique formatting and limited transferability. Attempts have been made at integrating these resources, yet it remains difficult to communicate and share data across several online resources. A disadvantage of the lack of coordination between development efforts is the inability of the user community to create reproducible analyses (workflows). The latter results in the more serious unreliability of glycomics metadata.

Results: In this paper, we realize the significance of connecting multiple online glycan resources that can be used to design reproducible experiments for obtaining, generating and analyzing cell glycomes. To address this, a suite of tools and utilities, have been integrated into the analytic functionality of the Galaxy bioinformatics platform to provide a Glycome Analytics Platform (GAP).

Using this platform, users can design *in silico* workflows to manipulate various formats of glycan sequences and analyze glycomes through access to web data and services. We illustrate the central functionality and features of the GAP by way of example; we analyze and compare the features of the N-glycan glycome of monocytic cells sourced from two separate data depositions.

This paper highlights the use of reproducible research methods for glycomics analysis and the GAP presents an opportunity for integrating tools in glycobioinformatics.

Availability and Implementation: This software is open-source and available online at https://bitbucket.org/scientificomputing/glycome-analytics-platform

Contacts: chris.barnett@uct.ac.za or kevin.naidoo@uct.ac.za

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Changes in the dynamic expression of cellular glycans (sugars) and glycoconjugates have been implicated in the initiation and progression of neoplastic disease (Campbell *et al.*, 2014a; Dube and Bertozzi, 2005; North *et al.*, 2010; Pierce, 2010; Redelinghuys and Crocker, 2010). This finding has driven research interest in defining the glycome *the complete repertoire of glycans and glycoconjugates that cells produce under specified conditions of time, space and environment.* (Bertozzi, 2009) Experimental techniques such as NMR, MALDI-TOF mass spectrometry (MS) and Liquid chromatography-MS are

used to determine the sequence and structure of glycans (Raman *et al.*, 2006; Ranzinger *et al.*, 2015; Satomaa *et al.*, 2009); while the specificity of glycan binding is analyzed by glycan arrays (Heimburg-Molinaro *et al.*, 2011) or lectin microarrays (Hirabayashi *et al.*, 2013) and computational methods in informatics and machine learning techniques are used to identify glycans and categorize their complexity (Li *et al.*, 2010).

While a chemically rigorous definition of carbohydrates is given elsewhere (McNaught, 1997) that is not of relevance here. Compared with DNA, RNA and proteins, which are composed of nucleotides and amino acids, a glycan is structurally and conformationally more complex.

The alphabets between these biomolecules differ. The 4 DNA monomers are the nucleotides derived from the A, C, T, G bases and the protein alphabet consists of 20 amino acids. The glycan alphabet is large. The monosaccharide database (http://www.monosaccharidedb.org/start.action) catalogues 776 monomers, GlyTouCan (Aoki-Kinoshita *et al.*, 2015) the international glycan structure repository lists 800 monomers and the Bacterial Carbohydrate Structure Database (http://csdb.glycoscience.ru/bacterial/) registers 357 bacterial glycan monomers. Comparing the bacterial and mammalian registries of monosaccharides (Herget *et al.*, 2008) reveals that bacteria have a much greater monomer diversity than humans. By defining the glycan alphabet in terms of the basetype (essentially the variation of the core, see formal definition in the reference (Herget *et al.*, 2008)) of a monosaccharide the human glycan alphabet consists of 14 monomers which is similar to that reported by Cummings, who suggests that the diversity in human glycans stems from merely 9 monomers (Cummings, 2009). While the human glycan alphabet is small in number (9), the complexity in glycan structure comes from the multiple linkages possible between monomers, the stereochemical possibilities for each linkage as well as the conformational flexibility of the monomers themselves. This complexity is evident by comparing the 9 combinatorial possibilities when linking three *different* amino acids to as many as 27 648 possibilities from three different monosaccharides (Laine, 2008).

A final level of complexity is the synthesis of glycans. Unlike DNA and protein synthesis, glycan synthesis is non-templated. Protein enzymes are responsible for glycan synthesis. Glycosyl transferases extend glycans by adding monomers to existing glycans, while Glycosyl Hydrolases lyse glycosidic linkages and divide glycans into smaller sections. The activity of these enzymes is influenced by factors including cellular metabolism, life stage and nutrient availability (Walt *et al.*, 2012).

There are a number of well-developed, interoperable, robust tools and databases for gene and protein analysis (Brooksbank *et al.*, 2003; Goecks *et al.*, 2010; Smedley *et al.*, 2009). This is due in part to the templated synthesis of these polymers that make relatively robust methods for synthesis (notably PCR) and sequencing possible. The combined effect of the biochemical synthesis and conformationally complex, multi-branched nature of glycans makes the purification, extraction and computational analysis of the structural data far more challenging than in genomics and proteomics (Campbell *et al.*, 2014a; Raman *et al.*, 2006; Ranzinger *et al.*, 2008). Some tools developed for the genomic and proteomic paradigm have been applied to glycomics and used in the design and insertion of N-glycosylation sites on proteins (Mazola *et al.*, 2011).

Although limited efforts have been made to bridge glycomics with other fields, glycobioinformatics remains hamstrung by the complexity and diversity of oligosaccharide structures (Campbell *et al.*, 2014a). Superimposed on this is lack of sustained support from funding agencies for the development of glycomics methods and data repositories. The development of glycomics methods therefore relies upon independent and sporadically funded efforts. The resulting databases and toolsets therefore lack interlinks between similar and related tools. More serious is the lack of an emergent set of developmental standards amongst databases. Major online resources such as Consortium for Functional Glycomics (CFG) (Raman *et al.*, 2006), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Hashimoto *et al.*, 2006), GlycomeDB (Ranzinger *et al.*, 2008, 2011), Resource for Informatics of Glycomes at Soka (RINGS) (Akune *et al.*, 2010) and UnicarbKB (Campbell *et al.*, 2014b) store glycan data, translate between various formats and provide other useful tools.

A comprehensive overview of glycan sequence formats and databases has recently been reviewed, (Campbell *et al.*, 2014a) the limitations in current glycomics recognized and partial solutions proposed. For example, GlycomeDB (Ranzinger *et al.*, 2011) is designed to be a single web portal that provides access to glycan structures from several databases while UnicarbKB (Campbell *et al.*, 2014b) is a curated open access resource derived from EuroCarbDB that is itself a collaboration of several databases. Many web portals do not provide machine-readable web services; subsequently the Working Group on Glycomics Database Standards (WGGDS) initiated (i) the development of standards for inputs and outputs of glycomics web services and (ii) the development of guidelines to report glycomics data (Campbell *et al.*, 2014a; York *et al.*, 2014). The Resource Description Format (RDF) is a model for exchanging data on the web that uses subject-predicate-object expressions (triples) to express formal naming and to describe relationships between pieces of information. The glycan RDF subgroup defined GlycoRDF, an ontology used for exporting glycomics data via the Semantic web (Ranzinger *et al.*, 2015). The SPARQL Protocol and RDF Query Language (SPARQL) along with the RDF Query Language is used to query the RDF data from curated endpoints (Aoki-Kinoshita *et al.*, 2013a,b). The GlyTouCan (https://glytoucan.org/) initiative aims to standardize glycan data that will result in an accessible international glycan structure repository (Aoki-Kinoshita *et al.*, 2015).

Conducting an *in silico* systems biology experiment requires access to data repositories, computational tools and a knowledge base to connect these resources (Oinn *et al.*, 2004). Access to remote web services and ensuring reproducibility of computational experiments requires a working knowledge of computer scripting and programming. Workflows (Goecks *et al.*, 2010) are useful for streamlining data analysis where a repeatable process describing the organization of the individual steps can be automated. Computational analyses encompass tasks that retrieve data, transform or analyze it followed by a visualization of the results (Goble *et al.*, 2010). Workflows implemented in Taverna (Oinn *et al.*, 2004), Knime (Berthold *et al.*, 2008) and Galaxy (Goecks *et al.*, 2010) annotate each part of a computational experiment, that include the data source, data manipulation steps and assembly of necessary tools and parameters in a reusable template. Galaxy (Goecks *et al.*, 2010) provides a management system that requires no prior programming and web services experience. The workflows created on platforms such as this can be published using social research platforms such as myExperiment (Goble *et al.*, 2010).

2 Methods and implementation

Efforts to characterize glycomes make apparent the inherent gaps in accessibility to software, resources and databases. Methods to increase the reproducibility of *in silico* analysis are not widely

available. To address this, a single glycomics platform that is interoperable with multiple remote data sources and tools, provides tools for the design of reproducible and cohesive workflows, provides data sharing and addresses multiple glycan formats has been developed. This model relies on remote web resources that conform to web standards and provide a machine friendly web interface for sharing data and services. A REST or SOAP interface, or SPARQL endpoint provides effective programmatic access to web services. We recommend that all sites provide a web service interface with a standard API. Although web automation software techniques ('screen scraping') provide interaction with simple HTML websites, websites designed using AJAX, Java and JavaScript are not easily automated.

We have identified Galaxy as an existing open-source bioinformatics platform that meets these requirements. Galaxy is an open-source, web platform that incorporates workflow technology and in-built access to bioinformatics resources including remote data warehouses and tools. Several public instances of Galaxy are available and it can be installed locally or in the cloud. These multiple installation options denote that Galaxy is able to accommodate varying user requirements, for example - data security, no IT experience and no network transfer of data (depends on tools used). The Public Galaxy Toolshed is an appstore that hosts repositories containing Galaxy Tools, Galaxy Data Managers, custom Galaxy Datatypes and exported Galaxy Workflows. The Toolshed supplies tools to Galaxy and provides tool developers with a framework and common repository for sharing, updating and managing tools. To ensure that Galaxy is able to execute custom tools in a robust manner, a developer describes the underlying function of their custom tool via an XML framework contained in a tool definition file. Recently, a suite of open-source chemistry and cheminformatics tools (Open Babel (O'Boyle *et al.*, 2011)) has been integrated into Galaxy as part of the ChemicalToolbox (Lucas *et al.*, 2015).

We have created the GAP, Figure 1, which integrates existing glycoinformatics tools and resources along with our own glycomics utilities into the Galaxy platform. It contains tools for getting data listed under the 'GAP Get Data' heading, tools for manipulating data ('GAP Manipulate'), tools for joining or grouping data ('GAP Join, Subtract'), tools for converting between formats ('GAP Convert') and tools for extracting features from sets of glycans ('GAP Glycome Features').

We designed GAP by identifying glycan formats and web services pertinent to the analysis of cellular glycomes. This includes command-line Python scripts to interactively retrieve data from web services and in-house tools and utilities for glycan format manipulation. Wrapper XML scripts integrate these tools into Galaxy that are bundled together and installed to a Galaxy Instance via the Galaxy ToolShed. The XML scripts comprise Python module requirements that are automatically installed into a Python Virtual Environment. Testing was done using Python 2.7. While GAP can be used independently and is transferable to other workflow platforms, explicit interfaces and datatypes have been written for integration into Galaxy. Therefore these tools interoperate and function best on a Galaxy server with internet access.

Sources of individual glycan data as well as cell and tissue related glycomes are available from the CFG and databases such as KEGG, the GlycoGene Databank (Narimatsu, 2004), Glycosciences.de (Lutteke *et al.*, 2006) and GlycomeDB (Ranzinger *et al.*, 2008, 2011). The CFG contains glycan profiling, glycan array, gene-microarray and mouse phenotyping data. The glycan profiling data is derived from experiments that have used MALDI mass spectrometry and other analyses to characterize glycans from glycoconjugates

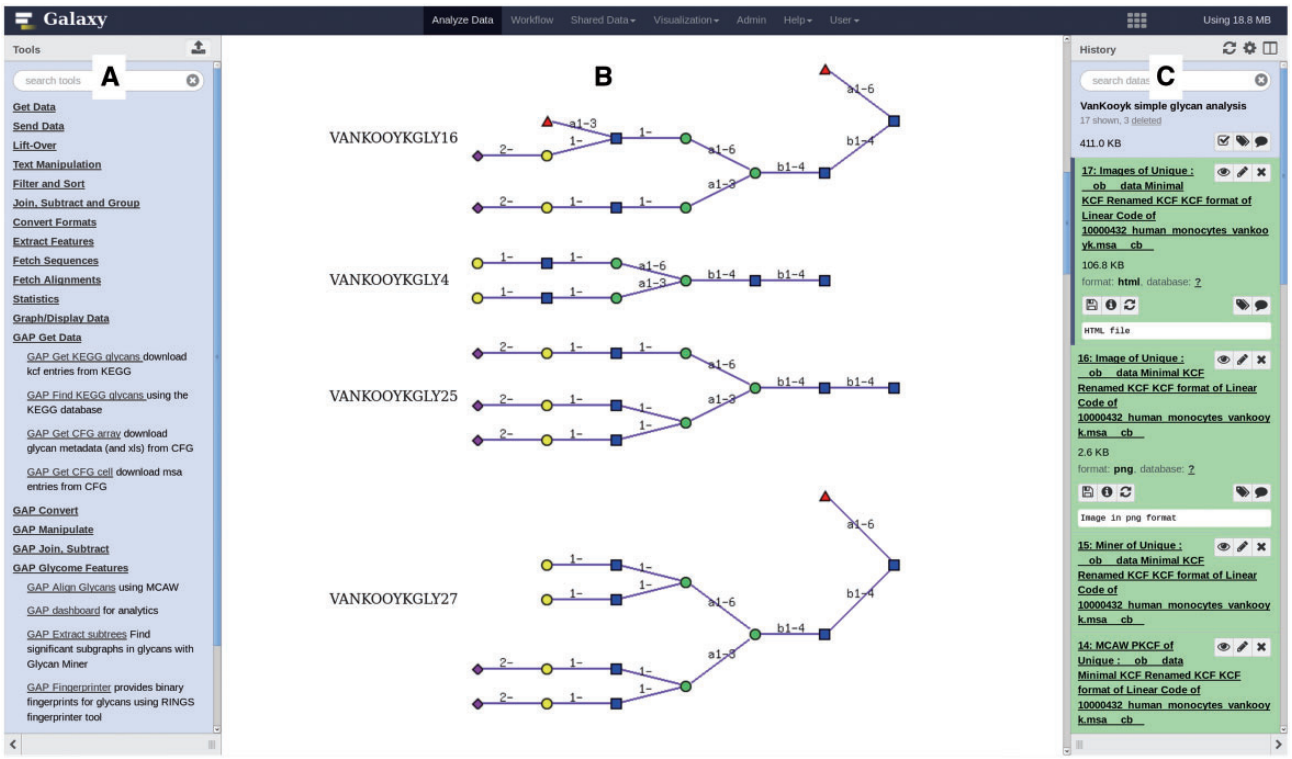

Fig. 1. The Galaxy interface with GAP. Access to workflows, shared data, help and user information is in the top panel. (**A**) Tools are in the left panel. (**B**) Selected data or tool information is shown in the large middle panel. (**C**) History is shown in the right panel. The GAP tools are shown in the tool panel. There are tools for accessing glycan data, converting formats, manipulating, joining, subtracting and grouping glycan data and for extracting glycome features

in cells and tissues. N-glycan, O-glycan, polar and non-polar glycolipid data from Human, Mouse and Chinese Hamster is available. The data from the CFG can be either manually accessed or via the 'Get CFG' tool within the GAP. The KEGG database contains numerous databases of systems, genomic, chemical and health information, for example KEGG GLYCAN, KEGG GENE, KEGG ENZYME, KEGG REACTION and KEGG PATHWAY. These are readily accessed through KEGG's REST interface that allows remote programmatic to its databases.

To illustrate KEGG's remote data manipulation and access via the GAP (Fig. 2), we build a workflow to access KEGG and search for glycan entries in KEGG GLYCAN that are related to the Enzyme ST6Gal1 (Beta-galactosamide alpha-2,6-sialyltransferase) found in KEGG ENZYME. The KEGG database cross-referencing tool is used to search for 'ec: 2.4.99.1' and return any glycans related to this enzyme. The glycan sequence data can then be downloaded in KEGG Chemical Function (KCF) format. Images of the glycan sequence can be created using 'KCF to Image'. Following this, variations in the glycan structures can be analyzed using tools such as MCAW (discussed later).

2.1 Data types and conversion

Glycans are considered as rooted, labelled, ordered tree structures (Aoki-Kinoshita, 2013a,b). Glycan sequence data is encoded using various formats and often contains ambiguous structures and repeating unit. These features cannot be encoded by all glycan formats (Campbell *et al.*, 2014a). There are several formats that linearize the glycan (IUPAC, LinearCode, LINUCS), while connection table approaches are also used (KEGG KCF, GlycoCT) and XML based formats (GlydeII, GlycoCT XML) have been adopted (Campbell *et al.*, 2014a). Recently the Web3 Unique Representation of Carbohydrate Structures (WURCS) format has been introduced. WURCS follows Semantic web guidelines and includes: (i) a linear notation which can be used as a URI (Uniform Resource Identifier) if needed and (ii) a unique notation such that any published glycan structure can be represented distinctively (Tanaka *et al.*, 2014).

We have designed support for the following glycan formats (data types) in Galaxy, KEGG KCF, GLYDEII, GlycoCT, GlycoCT (XML), LINUCS, LinearCode, WURCS, IUPAC (condensed) and MSA (annotated mass spec—contains LinearCode).

Galaxy is able to automatically deduce the type of input data using a 'sniffer' function; this function identifies signatures of a particular format for identification. We have designed 'sniffer' functions for each glycan data type. Initially the data is sent to the RINGS webservice. If this is unavailable then a simple file signature recognition code is called for each type. Failing this the user can upload data with the correct data type suffix, e.g. '.iupac', or specify the type. Data can be directly uploaded or copy-pasted into Galaxy using 'Get Data'. For example uploaded annotated mass spec profiles are automatically assigned as '.msa' datatype. Only the tools that support the '.msa' datatype as an input will show this data in their input selection panel.

Glycan builder (integrated with UnicarbKB) illustrates and interconverts between several glycan formats (for example it can import BCSDB and export GlycoCT XML format) (Damerell *et al.*, 2012). The RINGS web resource includes utilities for converting between glycan formats. There is a new converter tool which supports multiple format interconversions and specific interconversion utilities, for example 'LinearCode to KCF' and 'KCF to image'. Other resources are available (c.f. Table 3 (Campbell *et al.*, 2014a)) and we provide interfaces to these tools.

2.2 Analyzing glycomes

Glycan Miner is an efficient method for mining motifs or significant subtrees from a set of glycans that provides the novel ability to mine closed frequent subtrees (Aoki-Kinoshita, 2013a,b). This functionality depends on the definition of a support and a screening factor (alpha) to modulate the properties of the subtrees mined from the

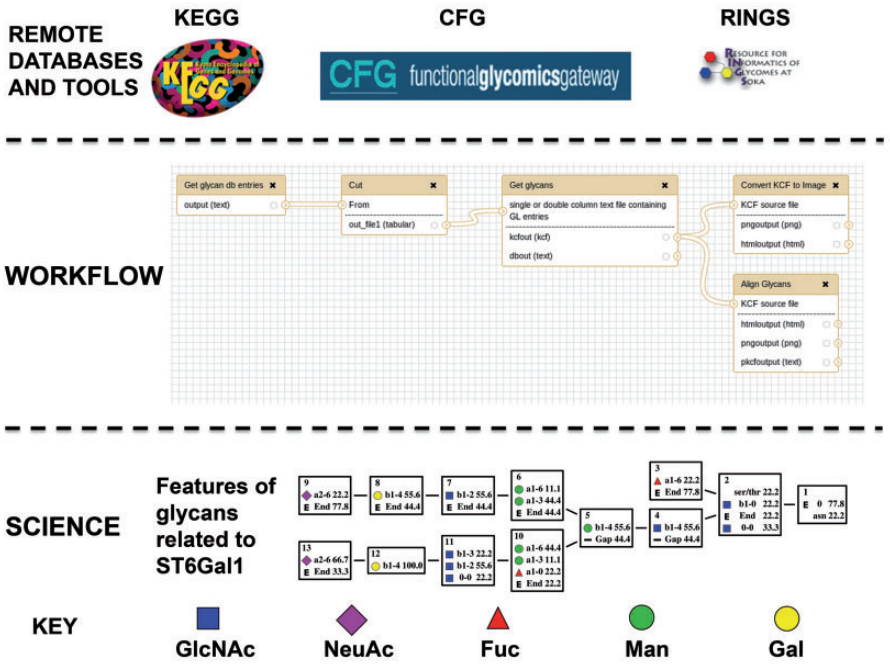

Fig. 2. The linking of remote databases and tools allows the creation of workflows which can be used to aid in scientific discovery. See Supplementary Figure S1 for further workflow examples. The key illustrates a commonly used symbol nomenclature for monosaccharides (the CFG notation). Abbreviations used in the key are as follows: N-Acetyl glucosamine (GlcNAc), N-Acetylneuraminic acid (NeuAc), Fucose (Fuc), Mannose (Man), Galactose (Gal)

set of glycans. For a given set of glycans G, the support of a glycan g is the number of glycans containing g. The glycan g is considered frequent if its support is greater or equal to a cut off value viz, *the minimum support* (minsup). The minsup value represents the (integer) number of glycans (Aoki-Kinoshita, 2013a, b; Hashimoto *et al.*, 2008).

A shortcoming of this mining algorithm is that the subtree frequency depends on the distribution of monosaccharides. Subsequently high frequency subtrees do not adequately represent structural significance. Thus, a control dataset gets generated from the input dataset and is used to predict and rank the significance of the mined subtree patterns (Aoki-Kinoshita, 2013a,b; Hashimoto *et al.*, 2008).

Multiple Carbohydrate Alignment with Weights (MCAW) aligns glycans using a tree alignment algorithm (Hosoda *et al.*, 2012). From an input of several glycans in KCF format (Hashimoto *et al.*, 2006), MCAW applies a pairwise glycan comparison algorithm and progressively builds a multiple glycan alignment. This Profile KCF (PKCF) alignment contains information on the alignment order, position and state of each node in the tree. The PKCF is rendered as an aligned glycan profile image presenting statistics of the alignment order, position and state of each node such that common patterns in the set of glycans can be identified. Multiple profiles can be compared to reveal glycan recognition patterns across differing glycomes.

The Glycoviewer (http://www.glycoviewer.babs.unsw.edu.au/) (Joshi *et al.*, 2010) produces similar multiple tree analysis to MCAW but is not included in this implementation. This is because it requires glycan sequences in IUPAC format that have fully specified linkage information and so does not function efficiently when input data has multiple unknown linkage information.

2.3 Additional utilities

To maintain all analyses as steps of a workflow inside of Galaxy, additional utilities were designed and implemented that overcome error conditions resulting from inconsistent glycan data formatting. This is required to maintain reproducible computational analyses.

We include a tool to extract LinearCode data from .msa files (supports several versions of the msa file format specification). This tool corrects format inconsistencies in the Linear Code, for example spaces and missing outer brackets. The KCF format has the flexibility to include complex ENTRY specifications or multiple REFERENCE and COMPOSITION entries this leads to the malfunctioning of several analysis tools. The GAP has the functionality that simplifies KCF by returning only the ENTRY NODE and EDGE specifications and a tool for renaming and numbering the ENTRY section.

2.4 Comparing sets of glycans

When comparing sets of glycans (virtual glycomes) the Miner tool is used to distinguish common subtrees in each set and the MCAW tool to overview the similarities and differences throughout the leaves and the core. Nevertheless, even when summarized, the complexity and multitude of glycans in a set can easily overwhelm the user and reduce the usefulness of a particular analysis.

Our approach for comparing the glycan sets A, B is as follows. Remove duplicate glycans in each set and return the unique sets A, B. Determine the glycans that are in common between these sets by the intersection set A∩B. Determine the glycans unique to A and unique to B by the relative differences A–B and B–A. Miner and MCAW analysis of the glycans in the relative difference sets A–B and B–A then more clearly presents the unique and significant features of these glycomes.

2.5 Creation of glycomics workflows

During exploratory analyses an environment that favors ad hoc discovery rather than predefined pipelines is preferred. This is supported within the GAP Galaxy implementation where the process is simplified through the generation of workflows from user histories. The dynamic exploration and analysis of data (user history) can be bundled into re-usable and shared workflows (predefined analysis pipelines). This was the method used to create a Simple Glycome Analysis workflow (Supplementary Fig. S1A). There, glycan msa data is loaded, converted to Linear Code, converted to KCF, cleaned. This data was then converted to an image. Duplicates were removed and a MCAW and Miner analysis was carried out.

An additional workflow was created for the comparison of two glycome datasets (Supplementary Fig. S1B). First, the Simple Glycome Analysis workflow is applied to both datasets to yield two sets containing unique glycan data. These are inputs for the new analysis. The intersection and relative difference of these sets is calculated, images of the glycans are created and MCAW and Miner analyses are carried out.

3 Results

Glycomes comprised of N-glycans from healthy human monocytic cells were analyzed. These were obtained from the CFG (*N-Glycans from Human Monocytes, Paul Crocker, cfg_rRequest_1686, 10000006.msa and N-Glycans from Human Monocytes, Yvette Van Kooyk, cfg_rRequest_1687, 10000432.msa*). These datasets are not duplicative and the cells were sourced from different laboratories. The data were loaded into Galaxy and the simple glycome analysis workflow was applied to both datasets. An additional workflow for comparing these sets was then applied. A MCAW and miner analysis were carried out on the common glycans (intersection set) and on the glycans specific to the Crocker set.

There are 26 unique glycans (all have a unique sequence) in the Crocker set (excluding the ambiguously attached fragments found on certain glycans) and 25 unique glycans in the Van Kooyk set. A miner analysis (alpha: 1.0, minsup: 5) found 7 subtrees in the Crocker set and 8 subtrees in the Van Kooyk set. The subtree with the most support (it is found in all glycans in the set) is the common N-glycan core. In general the subtrees between these sets are similar and found in similar proportions, see Supplementary Figure S2. The exceptions are subtree 4 which is not found in the Van Kooyk sample and subtrees 8 and 9 which are not found in the Crocker sample. Subtrees 8 and 9 include additional sialylation at the leaves.

A MCAW analysis of the Crocker and Van Kooyk datasets shows the conserved common N-glycan core and variability in the leaves (Fig. 3). From these MCAW analyses we see that the Crocker set has triantennary glycans not found in Van Kooyk's data (see node 18 in Fig. 3A). Note that according to the CFG glycan profiling protocols low level MS is performed and wherever possible ES-MS/MS or MALDI-TOF/TOF-MS/MS is used to further identify the glycan sequence. This profiling is dependent on the experimentalist and the annotator of the dataset. It is important to note that although the triantennary glycans are not seen in Van Kooyk's data it does not mean that they do not exist. The Van Kooyk data has variability in the mannose linkages at depth 4 (nodes 4 and 10 in

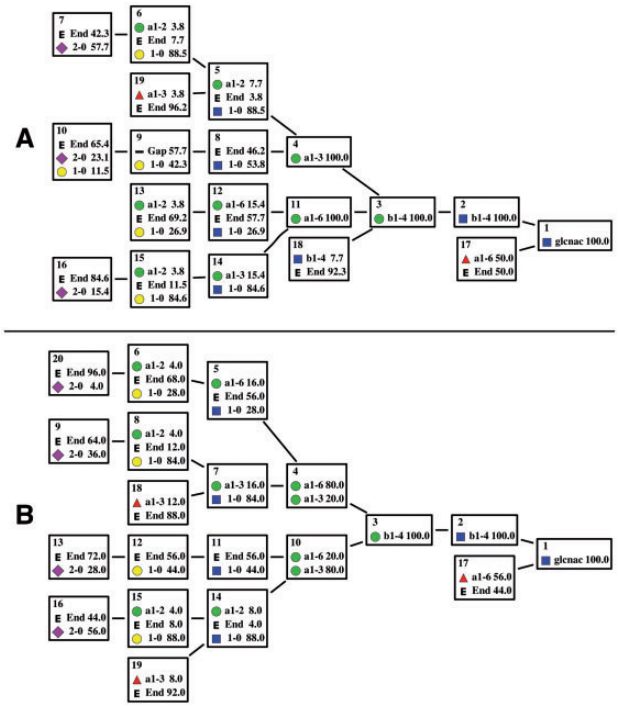

Fig. 3. MCAW analysis. The Crocker set (**A**) and the Van Kooyk set (**B**). If not indicated the anomeric information is not known. A '0' indicates a wildcard and denotes that the linkage information is not known. 'E' denotes the end of a sequence, that is a certain percentage of aligned glycans terminate. Numbers are in percent, for each node these sum to unity

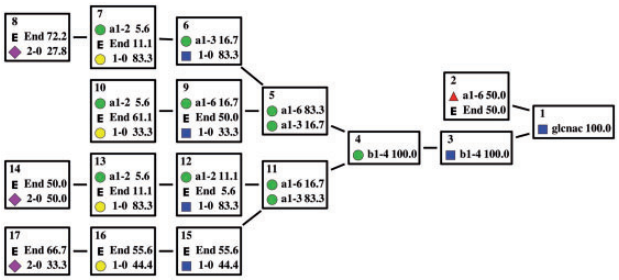

Fig. 4. MCAW analysis of the intersection of Crocker and Van Kooyk set. Half of these glycans are core-fucosylated. The Sialyl Lewis and Lewis motifs are not present

Fig. 3B); although this may be an artifact of the ordering process in the alignment algorithm as 21 of 25 structures are composed of subtree 3 (Supplementary Fig. S2). This data also shows additional fucosylation (nodes 18 and 19, Fig. 3B) and sialylation at the leaves (nodes 20, 9, 13, 16, Fig. 3B). From application of the comparison workflow, consider the MCAW analysis of the *intersection set*, the glycans common to both sets (Fig. 4). There are 18 glycans in common and these are not fucosylated at the leaves, although there is sialylation (nodes 8, 14, 17 in Fig. 4). Half of these glycan sequences are found to be core-fucosylated. The Lewis and sialyl Lewis-like motifs are not found to be in common.

To investigate the glycans that are unique to each set the relative difference between the sets is calculated. 8 unique glycans are found in the Crocker set and 7 unique glycans in the Van Kooyk set as shown in Figure 5 and Supplementary Figure S3 (Inspection of the data, reveal however, that glycan 25 with a predicted mass of 3691.1,

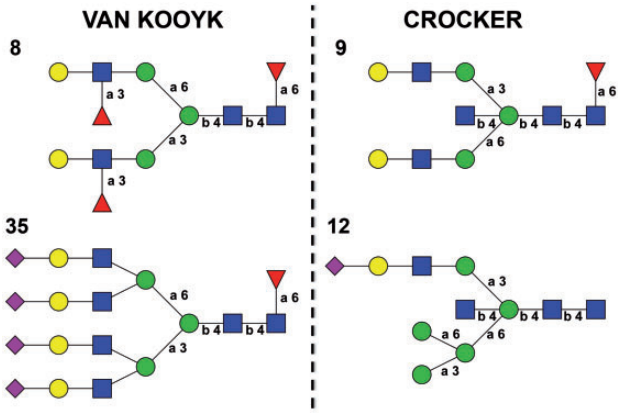

Fig. 5. Selected unique glycans from the Van Kooyk and Crocker sets. Each glycan is numerically labelled according to the ID generated during the analysis. Alpha and Beta glycosidic linkages are indicated as a and b in the figure. Images rendered using the GlycanBuilder (Damerell *et al.*, 2012)

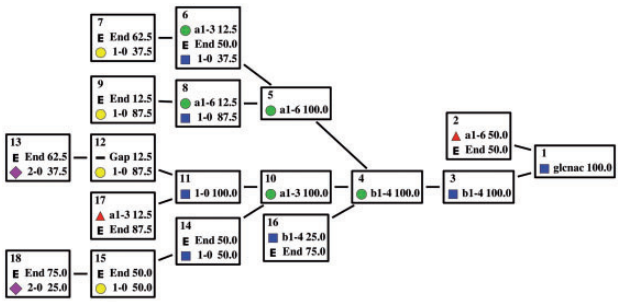

Fig. 6. MCAW analysis. The glycans unique to the Crocker set. '–' denotes a gap, that is a certain percentage of glycans do not contain the specified node contents, rather the parent of this node is directly connected to its grandchild

is a relatively minor peak in the MALDI MS data profile shown for Crocker dataset, and that mass appears to be present but not annotated in the van Kooyk dataset.). The features of the unique glycans derived from the Crocker set are then summarized using MCAW, see Figure 6.

The N-glycan set from the Crocker study was found to have additional bisected glycans, see Figures 5 and 6. The Van Kooyk set included additional glycans with sialylation at the leaves and more core fucosylation. The MCAW analysis of the Crocker set (Figs. 3A and 6) gives the impression that Sialyl Lewis motif may be present, but when considering the individual glycans this is not the case. Instead only the Lewis motif is found on one of these glycans. In contrast, the Van Kooyk set contains one glycan with Sialyl Lewis, two with Lewis and one with two Lewis motifs. Inadequate linkage information is available to concretely identify the Lewis motifs, although it is most likely Lewis^{X} (Lo *et al.*, 1997). A MCAW and miner analysis of these sets more easily identifies the unique aspects of each glycome.

4 Conclusions

The use of reproducible research methods for glycomics analysis is key to the acceleration of glycome assembly in diseased and healthy cellular states. Once a glycome is assembled it can be mined for biomarker discovery. The GAP introduced here offers the opportunity for open source development that integrates multiple web resources into

reproducible workflows using Galaxy. The GAP's integration of glycomics tools into an established bioinformatics web platform to improve their accessibility, usability, re-usability improves the sharing and reproducibility of computational analysis and data in glycomics. We illustrated the analyses and pattern matching capabilities within GAP on N-glycan sets found in two healthy monocyte cell studies. From this 18 glycans were found that were common to both sets. The GAP analyses revealed that the N-glycan set from the Crocker study has additional bisected glycans; while that from Van Kooyk's study was found to have additional glycans with sialylation at the leaves and more core fucosylation as well as the Sialyl Lewis and Lewis motifs.

In the future we intend to connect semantic web services and include access to UnicarbKB and GlyTouCan. Linking to glycogene information for each sample would aid in an improved understanding of this data.

Funding

This work is based in part upon research supported by the South African Research Chairs Initiative (SARChI) of the Department of Science and Technology (DST) and National Research Foundation (NRF) grant 48103 (KJN) and the National Bioinformatics and Functional Genomics (NBIG) grant 86944 (KJN). CBB thanks the University of Cape Town's Research Committee (URC) for travel and contact support to Soka University; this publication is based on research that has been supported in part by the University of Cape Town's Research Committee (URC) and the NRF grant, 87956 (CBB).

Conflict of Interest: none declared.

References

Akune,Y. *et al.* (2010) The RINGS resource for glycome informatics analysis and data mining on the Web. *Omics: J. I. Biol.*, **14**, 475–486.

Aoki-Kinoshita,K. (2013a) Mining frequent subtrees in glycan data using the rings glycan miner tool. In: Mamitsuka,*et al.* (eds) *Data Mining for Systems Biology*. Humana Press, London, UK, pp. 87–95.

Aoki-Kinoshita,K.F. *et al.* (2013b) Introducing glycomics data into the Semantic Web. *J. Biomed. Semant.* BioMed Central, London, UK, **4**, 39.

Aoki-Kinoshita,K. *et al.* (2015) GlyTouCan 1.0 – The international glycan structure repository. *Nucleic Acids Res.* **D1**, pp. D1237–D1242.

Berthold,M. *et al.* (2008) KNIME: The Konstanz Information Miner. In: Preisach,C. *et al.* (eds) *Data Analysis, Machine Learning and Applications*. Springer, Berlin, Heidelberg, pp. 319–326.

Bertozzi, C.R.Sasisekharan,R. (2009) Glycomics. In: Varki,A *et al.* (eds) *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.[TQ2]

Brooksbank,C. *et al.* (2003) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, **31**, 43–50.

Campbell,M.P. *et al.* (2014a) Toolboxes for a standardised and systematic study of glycans. *BMC Bioinformatics*, **15**, S9–S9.

Campbell,M.P. *et al.* (2014b) UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.*, **42**, D215–D221.

Cummings,R.D. (2009) The repertoire of glycan determinants in the human glycome. *Mol. bioSyst.*, **5**, 1087–1104.

Damerell,D. *et al.* (2012) The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biol. Chem.*, **393**, 1357–1362.

Dube,D.H. and Bertozzi,C.R. (2005) Glycans in cancer and inflammation [mdash] potential for therapeutics and diagnostics. *Nat. Rev. Drug Discov.*, **4**, 477–488.

Goble,C.A. *et al.* (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.*, **38**, W677–W682.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Hashimoto,K. *et al.* (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63R–70R.

Hashimoto,K. *et al.* (2008) Mining significant tree patterns in carbohydrate sugar chains. *Bioinformatics (Oxford, England)*, **24**, i167–i173.

Heimburg-Molinaro,J. *et al.* (2011) Preparation and analysis of glycan microarrays. *Curr. Protoc. Protein Sci.*, Unit12.10.

Herget,S. *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**, 35-35.

Hirabayashi,J. *et al.* (2013) Lectin microarrays: concept, principle and applications. *Chem. Soc. Rev.*, **42**, 4443–4458.

Hosoda,M. *et al.* (2012) Alignment with weights applied to carbohydrates to extract binding recognition patterns. In: Shibuya,T. *et al.* (eds) *Pattern Recognition in Bioinformatics*. Springer, Berlin, Heidelberg, pp. 49–58.

Joshi,H.J. *et al.* (2010) GlycoViewer: a tool for visual summary and comparative analysis of the glycome. *Nucleic Acids Res.*, **38**, W667–W670.

Laine,R.A *et al.* (2008) The Information-Storing Potential of the Sugar Code. In: Gabius,H.J. *et al.* (eds) *Glycosciences: Status & Perspectives*. Wiley-VCH Verlag GmbH, Weinhem, pp.1–14.

Li,L. *et al.* (2010) A weighted q-gram method for glycan structure classification. *BMC Bioinformatics*, **11**, S33.

Lo,S.K. *et al.* (1997) Engagement of the Lewis X Antigen (CD15) results in monocyte activation. *Blood*, **89**, 307–314.

Lucas,X. *et al.* (2015) The Purchasable Chemical Space: A Detailed Picture. *Journal of Chemical Information and Modeling*, **55**, 915–924.

Lutteke,T. *et al.* (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, **16**, 71r–81r.

Mazola,Y. *et al.* (2011) Integrating bioinformatics tools to handle glycosylation. *PLoS Comput. Biol.*, **7**, e1002285.

McNaught,A.D. (1997) Nomenclature of carbohydrates (recommendations 1996). *Adv. Carbohydr. Chem. Biochem.*, **52**, 43–177.

Narimatsu,H. (2004) Construction of a human glycogene library and comprehensive functional analysis. *Glycoconjugate J.*, **21**, 17–24.

North,S.J. *et al.* (2010) Chapter 12 – Mouse and human glycomes. In: Pierce,R.D.C.M. (ed) *Handbook of Glycomics*. Academic Press, San Diego, pp. 263–327.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminf.*, **3**, 33.

Oinn,T. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England)*, **20**, 3045–3054.

Pierce,J.M. (2010) Chapter 16 – Cancer glycomics. In: Pierce,R.D.C.M. (ed) *Handbook of Glycomics*. Academic Press, San Diego, pp. 397–429.

Raman,R. *et al.* (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, **16**, 82r–90r.

Ranzinger,R. *et al.* (2015) GlycoRDF: an ontology to standardize glycomics data in RDF. *Bioinformatics (Oxford, England)*, **31**, 919–925.

Ranzinger,R. *et al.* (2011) GlycomeDB—a unified database for carbohydrate structures. *Nucleic Acids Res.*, **39**, D373–D376.

Ranzinger,R. *et al.* (2008) GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinformatics*, **9**, 384-384.

Redelinghuys,P. and Crocker,P.R. (2010) Chapter 11 – Glycomics of the immune system. In: Pierce,R.D.C.M. (ed) *Handbook of Glycomics*. Academic Press, San Diego, pp. 235–261.

Satomaa,T. *et al.* (2009) Analysis of the human cancer glycome identifies a novel group of tumor-associated N-acetylglucosamine glycan antigens. *Cancer Res.*, **69**, 5811–5819.

Smedley,D. *et al.* (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Tanaka,K. *et al.* (2014) WURCS: The Web3 Unique Representation of Carbohydrate Structures. *J. Chem. Inf. Model.*, **54**, 1558–1566.

Walt,D. *et al.* (2012) *The National Academies Collection: Reports Funded by National Institutes of Health. In, Transforming Glycoscience: A Roadmap for the Future*. National Academies Press (US) National Academy of Sciences, Washington, DC.

York,W.S. *et al.* (2014) MIRAGE: The minimum information required for a glycomics experiment. *Glycobiology*, **24**, 402–406.