http://www.glycopedia.eu/e-chapters/

A Traveler's Guide to Complex Carbohydrates in the Cyber Space

David Alocci, Frederique Lisacek & Serge Pérez

https://glycopedia.eu/e-chapters/a-traveler-s-guide-to-complex-carbohydrates-in-the-cyber-space/Glyco-Cyber-Space

Contents

1. ABSTRACT

Glycoscience is a rapidly developing and emerging scientific discipline. Like many other scientific disciplines, glycoscience is adapting to the exciting rise of accessible scientific data, which now impacts research and modifies its practice. The accumulation of information along with the development of enabling technologies has laid the foundation of a rich computational toolbox tailored for the detection and high-resolution determination of complex glycans. In parallel, a variety of online resources essentially in the form of databases covering glycan and glycoproteins structures have been developed by independent research groups worldwide. At present, more than 150 entries are freely available on the internet yet these often produced independently of one another. With the aim of facilitating glycoscience research, we have clustered these different tools according to their major field of applications. As a result, the following entries can be accessed: Portals. Genome and Glycome; Representations; Experimental Results; Glycans; Glycoproteomics; Functional Glycomics; Glycolipids; CAZYmes; Polysaccharides

Cross-talk between these computational resources is needed. To illustrate this point, one section of the chapter is devoted to the practical usage of integrative tools to guide the traveler in the navigation, investigation and the quest for correlations between structure and function in glycobiology.

Some fundamental principles of bioinformatics about data handling are presented in three consecutive annexes. These cover: Data Integration, Data Integration Strategies and their implementation in Bioinformatics.

2. INTRODUCTION

Two recently published monographs, "*A road-map for Glycoscience in Europe*" and "*Transforming Glycoscience : A Roadmap for the future*" published respectively under the auspices of the European Science Foundation and the National Academies USA, identified a selected number of goals. One of particular importance was the need for the "establishment of long-term databases and bioinformatics and computational tools to enable accurate carbohydrate and glycoconjugate structural predictions". This recommendation highlights a major challenges facing Glycoscience namely, the development and implementation of robust and validated informatics toolbox enabling accurate and fast determination of complex carbohydrate sequences extendable to 3D prediction, computational modeling, data mining, and profiling.

Another key challenge for Glycoscience is to pull it out of its isolation. Glyco-molecules are functionally important in many biological processes yet their occurrence, let alone their role, is rarely considered. Nonetheless the recommended shift towards bioinformatics as stated above, is a chance to bridge Glycoscience with other fields. Indeed, the concomitant expansion of stable and integrated databases, cross-referenced with popular bioinformatics resources should contribute to connecting glycomics with other –omics.

Proteomics, transcriptomics and genomics are at the core of biology and medicine whereas lipidomics, metabolomics and glycomics seldom are taken into account.

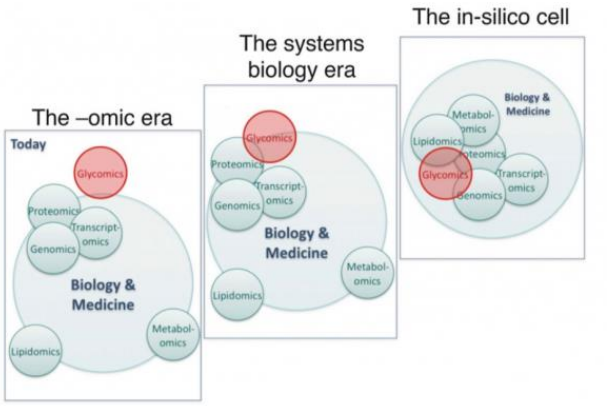


Figure 1. A prospect of the integration of 'omics' research in Biology and Medicine. Courtesy of "Databases and Associated Tools for Glycomics and Glycoproteomics" (Lisacek et al., 2016).

Bioinformatics played an essential role in the integration of genomics, proteomics and transcriptomics (Manzoni et al., 2018). The development of an integrated approach to assemble and annotate newly sequenced genomes from transcriptome and proteome data is just one example of the importance of bioinformatics (Prasad et al., 2017). However, other issues like the extent of proteoforms (Aebersold et al., 2018) or that of the metabolome still hamper data integration, leaving the puzzle incomplete. Furthermore, several other -omics fields such as lipidomics or glycomics are remote in this picture mainly because of technical and knowledge gaps. This is particularly the case in glycomics that remains internally fragmented and misses solid bioinformatics support. Only the integration of all the 'omics' will lead to a in-silico simulation on a living cell.

The upcoming scientific production in Glycoscience amounts to the yearly publication of about 70,000 research articles. Amplified by the availability of sophisticated and powerful high-performance computing and searching capacities, the space of accessible information has substantially increased such that data mining opens new prospects of discovery. This new view not only emphasises the worth of analyzing raw data from published work, but also points at untapped wealth that may be harvested in collected data sets from which extracted information can be transformed into knowledge. The field of structural glycobiology has partially benefited from such advances with the development of tools and databases for structural analysis of carbohydrates. A variety of other online resources mainly in the form of databases covering glycan and glycoproteins structures, enzymes responsible for their biosynthesis and degradation, glycan binding to human pathogens, glyco-epitope and their antibodies,... has been developed by independent research groups, worldwide.

The present chapter covers the description of resources, in the form of tools and databases, that are freely available (most of them being web-based) and are regularly updated and improved. With the aim of facilitating glycoscience research, we have clustered these different tools according to their major field of applications. As a result, the following entries can be accessed : Portals ; Genome and Glycome ; Representations ; Experimental Results ; Glycan ; Glycoproteomics ; Functional Glycomics ; Glycolipids ; CAZYmes & Polysaccharides.

3. GLYCOMICS

Carbohydrates are a class of molecules utterly crucial for the assembly of complex multicellular organisms which requires interactions among cells. Beside the mediator role in cell-cell, cell-matrix, and cell-molecule interactions, glycans play an essential function in host-pathogen interaction. This is not surprising since all types of cell in nature are covered by a glycocalyx, a sugar shell that could reach the thickness of 100 nm at the apical border of some epithelial cell (Le Pendu et al., 2014).

Cell membranes are built by several molecules, many of which are carrying a set of carbohydrates (glycoconjugates). Each of these sugars consists of a single (monosaccharides) or multiple (oligosaccharide) units. The presentation follows the rule set by the Essential of Glycobiology, which uses the term "glycan" to "refer to any form of mono-, oligo-, polysaccharide, either free or covalently attached to another molecule.

Glycans are linear or branched molecules where each building block is linked to the next using a glycosidic bond. The units which compose glycans are called monosaccharides, and they are identified as units since they cannot be hydrolysed to a simpler form (Varki et al., 2010). A list of all monosaccharides

available is shown. However, each biological system can only use a subset of those.

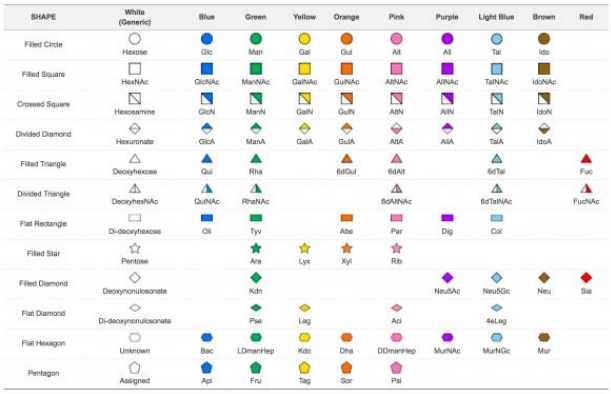


Table 1: The SNFG encoding table for monosaccharide. Each row represents a class of monosaccharide which are represented by one specific shape (in white). Each column, instead, identifies a type of monosaccharide. For example, glucose and all its modifications are coloured in blue. This table contains all monosaccharide found in Nature and has been extracted from "Symbol Nomenclature for Graphical Representation of Glycans" (Varki et al. 2015).

Depending on the stereochemistry of the two building blocks, the glycosidic linkage is defined as α-linkage (same stereochemistry) or β-linkage (different stereochemistry). A variation on the linkage type has a major impact on structural properties and biological functions of oligosaccharides with the same composition. This is evident for starch and cellulose which share the same composition but have a different type of linkages (α1-4 for starch, β1-4 for cellulose). Besides the linkage type, building blocks have multiple attaching points leading to an extensive collection of possible glycan structural configurations.

Elucidating glycan structures is crucial to determine whether they can be recognized by a glycan-binding protein (**GBP**). These proteins are scanning glycan structures searching for a specific region, defined as glycan determinant. We can imagine this process as the key and lock mechanism introduced by Emil Fischer in 1894 (Cummings, 2009). Blood group antigens (A, B and O) are the first example of glycan determinants, and they give an idea of the importance of glycobiology.

Glycans can freely stand in body fluids or be attached to other molecules. The combination of a single or multiple glycan(s) with a molecule is named glycoconjugate. In particular, when sugars are attached to proteins, we define the whole as glycoproteins whereas the combination of carbohydrates and lipids are called glycolipids. This article focus on glycans and glycoproteins as a start of a work which will eventually evolve towards glycolipids and proteoglycans.

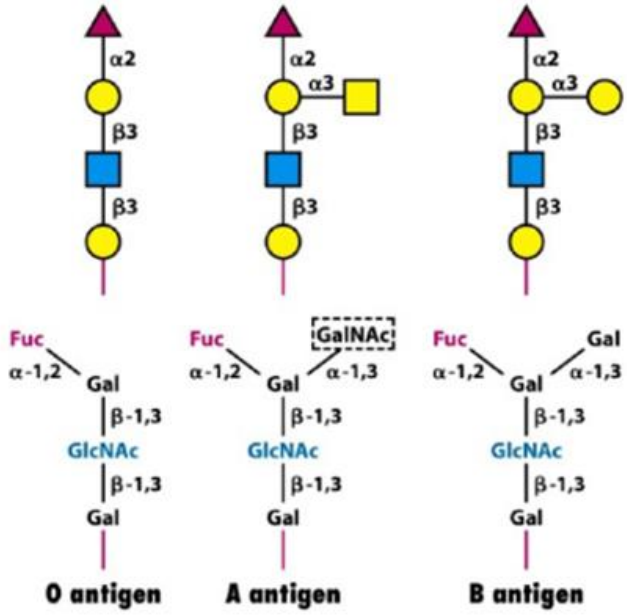


Figure 2: ABO antigens exist on glycoproteins and glycolipids in red cell membranes and also on most cells and tissues in humans, and in animal tissues.

Branched glycans attached to proteins are divided into two main groups: N-linked and O-linked. In the former group, glycans are attached to an asparagine which has to be located in a specific consensus sequence, usually Asn-X-Ser or Asn-X-Thr (in rare cases Asn-X-Cys). In the latter group instead, sugars are attached to serine, threonine, tyrosine, hydroxylysine or hydroxyproline (Lauc et al., 2016). N-linked glycans have a common pentasaccharide core (three Mannose and two N-Acetylglucosamine), and they can be classified into three groups : high-mannose, complex and hybrid. On the contrary, O-linked glycans have eight types of cores which start with an N-Acetylgalactosamine.

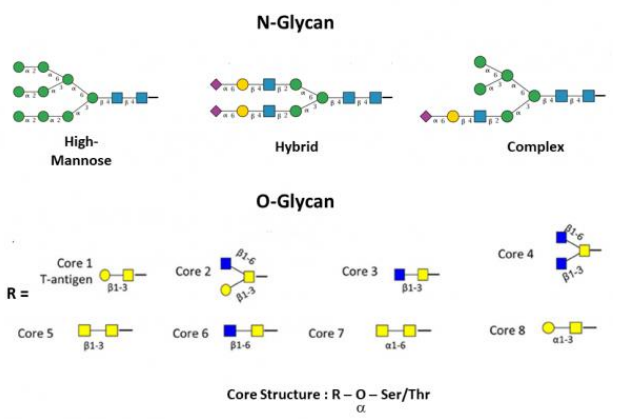


Figure 3: The N-glycan section shows the three types of N-linked glycans : high-mannose, complex and hybrid. The O-glycan section, instead, presents the eight type of cores available. All structures follow the SNFG standard.

The collection of all glycan structures expressed in a specific biological system is defined as the glycome (Packer et al., 2008). In analogy with genomics and proteomics, the systematic study of the glycome is called glycomics. In contrast to proteins, glycans cannot be directly predicted from a DNA template. Also, technical difficulties to correctly determine their complex and branched structures are the reasons why glycomics is lagging far behind other "omics".

A typical glycomics experiment releases and catalogs all the glycan structures present in a cell, tissue, etc. using techniques like mass spectrometry. In a more sophisticated version of this experiment, glycans are analysed while still attached to trypsin-digested peptides (glycoproteomics experiments). In addition to technical difficulties in elucidating complete structures, glycans are studied separately from their environment losing all the information about the actual landscape (Varki et al., 2010). Although qualitative techniques are improving, the quantitative aspect remains somewhat untouched, since quantitative experiments are still bound to composition more than structure elucidation.

4. DATA INTEGRATION GLYCOMICS

This section presents the steps needed to integrate biological data within glycomics and with other "omics".

Glycan formats. Glycans are inherently more complex than nucleic acids and proteins, so defining a format to store the molecular information correctly is not a trivial problem. The complexity of the glycans resides in their branched structure and the collection of building blocks available. In contrast with proteins and nucleic acids which are made of respectively 4 and 20 building blocks, glycans can be built with many different monosaccharides. Additionally, information about monosaccharide anomericity, residues modification and substitution, glycosidic linkages and possible structure ambiguities must be taken into account. Without commenting on the different nomenclatures available to represent each monosaccharide, encoding a glycan structure into a file is required.

The use of several formats has made data integration across different sources almost impossible. To tackle this problem, bioinformaticians have developed a collection of tools to parse and translate glycan structures across different encoding systems. Four main projects have been focus on data integration and tool development : RINGS (Akune et al., 2010, Glycoscience_de (Lutteke et al., 2006), EuroCarbDB (von der Lieth et al., 2011) and GlycomeDB (Ranzinger et al., 2008) which now has been replaced by GlyTouCan (Tiemeyer et al., 2017). In the recent years, different databases like MatrixDB (Launay et al., 2015), Glyco3D (Perez et al., 2015, 2016), UniLectin (Bonnardel et al., 2018) and the Glycomics@ExPASy initiative have introduced GlycoCT as a standard encoding format.

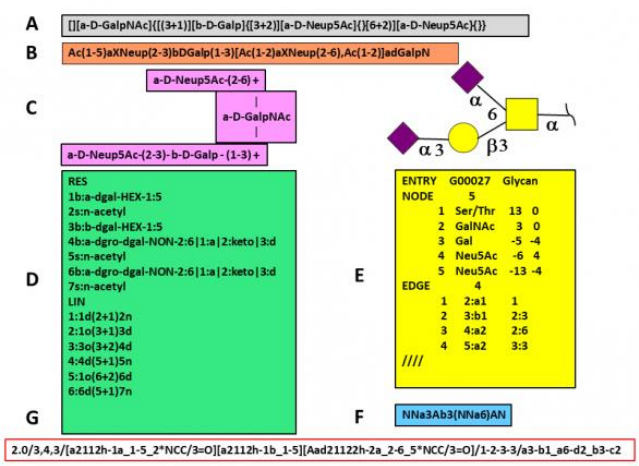


Figure 4: An O-Glycan (Glyconnect ID 2641 and GlyTouCan accession number G01614ZM) encoded with seven different glycan format. (A) LINUCS sequence format as used in GLYCOSCIENCES.de. (B) BCSDB sequence encoding. (C) CarbBank sequence format. (D) GlycoCT sequence format as used in GlycomeDB and UniCarbDB. (E) KCF format used in the KEGG database. (F) LinearCode® as used in the CFG database. (G) WURCS format used in GlyTouCan. Adapted from Toolboxes for a standardised and systematic study of glycans (Campbell et al. 2014).

Glycan graphical representations; Data formats presented in the previous paragraph are essential for storing information and exchanging data across software applications but are not suitable for humans.

Consequently, glycobiologists have proposed different graphical representations where symbols or chemical structures replace monosaccharides. A collection of depictions of the O-glycan presented above is given. Each of these representations has some peculiarities. The chemical depiction is used by researchers that are interested in glycan synthesis or use NMR to elucidate glycan structures. The Oxford nomenclature (Harvey et al., 2011) defines linkages using angles and encodes monosaccharide anomericity with dashed or solid lines.

The nomenclature used by the first version of Essential of Glycobiology encodes monosaccharides using shapes and colour is the most user-friendly. With the publication of the third edition of Essential of glycobiology (Varki et al.,2017) have made an effort to define a standard nomenclature called the Symbol Nomenclature for Glycans (SNFG) which should replace previous graphical representations (Varki et al., 2015).

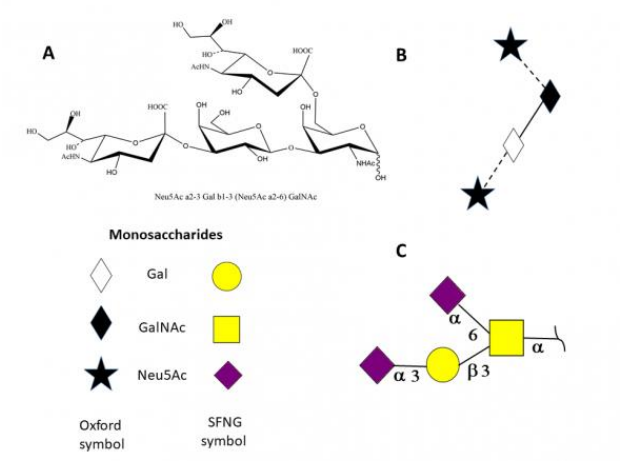


Figure 5: ABO antigens exist on glycoproteins and glycolipids in red cell membranes. ABO antigens exist on glycoproteins and glycolipids in red cell membranes Graphic representations of the O-glycan. (A) Chemical representation. (B) Oxford cartoon representation. (C) SNFG representation. The legend shows the difference encoding of monosaccharides between Oxford and SNFG. Adapted from Toolboxes for a standardised and systematic study of glycans (Campbell et al. 2014).

Glycan composition format. In many cases, experimental techniques can only distinguish glycan compositions, losing the information about the 2D structure. A glycan composition, indeed, reports the amounts of different monosaccharides present in the glycan without saying anything about their position in the space. An example is H2N1S2 which refers to a glycan with 2 Hex, 1 HexNAc, and 2 NeuAc. Sometimes, in addition to the composition, it is possible to extract information about the number of antennas, the presence of a bisected HexNAc and the amount of galactose. In this case, the researcher uses a nomenclature called Oxford which allows encoding of these additional data with the composition. This nomenclature should not be confused with the Oxford graphical representation for glycan structure described in the previous paragraph. In Oxford nomenclature, A2G2S2 refers to a glycan which has 2 antennas, 2 galactose and 2 NeuAc resulting in a composition of 5 Hex, 4 HexNAc, and 2 NeuAc. However, Oxford nomenclature works only with N-glycans and there are many dialects which are currently in use.

Identifiers. The community acceptance of unique identifiers for genes and proteins has facilitated data integration and data exchange boosting the research in genomics and proteomics. At the same time, some initiatives have taken place to fill this gap in glycomics. In 1989, the Complex Carbohydrates Structure Database, even known as CarbBank, (Doubet et al.,1989) was the first attempt to establish unique identifiers for glycan structures. However, in the late 1990s, CarbBank project ceased due to the end of funding. Consequently, different initiatives like CFG and KEGG have tried to continue CarbBank mission leading to a multiplicity of identifiers for each glycan structure. From 2013, researchers in glycomics brought back the need for a glycan structure registry which would ease data sharing and increase data integration among different platforms. Following this pressing wish, in 2015, Aoki-Kinoshita et al., (2016) unveiled GlyTouCan, a central registry for glycan structures. GlyTouCan allows users to deposit glycan structures in exchange for a unique identifier. Since different glycomics techniques cannot fully elucidate glycan structures, GlyTouCan accepts submissions of incomplete structures. Recently, the glycomics community has strongly endorsed the work of Aoki-Kinoshita to secure the stability of the registry which has become a crucial resource in the glycomics panorama. Every scientist in glycomics is encouraged to submit glycan structures to the registry and use the unique identifiers in reports and manuscripts (Tiemeyer et al., 2017).

Reporting Guidelines. The innate complexity of glycan structures, often, needs orthogonal experimental techniques to be elucidated. In this scenario, each experiment solves only a part of the puzzle, and only the integration of multiple data source leads to an accurate annotation. Due to this fact, experimental guidelines are necessary to publish datasets that can be easily interpreted, evaluated and reproduced by other glycoscientists. From 2011, experts in the field of glycobiology, glycoanalytics and glycoinformatics have been working together, under the patronage of the Beilstein Institute, to define **the Minimum information Required for A glycomics Experiment** (MIRAGE) (York et al., 2014). This initiative follows the more popular initiatives MIAME and MIAPE which we have already discussed. At the time of writing, MIRAGE already has published guidelines for sample preparation , mass spectrometry analysis (Kolarich et al., 2013) and glycan microarray analysis (Liu et al. 2017). Additionally, a guideline for liquid chromatography analysis is in preparation.

Ontologies. The combination and integration of multiple experimental and knowledge-based sources are essential to define the role of glycans. However, data are spread across different databases which act as "disconnected islands". Ontologies provide a painless way to interconnect resources within glycomics and with other "omics". Sahoo et al. generated GlycO, "a glycoproteomics domain ontology for modelling the structure and functions of glycans, enzymes and pathways" (Sahoo et al., 2006). GlycO has a strong focus on the biosynthesis of complex glycan structures and their relationships with proteins, enzymes and other biochemical entities.

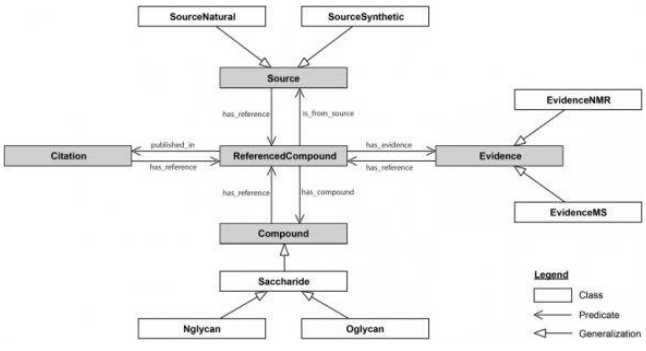


Figure 6: The diagram of the core classes of the GlycoRDF ontology. Grey and white boxes respectively identify classes and subclasses. Courtesy of GlycoRDF : an ontology to standardize glycomics data in RDF (Ranzinger et al. 2015).

Ranziger et al. developed GlycoRDF (Ranzinger et al., 2015). Contrary to GlycO, GlycoRDF has been designed with the precise goal of integrating all the information available in glycomics resources limiting the development of multiple RDF dialects. A detailed diagram of GlycoRDF is in Figure 6.

Visualisation. Despite its importance, data visualisation is still a challenge in glycomics. In the last decade, some initiatives have pushed the development of visual tools to improve some aspects of glycan identification and quantification. Glycoviewer (Joshi et al., 2010) is the first example of data visualisation tool which allows glycoscientists to visualise, summarise and compare different glycomes. GlycomeAtlas (Konishi & Aoki-Kinoshita, 2012), provides an interactive interface for exploring data produced by the Consortium of Functional Glycomics (CFG). To conclude, as stated in its website, GlycoDomainViewer (Joshi et al., 2018) is a visual "integrative tool for glycoproteomics that enables global analysis of the interplay between protein sequences, glycosites, types of glycosylation, and local protein fold / domain and other PTM context". GlycoDomainViewer integrates experimental data as well as knowledge data sources presenting the most extensive collection of information to explore the possible effect of glycosylation on a protein.

Despite the availability of these and more visual tools, the majority of glycoscientists are still using general purpose applications like Excel to publish experimental results. Therefore, results are hardcoded in figures or text and data is stored in tables which populate the supplemental material section. Additionally, integrative tools like GlycoDomainViewer, although very useful, are usually developed taking into account the needs of a specific research group limiting the possibility of reaching out to the entire community.

5. WEB DEVELOPMENTS

The integration of multiple data source is critical to have a complete overview of a living system. Then, once data are correctly combined, it is essential to develop software applications which allow the exploration of the integrated data sources and the design new hypotheses. Thus, besides data integration, software development represents another critical step towards a in-silico simulation of a living cell. Software development has many other aspects to be taken into account. One central question that comes along with the start of a new project is usually “Should I develop a desktop tool or a web application ?”. The answer to this question is crucial as the development process can be entirely different. The two main advantages of desktop applications are security and performance. Data are manipulated within the computer which can be detached from the network. Since most of the security issues come from the internet, desktop applications are more secure than web tools. Also, the use of machine-optimised code and the locality of the data increase the performance of desktop applications.

In the last years, the situation has changed and promoted the development of web applications over the desktop ones. Security issues are still there, but the spreading of technologies like HTTPS and SSL have provided a new security shield against fraudulent actions. On the performance side, the growing availability of cloud computing services, at a reduced cost, has given web applications a virtually endless computing power. For example, software like Hadoop, built for handling Big Data, can be plugged in a SOA and piloted via a web interface. In this scenario, the user-friendliness and possibility of updating contents remotely, introduced by web apps, are critical elements for developing a community-wide application that can be used by anyone even the less tech-savvy user. Another advantage of web tools is the possibility of using them regardless of the location (they only need an internet connection) and the platform which can range from a smartphone to a personal computer.

Web development is a broader term which implies the construction of a website. From the start of the Web in 1991 (History of the World Wide Web - Wikipedia, n.d.) websites have evolved from a collection of basic HTML pages to sophisticated applications built on top of multiple frameworks. A framework is an high-level solution which allows the reuse of software pieces. It also ensures the quality of the code and guarantees the same behaviours across multiple browsers and multiple platforms. The use of web framework has speeded up the development of new websites which, nowadays, fulfil high-quality standards. There are plenty of web frameworks on the market, therefore, choosing the right one has become a hard task. Web developers spend more and more time benchmarking different solutions since making a wrong choice would lead to a waste of energy and time. Each framework requires a period of study to learn its specific syntax and logic. Furthermore, some frameworks have short life leaving the developer without support and updates. In this ecosystem of frameworks which expands every day, we are going to focus on web components being one of the principal technologies used for developing the web tools.

Web Components. To fully understand the power of web components we need to introduce the concepts of HTML and DOM. HTML stays for HyperText Markup Language and is the standard programming language for building web pages. Together with Cascading Style Sheets (CSS) and javascript form the cornerstone technologies behind the web (HTML - Wikipedia, n.d.). The DOM, which stays for Document Object Model, is an application programming interface (API) which converts HTML, XHTML and XML into a tree structure where each node is an object and represent a part of the web page.

Web components are reusable interface widgets which are built with web technologies. In other words, web components directives allow developers to create new custom, reusable HTML tag in web pages and web apps. All the custom components built following the web components directives operate across modern browsers, and work in synergy with any JavaScript library or framework that support HTML (Webcomponents.Org - Discuss & Share Web Components, n.d.). Web components are standardised by the W3C, the international standard organisation for the World Wide Web, and they have been added to the specification of HTML and DOM.

Web components consist of four different technologies :

• Custom elements : A collection of JavaScript APIs that allow the user to define new custom HTML tags and their behaviour.

• Shadow DOM : A collection of JavaScript APIs to attach a private and encapsulated DOM to a custom element. Using shadow dom the code of each component become private and cannot be accessed from other ones. This technology removes the possibility of collision between different parts of the document.

• HTML templates : It introduce the custom tags 'template' and 'slot' used to create templates that are not displayed in the page. They can be adopted to create the custom element's basic structure.

• HTML Imports : It is the mechanism to import new custom components into plain HTML pages. HTML imports allow developers to encapsulate the logic and the interface of a web component into a separated file and import it where it will be used.

Developers are free to create "vanilla" web components which are built only using the standard by W3C. A second option is to use libraries which guide the development process and refine parts of the standard that are not mature. Google Polymer, Bosonic, SkateJS, X-Tag are some of the available libraries which help to create a web component ecosystem. Within the Glycomic@ExPASy initiative, we Google Polymer wa selected as one of the most stable and continuously updated libraries on the market. Also, Google Polymer developers are working together with W3C to enlarge and improve the web component standard. This lead to a reduction of the Polymer library each time the standard upgrade one of its technology. In the long run, Polymer components will be entirely built with the standard, and the Polymer library will slowly disappear.

6. Overviews : Tools & Databases

With the aim of facilitating glycoscience research, we have identified the tools and databases that are freely available on the internet and are regularly updated and improved. They have been clustered according to the major fields of applications : : **Portals ; Genome and Glycome ; Representations ; Experimental Results ; Glycan ; Glycoproteomics ; Functional Glycomics ; Glycolipids ; CAZYmes & Polysaccharides, Glycolipids**.

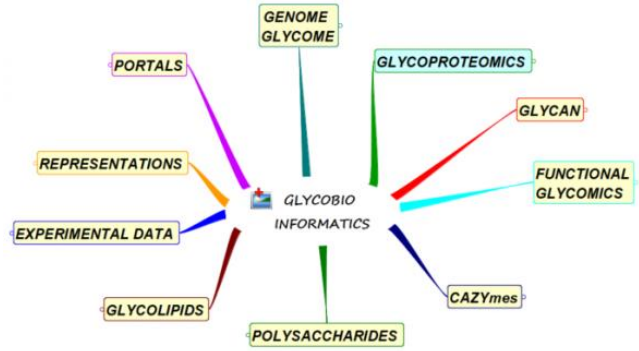


Figure 7: Schematic representations of the areas of research in glycoscience, where tools and databases are available

7. PORTALS :

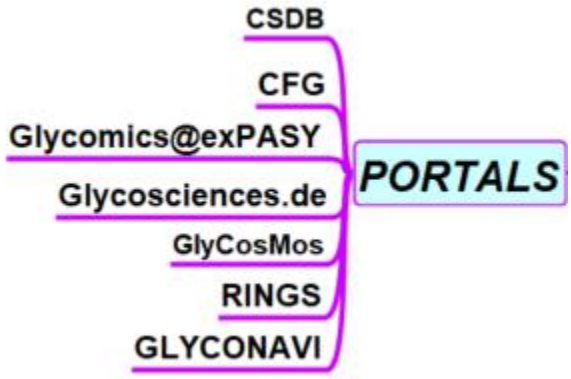


Figure 8: Some of the important portals in glycoscience**.**

Carbohydrate Structure Data Base CSDB covers information on structures and taxonomy of natural carbohydrates published in the literature and mostly resolved by nuclear magnetic resonance (NMR). CSDB is composed of two parts : Bacterial & Archeal (BCSDB) and Plant & Fungal (PFCSDB).
http://csdb.glycoscience.ru/database/core/help.php?topic=rules

Consortium for Functional Glycomics- CFG. The CFG serves to combine the expertise and glycomics resources to reveal functions of glycans and glycan-binding proteins (GBPs) that impact human health and disease. The CFG offers resources to the community, including glycan array screening services, a reagent bank, and access to a large glycomics database and data analysis tools.
http://www.functionalglycomics.org

ExPASy : Glycomics@ExPASy is the glycomics tab of ExPASy, the server of SIB Swiss Institute of Bioinformatics. It centralizes web-based glycoinformatics resources developed within an international network of glycoscientists.
https://www.expasy.org/glycomics

GlyCosMos : Portal development entails the integration of more diverse –omics data including glycogenes, glycoconjugates such as glycoproteins and glycolipids, molecular structures, and pathways.
https://glycosmos.org

GLYCONAVI : A web site for carbohydrate research. It consists of the 'GlycoNAVI DataBase' for molecular information of carbohydrates, and chemical reactions of carbohydrate synthesis, the *'Route Searching System for Glycan Synthesis'*, and 'GlycoNAVI Tools' for editing two-dimensional molecular structure of carbohydrates.
http://www glyconavi.org/

Glycoscience.de : The portal provides databases and tools to support glycobiology and glycomics research. Its main focus is on 3D structures, including 3D structure models as well as references to PDB entries that feature carbohydrates.
http://www.glycosciences.de

RINGS: A web resource providing algorithmic and data mining tools to aid glycobiology research.
http://rings.t.soka.ac.jp/

8. GENOMES-GLYCOMES :

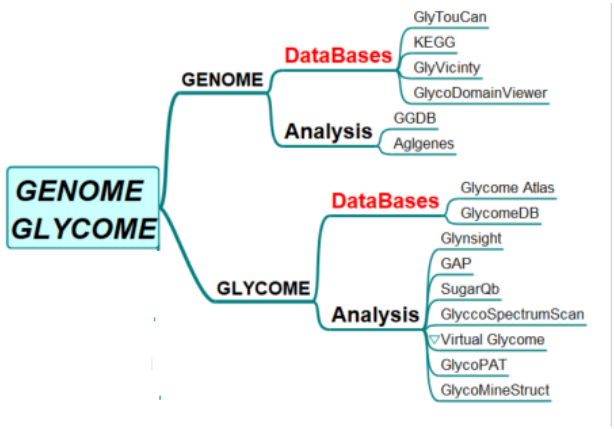


Figure 9: Databases and analytical tools for glycol genomics and glycoporteomics.
GlycoDomainViewer: a bioinformatics tool for contextual exploration of glycoproteomes. It presents a first map of the human O-glycoproteome with almost 3000 glycosites.
https://glycodomain.glycomics.ku.dk/

Glycome Analytics Platform:
https://bitbucket.org/scientificomputing/glycome-analytics-platform

GlycomeAtlasV5 : Visualization of glycome profiling data on human and mouse tissue samples.
http://rings.t.soka.ac.jp/GlycomeAtlasV5/index.html

GlycoGeneDatBase : Glycogene includes genes associated with glycan synthesis such as glycosyltransferase, sugar nucleotide synthases, sugar-nucleotide transporters, sulfotransferases, etc
https://acgg.asia/ggdb2

GlycoMinestruct : A bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features.
http://glycomine.erc.monash.edu/Lab/GlycoMine_Struct/

Glynsight : offers visualization and interactive comparison of glycan expression profiles. The tool was initially developed with a focus on IgG N-glycan profiles but it was extended to usage with any experiment, which produces N- or O-linked glycan expression data.
https://glycoproteome.expasy.org/glynsight/

GlycoSpectrumScan: A web-based bioinformatic tool designed to link glycomics and proteomics analyses for the characterization of glycopeptides. GlycoSpectrumScan is MS-platform independent, freely accessible, and profiles glycopeptide MS data using beforehand separately acquired released glycan and proteomics information. Both N- and O-glycosylated peptides as well as multiply glycosylated peptides can be analyzed
https://github.com/wliu1197/glycospectrumscan

GlyTouCan : GlyTouCan is the international glycan structure repository.
https://glytoucan.org/

GlyVicinty : Analysis of Amino Acids in the Vicinity of Carbohydrate Residues contained in the Protein Data Bank (PDB).
http://www.glycosciences.de/tools/glyvicinity/

KEGG: GLyCan database KEGG (Kyoto Encyclopedia of Genes and Genomes) is a bioinformatics resource for linking genomes to life and the environment.
https://www.genome.jp/kegg/glycan/

SugarQb: enables genome-wide insights into protein glycosylation and glycan modifications in complex biological systems. This is a collection of software tools (Nodes) which enable the automated identification of intact glycopeptides from HCD-MS/MS data sets, using commonly use peptide-centric MS/MS search engines.
http://www.imba.oeaw.ac.at/SugarQb SugarQb

Virtual Glycome: This website is focused on presenting selected computational tools and experimental resources that can be used to better understand the processes regulating cellular glycosylation at multiple levels.
https://virtualglycome.org/

9. REPRESENTATIONS :

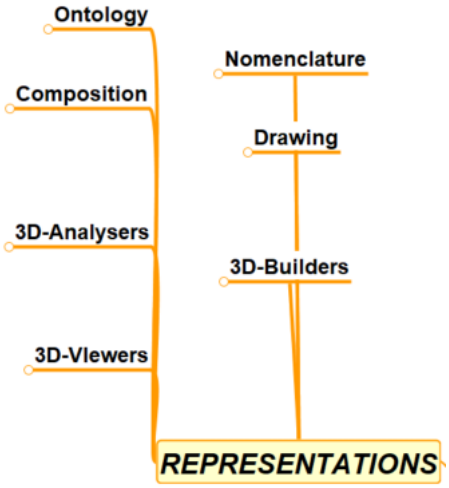


Figure 10: Informatics tools for complex carbohydrate represantions

Nomenclature

IUPAC Carbohydrate Nomenclature
www.sbcs.qmul.ac.uk/iupac/2carb/
SNFG Symbol Nomenclature for Glycans
https://www.ncbi.nlm.nih.gov/glycans/snfg.html,

GlycoCT:- A unifying sequence format for carbohydrates.

WURCS : the Web3 Unique Representation of Carbohydrate Structures.

KCF : Format used in the KEGG database.

LINUCS : Linear Notation for Unique Description of Carbohydrate Structures
http://www.glycosciences.de/modeling/sweet-remote/

Glyde II GLYDE-II : The GLYcan Data Exchange format PubChem Sketcher From drawn structures generates

PubChem Sketcher: From drawn structures generates SMILES, InChI and InChIKey
https://pubchem.ncbi.nlm.nih.gov/edit2/

Conversion : Utilities

GGlycoCT = WURCS
Convert Tool converts from any format (GlycoCT condensed,
https://glytoucan.org/Structures/structureSearch

KCF, IUPAC, LinearCode, LINUCS) to (almost) any other format (depending on input)
http://rings.t.soka.ac.jp/cgi-bin/tools/utilities/convert/index.pl

GlycoCTcondensed to KCF : converts a glycan structure in GlycoCT format to KCF
http://rings.t.soka.ac.jp/cgi-bin/tools/utilities/GlycoCTtoKCF_au/glycoct_index_au.pl

GLYDE2 to KCF : converts a glycan structure in GLYDE2 to KCF
http://rings.t.soka.ac.jp/cgi-bin/tools/utiliies/GLYDE2toKCF/glyde2_index.pl

IUPAC to KCF : converts a glycan structure in IUPAC format to KCF
http://rings.t.soka.ac.jp/cgi-bin/tools/utilities/IUPACtoKCF_au/iupactokcf_index_au.pl

LINUCS to KCF converts a glycan structure in LINUCS format to KCF
http://rings.t.soka.ac.jp/cgi-bin/tools/utilities/LINUCStoKCF/ linucs_to_kcf_index.pl

CSDB Linear to GlycoCT, Glyde-II, LinUCS, WURCS, GLYCAM, SMILES, 3DMOL
http://csdb.glycoscience.ru/database/core/translate.html

CSDB Linear to SweetDB, SNFG, structural formula image
http://csdb.glycoscience.ru/database/core/check_structures.html

Composition

GlycanMass GlycanMass is a tool which allows calculating the mass of an oligosaccharide structure
https://web.expasy.org/glycanmass/

SwissMassAbaccus Swiss Mass Abacus is a calculator of peptide and glycopeptide masses.
https://glycoproteome.expasy.org/swiss-mass-abacus/

Drawing

GlycanSketcher SugarSketcher : Quick and Intuitive Online Glycan Drawing
https://github.com/alodavide/sugarSketcher,

Glycan Builder An interface for building and displaying glycan structures
http://sugarbind.expasy.org/builder

Draw Glycan SNFG The symbol nomenclature for glycans (SNFG) contains 67 different monosaccharides represented using various colors and geometric shapes.
http://www.virtualglycome.org/DrawGlycan/

DrawRingS 5-based glycan structure drawing tool for generating KCF and IUPAC format and/or querying the RINGS database, ...
http://www.rings.t.soka.ac.jp/DrawRINGS: 2D HTML

LiGraph : generates schematic drawings of oligosaccharides which are often used to display glycan structure.

3D Analysers

Carp generates Ramachandran-like plots of carbohydrate linkage torsions in pdb-files.
http://www.glycosciences.de/tools/carp/

Azahar a PYMOL plugin for construction, visualization and analysis

pdb-care checks carbohydrate residues in pdb-files for errors.
www.glycosciences.de/tools/pdb-care/

pdb2linucs: automatically extracts carbohydrate information from pdb-files and displays it using the LINUCS-Code
www.glycosciences.de/tools/pdb2linucs/

3D-Builders

doGlyca:ns: Tools for Preparing Carbohydrate Structures for Atomistic Simulations of Glycoproteins, Glycolipids, and Carbohydrate Polymers for GROMACS.

Glycan-Web; Carbohydrate, Glycoprotein, GAGs Builder

Sweet: A program for constructing 3D models of saccharides from their sequences using standard nomenclature

PolysGlycanBuilder: Algal, Bacterial, GAG, Plant Polysaccharides, N-O linked Glycan

CHARMM Gui An option within the Glycan Reader & Modeler section of the general menu
http://www.charmm-ui.org/?doc=input/glycan

Carbuilder:
https://people.cs.uct.ac.za/~mkuttel/CBresidues.html

Rosetta Carbohydrate:
https://www.rosettacommons.org/docs/latest/application_docum entation/carbohydrates/WorkingWithGlycans

Others Several Molecular Modeling software have provision to build 3D structures of glycan

3D-Viewers

LiteMol Powerful and blazing-fast tools for handling 3D macromolecular data in the browser
https://webchemdev.ncbr.muni.cz/LiteMol/

Glycan-Web 3D-Symbol Nomenclature for Glycans
http://glycam.org/

SweetUnityMol : Molecular 3D structures (SNFG compliant) and networks viewer . Virtual Reality. Manual
https://glycopedia.eu/resources/sweet-unity-mol-3d-visualization/article/presentation

Others Several Molecular Viewer platforms have provision to display 3D structures of glycan

10. EXPERIMENTAL DATA

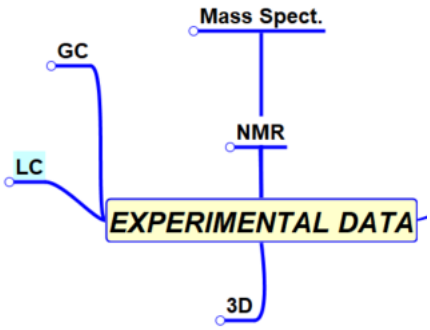


Figure 11: Tools for the structural analysis of complex carbohydrates

A Practical Guide to Structural Analysis of Carbohydrates
http://www.stenutz.eu

Mass Spectrometry

CCRC Spectral Database for PMAA's
https://www.ccrc.uga.edu/specdb/ms/pmaa/pframe.html Database of Partially Methylated Alditol Acetate

GMDB https://jcggdb.jp/rcmg/glycodb/Ms_ResultSearch GMDB is a database of glycan mass spectral data.

GlycoMob: An ion mobility-mass spectrometry collision cross section database for glycomics.
http://www.glycomob.org

NIST Mass Spectrometry Data Center The NIST Glycan Mass Spectral Reference Library is a collection of tandem mass spectral data (MS2, MS3, MS4) of glycans extracted from the NIST 17 Tandem MS library.
https://chemdata.nist.gov/glycan/spectra

UniCarb-DB UniCarb-DB is a structural and mass spectrometric database used in glycomics that provides over 1000 LC-MS/MS spectra for N- and O-linked glycans.
http://unicarb-db.expasy.org/

Gaz Chromatography

Glycopedia e-chapter: Mini database of MS Spectra of 4 types of derivatives (Acetylated alditols, acetylated methyl glycosides, partially methylated and acetylated alditols, acetylated octyl or butyl glycosides.
www ://http glycopedia.eu

Liquid Chromatography

GlycanR: A tool for analysing N-GlycanData which is mostly oriented to data obtained by UPLC (Ultra Performance Liquid Chromatography) and LCMS (Liquid chromatography–mass spectrometry) analysis of Plasma and IgG glycome.
project.org/web/packages/glycanr/index.html

Glycostore: A curated chromatographic, electrophoretic and mass-spectrometry composition database of N-, O-, glycosphingolipid (GSL) glycans
https://glycostore.org/

NMR

CASPER CASPER is a web-based tool, facilitates prediction 1H and 13C NMR chemical shifts of oligo- or polysaccharide .
www.casper.organ.su.se/casper/

GlycoQuest: The integrated search engine for glycans https://www.bruker.com/products/mass-spectrometry-and-separations/ms-software/proteinscape/glycoquest.html

GODESS: NMR spectrum simulation service for carbohydrate-containing molecules (including polymers and glycoconjugates).
http://csdb.glycoscience.ru/database/core/nmrsim.html

GRASS: provides semi-automated NMR-based structure elucidation of saccharides.
http://csdb.glycoscience.ru/biopsel/genstruc.php

3D Structures

GFDB: A glycan fragment database : a database of PDB-based glycan 3D structures.
http://www.glycanstructure.org

GLYCAM: Primary sequences for a number of common glycans have been pre-built and their predicted 3D structures are available. High Mannose, Hybrid N-Glycan, Complex Type, Sialyl/Fucose Complexes.
http://glycam.org

GlycoMapsDB: A database containing more than 2500 calculated conformational maps for a variety of di- to pentasaccharide fragment.
http://www.glycosciences.de/modeling/glycomapsdb/

Glyco3D: A single entry to access 3D structures of mono, di, and oligosaccharides, polysaccharides, lectins, gag-interacting proteins.
http://glyco3d.cermav.cnrs.fr

CHARMM GUI :
http://charmm-gui.org/input/glycan

11. GLYCOPROTEOMICS

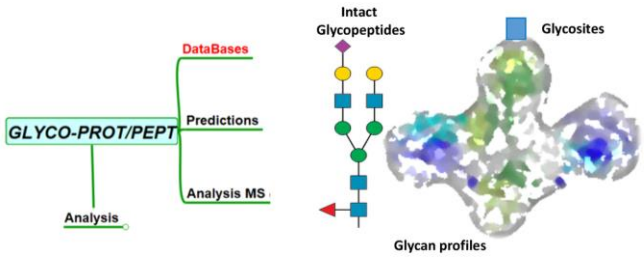


Figure 12: Tools for the structural analysis of complex carbohydrates

Glyconnect: A platform integrating sources of information to help characterise the molecular components of protein glycosylation.
https://glyconnect.expasy.org/

GlyDB: N-Glycan Structure Annotation of Gcoptides Using a Linearized Glycan Structure Database

GlycoFish: A Database of Zebrafish N-linked Glycoproteins Identified Using SPEG Method Coupled with LC/MS.
http://betenbaugh.jhu.edu/GlycoFish

GlycoFly Drosoph : A database for Drosophil N-linked glycoproteins identified using SPEG–MS techniques.
http://betenbaugh.jhu.edu/GlycoFly

GlycoPAT http://www.

GlyProt: A web-based tool that enables meningful N-glycan conformations to be attached to all the spatially accessible potential N-glycosylation sites of a known three-dimensional (3D) protein structure.

GlycoSuiteDB: A relational database that curates information from the scientific literature on glyco-protein derived glycan structures.
http://www.glycosuite.com

O-GlycBase A database of glycoproteins with O-linked glycosylation sites.
http://www.cbs.dtu.dk/databases/OGLYCBASE/

UniCarbKB: New database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations.
http://unicarbkb.org UniCarbKB :

UniPep: A database for human N-linked glycosites : a resource for biomarker discovery.
http://www.unipep.org

Predictions

N-Glycosylation

I-GPA: Integrated GlycoProteome Analyzer (I-GPA) for Automated Identification and Quantitation of Site-Specific N-Glycosylation.

MAGIC: Identifies intact N-glycosylated peptides from a public protein database without requiring any prior information of proteins or glycans.
http://ms.iis.sinica.edu.tw/MAGIC-web/index.html

NetNGlyc The **NetNglyc** server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons?
http://www.cbs.dtu.dk/services/NtNGlyc/

GlycoProtO-Glycosylation

DictyOGlyc The **DictyOGly**c server produces neural network predictions for GlcNAc. O-glycosylation sites in Dictyostelium discoideum proteins.
http://www.cbs.dtu.dk/services/DictyOGlyc/

ISOGlyP isoglyp.utep.edu/ Provides isoform specific O-glycosylation prediction

NetOGlyc The NetOglyc server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins.
www.cbs.dtu.dk/services/NetOGlyc/

O-GlcNAc Pred A sensitive predictor to capture protein O-GlcNAcylation sites.
tthp://121.42.167.206/OGlcPred/

YinOYang The YinOYang WWW server produces neural network predictions for O-ß-GlcNAc attachment sites in eukaryotic protein sequences**.**
www.cbs.dtu.dk/services/YinOYang/

N-O-Glycosylation

GlycoPP GlycoPP is a webserver for predicting potential N- and O-glycosites in prokaryotic protein sequence(s).
http://crdd.osdd.net/raghava/glycopp/

GPP Algorithmically prediction of N-linked and O-linked glycosylation.
http://comp.chem.nottingham.ac.uk/glyco/

GlySeq http://www.glycosciences.de/tools/glyseq/**Uses** the PDB and SwissProt to perform statistical analysis of glycosylation site

C-Mannosylation

NetCGlyc Prediction of mammalian C-mannosylation sites
http://www.cbs.dtu.dk/services/NetCGhtlyc/

Lysine Glycation

Glypre In Silico Prediction of Protein Glycation Sites by Fusing Multiple Features and Support Vector Machine.
http://www.cds.dtu.dk/databases/GlycateBase-1.0/

Analysis

GlycCompSoft Software for Automated Comparison of Low Molecular Weight Heparins Using Top-Down LC/MS Data?
http://www.heparin.rpi.edu

SweetSeQer Simple de novo filtering and annotation of glycoconjugate mass spectra
http://software.steenlab.org

Analysis MS

GlycoDENovo GlycoDeNovo provides an interpretation-graph which designates how to interpret each peak until the candidate topologies of precursor ion.
https://github.com/hongpengyu/GlycoDeNovo.

GlycoFragwork A computational framework for identification of intact glycopeptides in complex samples.
http://darwin.informatics.indiana.edu/col/

GP Finder available from :
cblebrilla@ucdavis.edu

GlycoMiner A software tool to automatically identify tandem (MS/MS) spectra obtained in liquid chromatography/mass spectrometry.
http://www.szki.ttk.mta.hu/ms/glycominer/

GlcoMID A graph-based spectral alignment algorithm that can identify glycopeptides with multiple hydroxylysine O-glycosylation sites by tandem mass spectra.
http://proteomics.informatics.iupui.edu/software/glycomid/

GlycoPepDB http://hexose.chem.ku.edu/sugar.php A Tool for Assigning Mass Spectrometry Data of N-Linked Glycopeptides on the Basis of Their Electron Transfer Dissociation Spectra

GlycoPEP Detector A Tool for Assigning Mass Spectrometry Data of N-Linked Glycopeptides on the Basis of Their Electron Transfer Dissociation Spectra.
http://glycopro.chem.ku.edu/ZZKHome.php.

GlycoPEP Evaluator esearch. **GlycoPep** Evaluator, generates decoy glycopeptides de novo and enables accurate false discovery rate analysis for small data sets.
https://desairegroup.ku.edu/r

GlycoPepID: http://www.hexose.chem.ku.edu/sugar.php

GlycoPep Mass: An application that permits to compute theoretical glycopeptide masses while creating inclusion lists for targeted data acquisition on.

GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides.
https://www.biomarkercenter.org/gpquest

GPS: To explore site-specific N-glycosylation microheterogeneity of haptoglobin using glycopeptide CID tandem mass spectra and glycan database search.
http://edwardslab.bmcb.georgetown.edu/software/GlycoPeptideSearch.html

GlycoSeq GlycoSeq uses a heuristic iterated glycan sequencing algorithm that incorporates prior knowledge of the N-linked glycan synthetic pathway to achieve rapid glycan sequencing
https://github.com/dbaileychess/

GlycoX: To Determine Simultaneously the Glycosylation Sites and Oligosaccharide heterogeneity of Glycoproteins.

pGLCO: A software tool designed for the analysis of intact glycopeptides by using mass spectrometry.
http://pfind.ict.ac.cn/software/pGlyco1505/

Peptoonist; uses tandem mass spectrometry (MS/MS) to detect glycosylated peptides and single-MS to find the N-glycans present on each of these peptides.

PepSweetener: EXPASY A Web-Based tool to support manual annotation of intact glycopeptide MS spectra

ProSIGHT: ProSight Lite : graphical software to analyze top-down mass spectrometry data
http://www.prosightlite.northwestern.edu

SimGlycan: A predictive carbohydrate analysis tool for MS/MS data.
http://www.premierbiosoft.com/glycan

12. GLYCANS

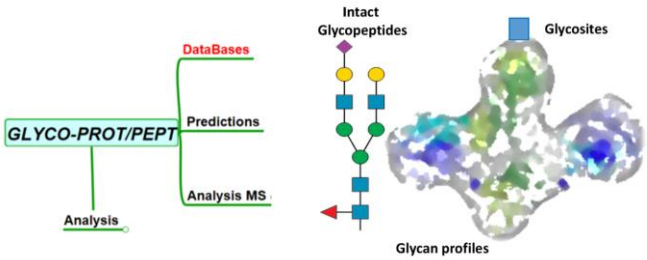


Figure 13: Tools and Databases for the structural analysis of glycans.

Analysis MS

BioOligo: As part of the suite of interlinked Databases of 3D Structures of Glycan, BioOligo. contains representations, 3D structures and NMR spectra of most occurring glycans.
http://glyco3d.cermav.cnrs.fr/search.php?type=bioligo

GlycoStore : Provides a centralised resource that combines glycan structure information with chromatographic separation and electrophoretic data.
http://www.glycostore.org

KEGG: Glycan: The KEGG GLYCAN structure database is a collection of experimentally determined glycan structures. It contains all unique structures taken from CarbBank, structures entered from recent publications, and structures present in KEGG pathways
https://www.genome.jp/kegg/glycan/

UniCarb-DB: UniCarb-DB is a structural and mass spectrometric database used in glycomics. UniCarb-DB provides over 1000 LC-MS/MS spectra for N- and O-linked glycans released from glycoproteins that were manually annotated.
http://unicarb-db.expasy.org/

Search

GlyS3: GlyS3 matches any substructure such as glycan determinants to a large collection of structures recorded in GlyConnect and SugarBindDB.
https://glycoproteome.expasy.org/substructuresearch/

Predictions

GS Align: A computational method for glycan structure alignment and similarity measurement. GS-align generates possible alignments between two glycan structures
http://www.glycanstructure.org/gsalign

GlycoFragment: A web tool to support the interpretation of mass spectraof complex carbohydrates.
http://www.glycosciences.de/tools/GlycoFragments/manual.pdf

GlycoForest: Glycoforest is a based on MS/MS spectra.
https://glycoforest.expasy.org/index.html

GNAT: Glycosylation Network Analysis Toolbox contains a variety of classes to describe glycans and glycosylation reaction networks. It also provides various methods to manipulate these classes
http://gnatmatlab.sourceforge.net/

GlyReSoft: GlyReSoft is a modular software tool for assigning site specific glycosylation from bottom-up mass spectrometry data sets.
https://github.com/GlycReSoft2

Analysis

Galaxy: Provides a useful method for an analytical procedure for N-glycan structures. Galaxy is a 2D/3D mapping method developed for the structural determination of asparagine-linked oligosaccharides (N-glycans) in glycoproteins.
http://www.glycoanalysis.info/galaxy2/manual/User_Manual.pdf

GIGTool GIG Tool is an application that extracts (i) precursor masses, (ii) oxonium ions and glycan fragments from tandem (liquid chromatography (LC)–MS/MS) mass spectra for glycan identification, and (iii) reporter ions from quaternary amine containing isobaric tag for glycan (QUANTITY) isobaric tags.
https://www.genome.jp/tools/kcam/

GLYCH: GLYCH (GLYcan CHaraterization) is package for glycan characterization using MS/MS.

GlycoMod: GlycoMod is a tool that can predict the possible oligosaccharide structures that occur on proteins from their experimentally determined masses
https://web.expasy.org/glycomod/

GlycoProfileAssigner: Automated structural assignment with error estimation for glycan LC data
http://glycananalyzer.neb.com

GlycoSearchMS: GlycoSearchMS takes a list of mass spectra peak values as input and searches for matches with the calculated fragments of SweetDB structures.
http://www.glycosciences.de/database/start.php?action=form_ms_search

GlycoWorkbench: A suite of software tools designed for the rapid drawing of glycan structures and for assisting the process of structure determination from mass spectrometry data.
https://glycoworkbench.software.informer.com/2.1/

GRITS: ToolBox GRITS Toolbox combines the analytical power of glycan permethylation with glycopeptide analysis, preserving site-specific information for comprehensive glycoprotein analysis
www.grits-toolbox.org/

KCaM: KEGG Carbohydrate Matcher s a tool for the analysis of carbohydrate sugar chains, or glycans. It consists of a web-based graphical user interface that allows users to enter glycans easily with the mouse
https://www.genome.jp/tools/kcam/

MultiGlycan: To help user to gather glycan profile information from LC-MS Spectra. It also reports quantity of specific glycan composition
https://bio.tools/MultiGlycan

13. Functional Glycomics

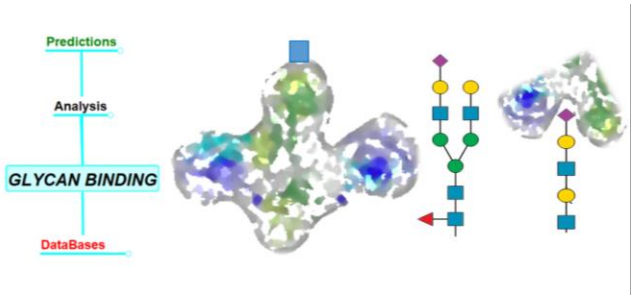


Figure 14: Tools and Databases for functional glycomics

Data Bases

Glyco-CD: Provides a collection of lectins and carbohydrates. Glyco-CD offers information on 63 clusters of differentiation (CD) antigens.
http://www.glycosciences.de/glyco-cd/

GlycoEpitope:ThisTHis database contains useful information on carbohydrate antigens, i.e. glyco-epitopes, and antibodies has been assembled as a compact encyclopedia.
https://www.glycoepitope

LectinDB: An integrated knowledge base (Lectindb, together with appropriate analytical tools
http://nscdb.bic.physics.iisc.ernet.in)

LfDB: Lectin Frontier DataBase (LfDB) provides quantitative interaction data in terms of the affinity constants (Ka) of a series of lectins toward a panel of pyridylaminated (PA) glycans
https://acgg.asia/lfdb2/

MCAW-DB: A database whereby users can view the multiple alignment analysis results obtained from the Multiple Carbohydrate Alignment with Weights (MCAW) tool.
https://mcawdb.glycoinfo.org/

PACDB Pathogen Adherence to Carbohydrate Database
https://jcggdb.jp/search/PACDB.cgi

SugarBindDB: provides information on known carbohydrate sequences to which pathogenic organisms (bacteria, toxins and viruses) specifically adhere.
https://sugarbind.expasy.org/SugarBind

UniLectin3D: A dedicated portal of databases and tools to study the lectins
https://unilectin.eu/

Predictions

PLecDom: A program for detection of Plant Lectin Domains in a polypeptide or EST sequence, followed by a classification of the identified domains into known families.
http://www.nipgr.res.in/plecdom.html

Analysis

Glydin: compiles and maps information relative to glycoepitopes (glycan determinants) as published in the literature or reported in databases.
https://glycoproteome.expasy.org/epitopes/

GlyQ-IQ: A software application for glycan MS based upon an algorithm centric nontargeted analyses approach.
https://github.com/PNNL-Comp-Mass-Spec/GlyQ-IQ

14. CAZYMES

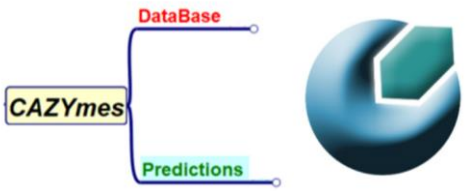


Figure 15: Databases and predictive tools for Carbohydrate Active Enzymes.

Data Bases

CAZy: A dedicated family classification system correlating structure and molecular mechanism of carbohydrate-active enzymes. It describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. The following classes of enzymes are covered : Glycoside Hydrolases (GHs) ; GlycosylTransferases (GTs) ; Polysaccharide Lyases (PLs) ; Carbohydrate Esterases (CEs) ; Auxiliary Activities (AAs) : Redox enzymes that act in conjunction with CAZymes & Carbohydrate-Binding Modules (CBMs)
http://www.cazy.org/

Predictions

GlycoDigest GlycoDigest a tool that simulates exoglycosidase digestion, based on controlled rules acquired from expert knowledge and experimental evidence available in GlycoBase.
www.glycodigest.org/

O-Glycologue O-Glycologue is a simulator of the enzymes of O-linked glycosylation.
https://www.boxer.tcd.ie/glycologue/

15. Polysaccharides

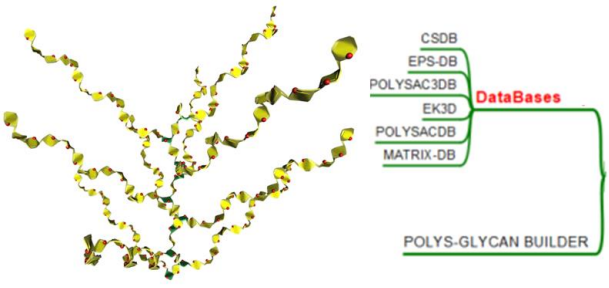


Figure 16: Polysaccharide Databases.

Carbohydrate Structure DataBase : CSDB merged from Bacterial (BCSDB) and Plant & Funi (PFCSDB) databases. Bacterial Carbohydrate Structure DataBase.
http://csdb.glycoscience.ru/database/index.html

EPS Database: The EPS Database provides access to detailed structural (1D-3D) taxonomic and bibliographic information on bacterial EPS
http://www.epsdatabase.com/

EK3D: An E. coli K antigen 3-dimensional structure database is a repository of 72 E. coli K antigens that provides information about the sugar composition, epimeric & enantiomeric forms and linkages between the sugar monomers.
http://www.iith.ac.in/EK3D/

MATRIX-DB: A biological database focused on molecular interactions between extracellular proteins and polysaccharides. It contains protein-protein interactions (PPIs) and also protein-glycosaminoglycan interactions, and 3D structures of GlycosAminoGlycans.
http://www.matrixdb.univ-lyon1.fr/

POLYSACDB: A comprehensive database of microbial polysaccharide antigens and their antibodies.
http://crdd.osdd.net/raghava/polysacdb/index.html

POLYSAC3DB: PolySac3DB is an annotated database that contains the 3D structural information and original fiber diffraction data of 157 polysaccharide entries that have been collected from an extensive screening of scientific literature.
http://www.polysac3db.cermav.cnrs.fr/

POLYS-GLYCAN BUILDER: An intuitive application to build 3D structures of polysaccharides (algea, bacteria, GAG, plants).
http://glyco3d.cermav.cnrs.fr/builder.php

16. Glycolipids

SphinGOMAP: SphinGOMAP is an evolving pathway map for sphingolipid biosynthesis that includes many of the known sphingolipids and glycosphingolipids arranged according to their biosynthetic origins.
http://sphingolab.biology.gatech.edu/

CHARMM Gui : simplifies the generation of various glycolipid structure and PSF files from a set of predefined glycolipids as well as a user-defined glycolipid sequence.
http://www.charmm-gui.org/?doc=input/glycolipid.

LIPID MAPS structure database (LMSD): A relational database encompassing structures and annotations of biologically relevant lipids. glycosphingolipids, saccharolipids.
http://www.lipidmaps.org/data/structure/

17. INTEGRATIVE TOOLS IN PRACTICE

The collection of tools and databases described in detail in the previous sections can be grouped into two categories. They can either be dedicated to solving a specific question or be used in an integrative way in several applications. In the context of exploring and understanding the biological functions where glycans are involved, toolboxes are required to navigate, investigate and correlate data. One such a facility is offered by the crosslinks of databases such as GlyS3, SugarBindDB, GlyConnect and the respective crosslinks to UniProt. A user can seek to establish the consistency of interactions taking place at the cell surface. In the following, three possible use cases are brought to the practitioner.

From MS to glycoprotein features.

This toolset is designed to match the expected boost of glycoproteomics (glycan composition at specific sites on complex mixtures of glycoproteins) data that is currently just reaching high throughput level. An example of how to integrate some of these dedicated tools for extracting glycoprotein features from MS data is shown below.

Figure 3

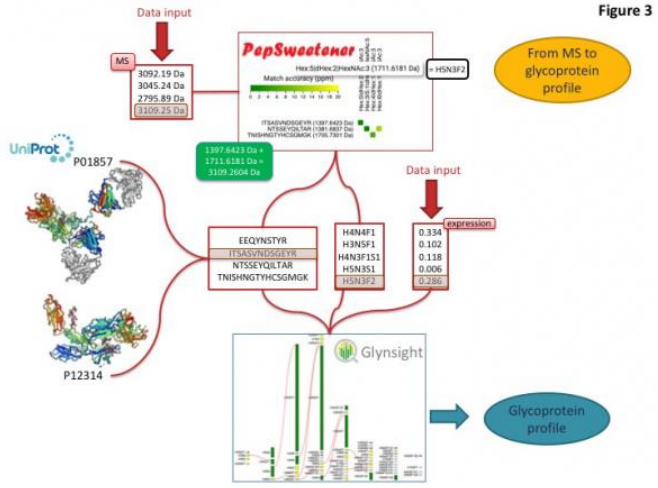


Figure 17: A typical scenario of the possible combination of PepSweetener and Glynsight to support the manual annotation of MS1 mass spectra of intact N-glycopeptides and integrate quantitative information when available. Users can process MS1 Spectra using PepSweetener to identify all the possible N-glycan compositions on a single human protein. Entire glycopeptide masses are broken into the respective contributions of the peptide and the glycan masses. Glynsight can be used to identify specific glycosylation patterns. The procedure can be repeated with a second protein and Glynsight will automatically generate the differential analysis of glycan profiles on the proteins. The integration with Glyconnect leads to displaying the potential glycan structures known to match the differentially expressed monosaccharide compositions.

From Mass Spectrometry data to glycoprotein profile.

A typical scenario of the possible combination of PepSweetener and Glynsight to support the manual annotation of MS1 mass spectra of intact N-glycopeptides and integrate quantitative information when available. Users can process MS1 Spectra using PepSweetener to identify all the possible N-glycan compositions on a single human protein. Entire glycopeptide masses are broken into the respective contributions of the peptide and the glycan masses. Glynsight can be used to identify specific glycosylation patterns. The procedure can be repeated with a second protein and Glynsight will automatically generate the differential analysis of glycan profiles on the proteins. The integration with Glyconnect leads to displaying the potential glycan structures known to match the differentially expressed monosaccharide compositions.

Predominant precursor masses in the MS spectra can be input into PepSweetener. This software supports the manual annotation of intact glycopeptides, using custom web visualization regardless of the instrument that produced the data. An interactive heat-map chart displays the results ; it features the combined mass contributions of theoretical (usually tryptic) peptides and attached glycan compositions. The variations in tile colours correspond to ppm deviations from the query precursor mass. Annotation can be refined through glycan composition filtering, sorting by mass and tolerance, and checking MS-MS data consistency via an in silico peptide fragmentation diagram (in-house fragmentation tool common with that of UniCarb-DB). PepSweetener is mainly designed as a complement or extension to software being developed for automatic analysis of glycoproteomics MS data and avoiding their dependency on a set workflow or type of instrument. The outcome of this study will guide the presentation of the Glycomics@ExPASy toolbox towards a more informative and instructive section on MS-based glycoproteomics data analysis tools.

Exploring glycoprotein features.

Current global glycome profiling experiments generate one or more set(s) of glycan compositions and structures with their respective expression on a protein, in a tissue or a cell. Tools and databases in Glycomics@ExPASy can be combined to explore distinctive glycan features that characterise glycoproteins as shown in figure xxx. In this case, the entry point of the workflow is GlyConnect to which a list of glycan compositions is submitted. The GlyConnect search tool will retrieve the possible related glycan structures and the proteins that have been reported to have these compositions/structures attached and stored in the databases.

A conceptual map displays the results ; the compositions sit in the middle and connect glycan structures and associated glycoproteins, respectively on the right and the left sides on the Figure 18.

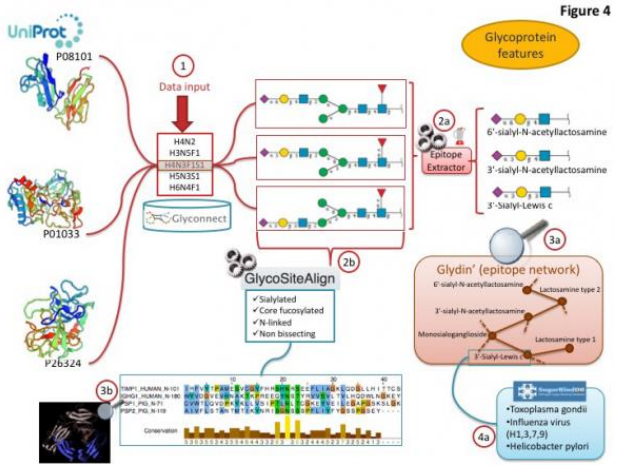


Figure 18: From composition to glycoprotein features:: An interactive way of extracting glycoprotein features from glycan compositions combining published data and ad hoc tools. A list of compositions is input in GlyConnect, which retrieves all the proteins reported as having these compositions attached to them (on the left) and reported glycan structures corresponding to this composition (on the right) annotated in this knowledge base. Glycan structures can be further processed to extract contained glycan epitopes using EpitopeXtractor. Glycoepitope results can be mapped on Glydin', an interactive epitope network. Glydin' aggregates glycan epitopes from four different sources (databases and literature reviews) and provides links to the original information. When epitopes are taken from SugarBindDB, further information on the pathogens can be browsed.

This visualization is well suited for understanding the potential relations between proteins and glycans. Activating the integrated EpitopeXtractor function provides a selection of glycan structures.

Glycan-mediated protein-protein interactions.

Using another combination of tools and databases in Glycomics@ExPASy, potential correlations between a glycan-binding protein (GBP) of a pathogen a host glycoprotein and a glycan structure can be made. In the first scenario, the starting point is a glycoepitope recognised by a specific GBP, a bacterial lectin described in SugarBindDB.

The blood group B antigen triose illustrates this point. A binding event in this database is always formed by A pair composed of a GBP/lectin and a glycoepitope part of a glycan present on the host surface defines a binding event in the database. Whenever possible, further information of a GBP/lectin is available via cross-reference to UniProt. The glycoepitope can be used as an input of the GlyS3 substructure search tool to match the full structures stored in GlyConnect that contain this specific ligand.

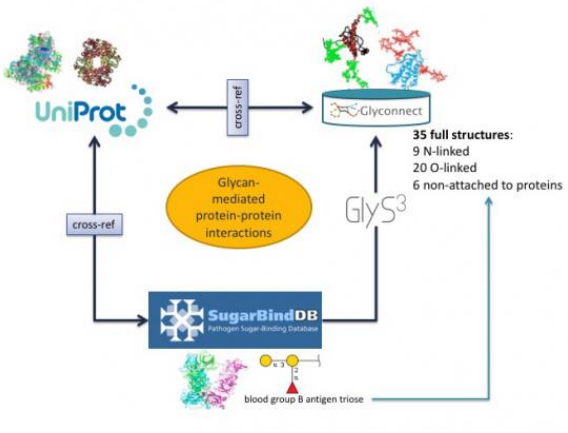


Figure 19. Glycan mediated protein-protein-interactions

The list of glycan structures retrieved by GlyS3 can be explored in GlyConnect that reports relationships between glycans and glycoproteins. The second scenario starts from a glycan structure in GlyConnect and relies on its reported relationships with glycoproteins.

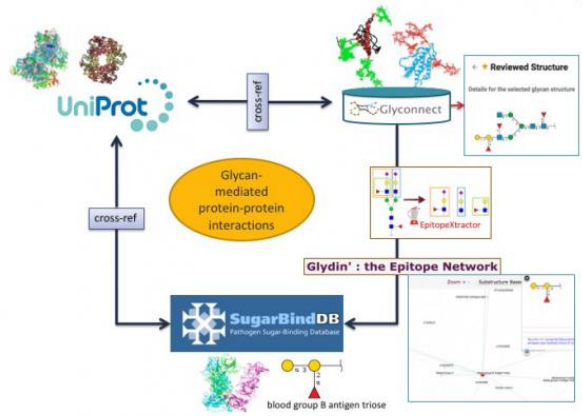


Figure 20: Glycan mediated protein-protein intearctions

Figure 19 shows how a new hypothesis on glycan mediated protein-protein interaction can be built using published data in Glyconnect and SugarBindDB : (a) In this scenario glycan binding protein (GBP) is selected in SugarBind. The information on the glycan ligand recognized by the GBP is used to perform a substructure search on all the structures in Glyconnect with the GlyS3 glycan substructure search tool. The structures identified by GlyS3 are used in Glyconnect to create a list of target proteins that can interact with the initial GBP. In the example of the blood group B antigen triose, there are 35 full structure types in GlyConnect that contain this glycoepitope (b) In this scenario a glycoprotein in GlyConnect is selected with its list of associated glycan structures. Glycans are processed with EpitopeXtractor to single out all the glycan epitopes contained. The Glydin' interactive map of structurally related glycoepitopes helps visualising the potential common substructures in the complete set of glycoepitopes. Then, extracted epitopes are used in SugarBind to identify all the reported GBPs that can possibly interact with the initial

glycoprotein. In this example, the VP1 capsid protein of the Norwalk virus is known to bind the blood group B antigen triose. Note that protein structures shown above UniProt and those shown above GlyConnect are not related to the example but simply illustrating the difference in the information that is stored on the unglycosylated protein in contrast with the stored information on intact glycoproteins

The figure shows an example of a reviewed N-linked glycan structure. GlyConnect also offers the option of running EpitopeXtractor to generate a selection of glycoepitopes contained in this starting glycan. Leveraging the binding data in SugarBindDB, the obtained glycoepitopes can be associated with a collection of GBPs/lectins that recognize one or more of these glycoepitopes. In the end, the workflow allows the selection of GBPs that could possibly interact with the glycoproteins on which the starting glycan has been reported to be attached. Cross-references of both glycoproteins in GlyConnect and GBPs/lectins in SugarBindDB to UniProt can be used to further rationalise potential interacting partners.

ANNEX I : INTEGRATIVE TOOLS IN PRACTICE

Data integration is one of the main challenges in bioinformatics. Although the term "data integration" often appears in research, especially in life science, a consolidated definition is still missing. We report here several definitions of "data integration" which will help in understanding the central concept and how it evolved.

In the beginning, the "data integration" problem was restricted to database research. Ziegler et al. describe the integration problem as the aim at "combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system" (Ziegler & Dittrich 2004). A higher level definition comes from Leser and Naumann which defines "data integration" as "a redundancy-free representation of information from a collection of data sources with overlapping contents" (Leser & Naumann 2007). More recently, Gomez-Cabrero et al. describe "data integration as the use of multiple sources of information (or data) to provide a better understanding of a system/situation/association/etc" (Gomez-Cabrero et al. 2014).

Out of these three options, the last definition is the most general and complete. The authors emphasized the use of several sources to discover new insights that are hidden when looking at a single source. We propose to adopt this definition of "data integration" and consider the two main challenges associated with it : 1) finding relevant data sources (data discovery) ; 2) using collected data to produce new findings (data exploitation) (Weidman and Arrison 2010). Additionally, we tackle the problem of data provenance, which arises once data sources are integrated and allow to trace the origin of each piece of information.

Data discovery

The data discovery challenge entails the search of relevant data sources for describing a system. Nowadays, finding data sources for a specific problem is easy. However, finding a relevant source for a system of interest is becoming more and more challenging. The ease of publishing data through the Web has contributed to an explosion of online resources and publicly accessible datasets. Furthermore, the picture becomes more and more fragmented as each sub-discipline provides its data representation. In biology, the problem is not only arising from aiming to combine different omics but also to reconcile information even within the same field. For example in glycomics, several different formats are available to describe glycan structures.

We list here the main three :

- IUPAC format which is regulated by the International Union of Pure and Applied Chemistry
- GlycoCT format which has been developed under the EUROCarbDB project.
- WURSC format which is developed in Japan and used for the structure repository GlyTouCan.

The creation of more and more heterogeneous data sources leads to what has been called "A loose federation of bio-nation" (Goble and Stevens 2008). In this context, the use of standards and community guidelines, which are going to be tackled in "Data integration in bioinformatics" paragraph, is the only way to stop the data fragmentation smoothing the way towards an efficient data discovery.

Data exploitation

Data exploitation is the process of exploring a set of data sources to provide new insights. Before starting the data exploitation, scientists must have a complete overview of each dataset including the units of measure and more subtle aspects such as environmental conditions, equipment calibrations, preprocessing algorithms, etc (Weidman and Arrison 2010). Having an in-depth knowledge of the data is the only way to avoid the phenomenon of "garbage-in-garbage-out" that could affect the data exploitation outcome.

Once each data set has been fully clarified, researchers can focus on developing methodologies to analyse the data. Methodologies can vary according to the different data types. In this regards, we distinguish between "similar" and "heterogeneous" types. According to Hamid et al. we consider "similar type" when they are produced by the same underlying source, for example, they are all gene expression data sets. If multiple data sources, like gene expression and protein quantification data sets, are taken into account, we refer to data as "heterogeneous type" (Hamid et al. 2009). Although researchers are developing more and more hybrid methodologies for data integration, a set of general techniques will be presented in the next paragraph.

The creation of data explorative tools is the last but not the least step of the data exploitation process. The development of user-friendly interfaces to navigate the outcome of data exploitation can decree the success of the whole process. Although scientist are usually putting most of their attention on methodologies and algorithms, the design of custom visualisation can be time-consuming especially due to data heterogeneity and data volume.

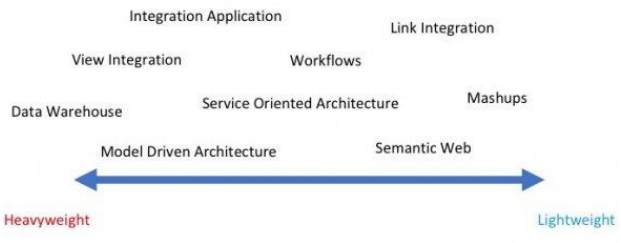
Data provenance

With the integration of more and more data sources, understanding the derivation of each piece of information becomes an issue. This is referred to as the data provenance issue in the literature. As in the case of data integration, data provenance has a different definition according to the field. In the context of database systems, Buneman et al. describe it as "the description of the origins of a piece of data and the process by which it arrived in a Database" (Buneman, Khanna, & Wang-Chiew 2001). However, data provenance is not only related to the data but also to the processes that led to the creation of the data. For this reason, Greenwood et al. (Greenwood et al. 2003) define data provenance as part of metadata which records "the process of biological experiments for e-Science, the purpose and results of experiments as well as annotations and notes about experiments by scientists".

We characterise data provenance as the record of the origin and all transformations that every dataset has undergone. It is quite a central question since the corresponding solutions enable users to trace back to the original data. Additionally, this concern for tracing information back provides the means to understand the different workflows that have been applied to integrate each dataset. For example, if a user is interested in removing all datasets that are produced from mouse, using the data provenance metadata, he/she can exclude these datasets and use the rest of information to perform new analyses (Masseroli et al. 2014). Data provenance is also essential in the assertion of the quality of a dataset, especially in fields where information is treated manually and the quality can vary according to the curator.

ANNEX II: DATA INTEGRATION STRATEGIES

In the last 20 years, software developers have explored a wide variety of methodologies for data integration. Each of this methodology has a technology of reference and uses one or more touch-points. A touch-point is the logical connection between different data sources which make integration possible, for example, data values, names, ontology terms, keywords, etc. In this section we list a few popular approaches that have been identified by Goble et al. These methods show a different level of integration which ranges from light solutions to heavyweight mechanisms.



Schema of the different data integration strategies placed according to their level of integration. From left to right we go from heavyweight to lightweight integration level.

Service oriented architecture

Service oriented architecture (SOA) is a way to integrate different data sources which can be accessed using a programmatic interface. The beauty of SOA resides in its loose coupling among different resources. The implementation of each resource is entirely masked behind its interface. As far as the interface remains the same, each resource provider can change and expand its resource without causing problems with the rest of the architecture. Usually, SOA relies on technologies like Simple Object Access Protocol (SOAP) and Representational State Transfer (REST) which allow the creation of web services (Dudhe & Sherekar 2014). Due to its simplicity, contrapose to the SOAP verbosity, REST emerged as a de facto standard for service design in Web 2.0 applications (Battle & Benson 2008).

Data sources can be accessed through programmatic interfaces which eradicate the problem of user simulation and screen-scraping. However, the poor stability of web services and the lack of documentation represent the major issues for SOA which leads to the impossibility of using the corresponding data.

Link integration

Link integration is a technique where an entry from a data source is directly cross-linked to another entry from a different data source. Because many data sources are websites and subsequently, entries are web pages, link integration can be renamed hyperlink integration. Users can navigate among different data sources using the hyperlinks present in each web page. The Uniprot "Cross-references" section, available in each entry, represents an example of link integration. In this section, a list of links connects the entry to sequence databases, 3D structure database, etc (Figure 4). Link integration is broadly used in bioinformatics and can be seen as a lite integration technique. This method relies on service provider agreement, and it is vulnerable to name clash, updates and ambiguous cases (Goble & Stevens 2008).

Cross-references

Sequence databases

Select the link destinations:	
EMBL	U39317 mRNA Translation: AAA91460.1
GenBank	L40146 Genomic DNA Translation: AAC41750.1
DDBJ	AY651263 mRNA Translation: AAX35690.1
	AF317220 mRNA Translation: AAK93958.1
	AK001311 mRNA Translation: BAG50891.1
	AK001428 mRNA Translation: BAG50911.1
	AC010378 Genomic DNA No translation available.
	CH471062 Genomic DNA Translation: EAW62095.1
	CH471062 Genomic DNA Translation: EAW62096.1
	CH471062 Genomic DNA Translation: EAW62097.1
	BC033349 mRNA Translation: AAH33349.1
CCDS	CCDS43369.1 [P62837-1]
	CCDS47275.1 [P62837-2]
PIR	I59365
RefSeq	NP_003330.1, NM_003339.2 [P62837-1]
	NP_862821.1, NM_181838.1 [P62837-2]
UniGene	Hs.108332

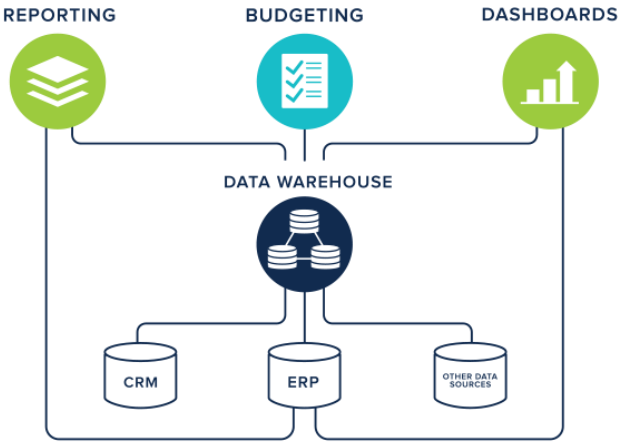
3D structure databases

Select the link destinations:	PDB entry	Method
PDBe	1UR6	NMR
RCSB PDB	1W4U	NMR
PDBj	2CLW	X-ray
	2ESK	X-ray
	2ESO	X-ray
	2ESP	X-ray
	2ESQ	X-ray
	3A33	X-ray
	3JVZ	X-ray
	3JW0	X-ray

The cross-reference section of a UniProt entry which represents an example of link integration.

Data warehousing

The data warehousing technique has its root in many companies. All the data produced within an organisation is extracted, transformed and loaded (ETL) into a new general data model (Figure 5). In this combined shape, data can be analysed to provide useful strategic information for business intelligence (Ponniah 2004). Data is stored and queried as a monolithic, integrated resource which relies on a predefined schema. In contrast to the previous methods, data warehousing represents a heavyweight mechanism for data integration. Due to the initial high costs of implementing a data warehouse and the fixed model, which can hardly change with time, this technique failed to last in life science applications.



Data warehouse model of BI360 by Solver https://www.solverglobal.com/it-it/products/data-warehouse This method is well-suited for companies which have the control over data production but becomes particularly unsafe when data is produced by third parties who potentially and unpredictably change their model at any time. When one or more data sources cannot be synchronized with the data warehouse, the only solution is to redesign the underlying data model from scratch, which is costly. A recent example of a data warehouse in bioinformatics is Geminivirus.org (Silva et al. 2017).

View integration

View integration is based on the same concept as data warehousing without providing a monolithic integrated resource. In this methodology, data is kept within the sources that are integrated on fly to provide a fresh integrated view. Users have the illusion of querying a unique resource, but, in the background, data is pulled from the several sources using ad-hoc drivers (Halevy 2001). The mediated schema of the view is defined at the beginning like in data warehousing, but drivers can be adapted to support changes in the data sources. However, drivers tend to grow with time making the maintenance more and more complicated. Additionally, the overall performance can be an issue since, in the view integration, the query speed is limited by the slowest source of information. TAMBIS (Stevens et al. 2000) can be considered the first example of view integration in bioinformatics. This software application was able to perform tasks using several data sources and analytical tools. TAMBIS used a model of knowledge built by a list of concepts and their relationships. However, the tool is not maintained anymore.

Model-driven service oriented architecture

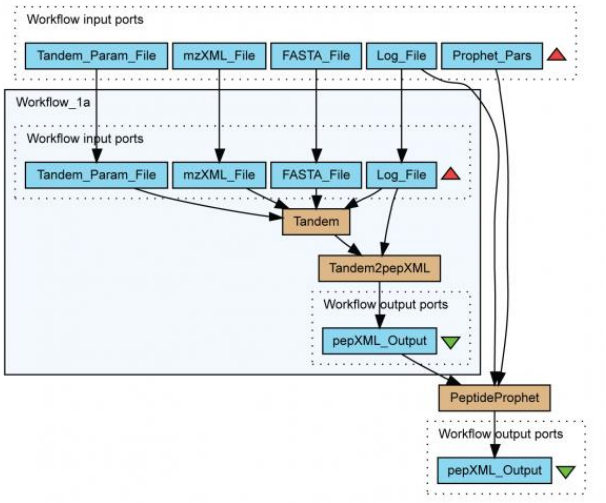
The model-driven service oriented architecture represents a hybrid technique which combines SOA and the integration view. Usually, this methodology is used by notable projects which are able to define a common data model for a particular context. In this way, data producers can join the infrastructure only if they are fully compliant with the predefined model. One example of model-driven SOA is caCORE, a software infrastructure which allows the creation of "interoperable biomedical information systems" (Komatsoulis et al. 2008). Using caCORE, scientists can access (syntactic interoperability) as well as understand (semantic interoperability) the data once it has been retrieved. caCORE has been used to create the cancer Biomedical Informatics Grid (caBIG) which consist of a data grid system where scientist can publish and exchange cancer-related data. In 2012, caBIG was followed by the National Cancer Informatics Program.

Integration applications

Integration applications are special tools designed to integrate data in a single application domain. Contrary to view integration and data warehousing, integration applications are far from being general integration systems. Software developers

tailor the application according to the needs of a specific subfield. In this way, the application is well suited to a specific application domain, but it cannot be transposed in another field. Due to their custom implementation, integration applications are usually a mix of several data integration methodologies. An excellent example of this technique is EnsEMBL (http://www.ensembl.org/) genome database project (Zerbino et al. 2018). EnsEMBL is a genome browser built on top of an open source framework which can store, analyse as well as visualise large genomes. It provides access to an automatic annotation of the human genome sequence which is based on several external data sources.

Workflows



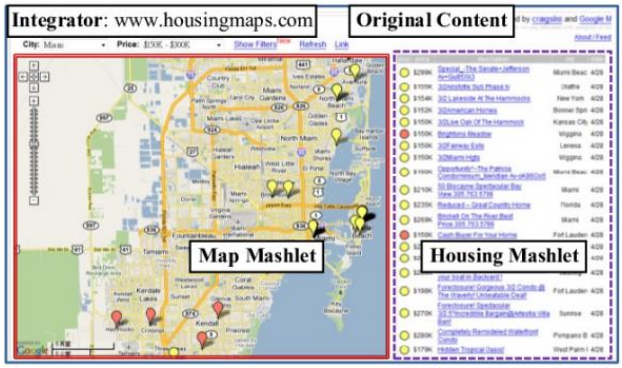
An example of Apache Taverna workflow for proteomics taken from http://ms-utils.org/Taverna/.

In particular, this workflow identifies peptides from mass spectrometry data using X !Tandem and validates assignments with PeptideProphet.

In data integration, a workflow describes the set of transformations which are applied to the data. A workflow can be built using a set of custom scripts or taking advantage of one of the workflow management systems available. When software like Knime (Fillbrunn et al. 2017), Apache Taverna (Oinn et al. 2004) or Galaxy (Afgan et al. 2016) are used, researchers can perform in silico experiments without becoming neither a software developer nor a scripting language expert (Karim et al. 2017).

Contrary to data warehouse and integration view, the integration process is defined by a series of transformations which are publicly exposed. The figure above describes a workflow for the identification of peptides in mass spectrometry data. Light blue boxes are file inputs whereas brown boxes are running scripts like PeptideProphet. Workflows can cope with unreliable data sources, and they can be adapted to face changes in data production. However, they are not the solution to every problem since the design of a workflow can be hard and its quality is strictly bounded to the data sources they integrate.

Mashups



An example of a mashup extracted from "A Secure Proxy-Based Cross-Domain Communication for Web Mashups" (Hsiao et al. 2011). The housing data feed is integrated with Google Maps to show the position of each entry.

All the methodologies proposed, except for link integration and workflows, require robust database and programming expertise. For some techniques, changing the data model or adding new data sources represent significant issues. For this reason, with the start of the Web 2.0, mashups have emerged. A Mashup is a web page or a web application where a collection of data sources are combined to create a new service. To create a mashup, we identify possible data sources that can be integrated to produce a novel functionality. Figure 7 shows a mashup done by integrating Craigslist apartment and housing listings (on the right) onto Google Maps.

Mashups provide a lite integration which is closed to aggregation more than integration. However, they produce useful lightweight applications which can be quickly built in a short amount of time with limited expertise. Mashup tools like Pipes (https://www.pipes.digital/) use a graphical workflow editor to bridge web resources easily. Data visualisation tools like Google Maps and Google Earth offer the possibility to display and combine georeferenced data on this principle as well.

The Semantic Web and RDF

The semantic web, even defined as the web of data, "provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries" (W3C Semantic Web Activity Homepage, n.d.). It is based on the Resource Description Framework (RDF), which consists of a standard model for data interchange on the web. The graph-based structure, provided by RDF, is responsive to change and support semantic descriptions of data. All data in RDF are in the form of triples, i.e., statements always composed by a subject, a predicate and an object. Each of these resources is identified by a Uniform Resource Identifier (URI) or a blank node. The latter identifies anonymous resources which can only be subjects or objects. Referring data with the same URI across several data sources, when these are semantically equal, allows

the creation of touch points. As mentioned earlier, touch points are the connection of multiple knowledge graphs and allow the creation of integrated resources. To store data in an RDF endpoint, it is necessary to draft an ontology which can be described using the RDF Schema (RDFS) or the Web Ontology Language (OWL). Additionally, ontologies are helpful to understand the graph structure of resources and to link resources which have shared information. Once data are stored in the triple store, SPARQL, a SQL-like query language, is used to retrieve data from each resource. Using federated queries, SPARQL interacts and retrieves data from multiple connected data sources. The logic is masked to the user who interacts only with one single interface.

To conclude, publishing data using RDF allows connecting different data sources that have one or multiple touch-points. The ultimate goal of the semantic web is to have a single web of connected endpoints which can be queried as a single resource. Cmapper, for example, is a web application that allows navigating all EBI RDF endpoint at once using the gene name as touch point (Shoaib, Ansari, & Ahn 2017).

ANNEX III : DATA INTEGRATION IN BIOINFORMATICS

All data integration techniques presented in the previous paragraphs need touch points to be implemented. In bioinformatics, a diversity of efforts have been carried out to provide standards and, as a matter of fact, touch points across data sources. In this paragraph, we identify some areas of study which are crucial to enforce standardisation and encourage data integration.

Standards

In life science, where data can be represented in many ways, widely adopted standards provide the only ground for data exchange and data integration. To show why standards are so relevant, we take the example of the amino acid naming convention. The 20 amino acids have a standard name that is recognised worldwide. Additionally, each amino acid has a one and three letter codes that are used by biologists around the globe. If a European biologist talks at a conference in Asia about the S amino acid, everyone in the audience understands it is about Serine.

Nowadays, lots of initiatives for developing standards are arising. We took a list of the most famous from a paper published by Lapatas et al. which is available in Table I (Lapatas et al. 2015). We will not explore each of these initiatives, but we provide URLs for more information.

To conclude, we want to strive the importance of standards for data sharing. Standards facilitate data re-use, limiting the work needed to integrate different data sources and the waste of potential datasets.

List of data standard initiatives. Courtesy of "Data integration in biological research: an overview" (Lapatas et al. 2015).

OBO The Open Biological and Biomedical Ontologies www.obofoundry.orgEstablish a set of principles for ontology development to create a suite of orthogonal interoperable reference ontologies in the biomedical domain PMID=17989687

CDISC Clinical data interchange standards consortium www.cdisc.orgEstablish standards to support the acquisition, exchange, submission and archive of clinical research data and metadata PMID=23833735

HUPO- PSI Human Proteome Organisation- Proteomics Standards Initiative www.psidev.info Defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification PMID=16901219

Alliance Global Alliance for Genomics and Health genomicsandhealth.org Create interoperable approaches to catalyze projects that will help unlock the great potential of genomic data PMID=24896853

COMBINE Computational Modeling in Biology co.mbine.orgCoordinate the development of the various community standards and formats for computational models PMID=25759811

MSI Metabolomics Standards Initiative msi-workgroups.sourceforge.net

Define community-agreed reporting standards, which provided a clear description of the biological system studied and all components of metabolomics studies PMID=17687353

RDA Research Data Alliance rd-alliance.org Builds the social and technical bridges that enable open sharing of data across multiple scientific disciplines

Ontologies

In the last twenty years, several ontologies have been created in the biological and biomedical fields (Bard & Rhee 2004 ; Hoehndorf, Schofield, & Gkoutos 2015 ; Kelso, Hoehndorf & Prüfer 2010). In philosophy, an ontology describes "what exists", whereas, in life science, it represents what exists in a specific context, for example, diseases (Turk 2006). **An ontology is defined as a collection of concepts and relationships used to characterise an area of concern.**

To consolidate and coordinate the rapid spread of ontologies, in 2001, Ashburner and Lewis established The Open Biomedical Ontology (OBO) consortium. The OBO ontologies form the basis of OBO Foundry, a collaborative experiment based on the voluntary acceptance by its participants of an evolving set of principles that extend those of the original OBO (Smith et al. 2007). As stated in its website, the OBO Foundry "*is a collective of ontology developers that are committed to collaboration and adherence to shared principles*", and its mission "*is to develop*

a family of interoperable ontologies that are both logically well-formed and scientifically accurate".OBO contains ten ontologies (June 2018) which are member ontologies like Gene Ontology (GO) and more than a hundred candidate ontologies. To become member, a candidate ontology has to be developed using OBO's shared principles and validated by OBO members (Quesada-Martínez et al. 2017).The growing use of OBO ontologies allows connecting more and more datasets using techniques like the semantic web. This enhances data integration and fosters the creation of the web of data.

Formats and reporting guidelines

As data is increasingly generated by high throughput techniques, computers, equipped with appropriate software, are required to store and analyse the information produced. In this scenario, data formats play a critical role as they provide instructions to store data in a file. However, the scarcity of well-designed data formats gives birth to many different standards that hamper data exchange and data integration. Therefore, bioinformaticians are forced to build converters, spending more time to clean data than to analyse them. Currently, to store Next Generation Sequence (NGS) data, there are six common file formats : FASTQ, FASTA, SAM/BAM, GFF/GTF, BED, and VCF (Zhang 2016). To ease and foster data sharing, converters between different data format have been developed in genomics, proteomics, etc. In this context, a good example is provided by the PRoteomics IDEntifications (PRIDE) project (Vizcaíno et al. 2016). PRIDE is a centralised repository for proteomics data which includes protein and peptide identification as well as post-translational modifications. Since mass spectrometry data can be encoded with several formats, the PRIDE development team have developed PRIDE Converter. This tool converts a large amount of mass spectrometry encoding formats into a valid PRIDE XML ready for submission to the PRIDE repository (Barsnes et al. 2009).With the gradual maturation of "omics", a huge step forward has been made by adopting clear guidelines to describe and depositing datasets. In this context, the first example of a concrete guideline is represented by The Minimum Information About a Microarray Experiment (MIAME) (Brazma et al. 2001). MIAME defines "the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified". The use of MIAME facilitates the creation of public repositories as well as the development of data analysis tools. Following the example of MIAME, in 2007, Taylor et al. published The minimum information about a proteomics experiment (MIAPE) (Taylor et al. 2007). Nowadays, proteomics guidelines are defined by HUPO Proteomics Standards Initiative (Hermjakob 2006) which additionally proposes data formats and controlled vocabularies (http://www.psidev.info).
In general, these guidelines focus on defining the content and the structure of necessary information to describe a specific experiment. Although they do not provide any technical solution for storing data, some of them suggest standard file formats.
To conclude, the use of minimum information guidelines together with suggested data formats enhance the data integration progress and the reusability of datasets.

Identifiers

An identifier is a short list of characters which identifies a data entry. For example, UniProt (Bateman et al. 2017) is using accession numbers, i.e. stable identifiers, to identify entries. When two or more entries are merged, all the accession numbers are kept. In this case, one is the "Primary (citable) accession number" whereas the others become "Secondary accession numbers". To avoid any source of uncertainty, it is not possible that one accession number refers to multiple proteins. In life science, the information is spread across multiple databases, and each of them has developed its identifiers. This leads to a multitude of identifiers to describe the same biological concept. However, to facilitate data integration, databases have cross-referenced their entries with external resources . In UniProt, for example, each protein has a cross-reference section which contains all external identifiers related to the same protein (3D structures, protein family, genome annotation, etc). If all life science databases used the same identifier to characterise a biological concept, data integration would be facilitated. However, the use of identifiers from established databases, like UniProt or GenBank, in research papers is already a sign of progress.In genomics, for example, most journals oblige researchers to deposit newly obtained DNA and amino acid sequences to a public sequence repository (DDBJ/ENA/Genbank - INSDC) as part of the publication process (Benson et al. 2018). Although this is happening in advanced fields like genomics and proteomics, disciplines like glycomics are lagging behind.

Visualisation

In biology, data visualisation is an essential part of the research process. Scientists have always relied on different visualisation means to communicate experimental results. Some domains of biology like phylogeny (Allende, Sohn, & Little 2015) and pathway analysis (Tanabe & Kanehisa 2012) have created specific visualisations that, nowadays, are considered a standard (i.e. phylogenic trees). The spreading of high throughput technologies has complicated the panorama. The increasing quantity of data and the integration of heterogeneous information have created new challenges for visualisation experts. For example, the rise of the next generation sequencing and the resulting availability of genome data has prompted the need for new custom visualisations to show sequence alignments, expression patterns or entire genomes (Gaitatzes et al. 2018).
Data visualisation is also becoming a crucial resource in the integration of multiple resources. General purpose tools are available to overlay data from different data sources. An example is Cytoscape (Shannon et al. 2003), an open source software platform for visualising interaction networks and biological pathways. Cytoscape gives the possibility to overlay networks with gene expression profiles, annotation and other quantitative and qualitative data.

REFERENCES

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., et al. (2018). How many human proteoforms are there ? Nat Chem Biol 14, 206–214.

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses : 2016 update. Nucleic Acids Res 44, W3–W10.

Akune, Y., Hosoda, M., Kaiya, S., Shinmachi, D., and Aoki-Kinoshita, K.F. (2010). The RINGS resource for glycome informatics analysis and data mining on the Web. Omics J. Integr. Biol. 14, 475–486.

Allende, C., Sohn, E., and Little, C. (2015). Treelink : data integration, clustering and visualization of phylogenetic trees. BMC Bioinformatics 16, 414.

Aoki, K.F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M. (2004). KCaM (KEGG Carbohydrate Matcher) : a software tool for analyzing the structures of carbohydrate sugar chains. Nucleic Acids Res 32, W267–72.

Aoki-Kinoshita, K., Agravat, S., Aoki, N.P., Arpinar, S., Cummings, R.D., Fujita, A., Fujita, N., Hart, G.M., Haslam, S.M., Kawasaki, T., et al. (2016). GlyTouCan 1.0—The international glycan structure repository. Nucleic Acids Res. 44, D1237-1242.

Banin, E., Neuberger, Y., Altshuler, Y., Halevi, A., Inbar, O., Nir, D., Dukler, A., and 献一笠井 (2002). A Novel Linear Code® Nomenclature for Complex Carbohydrates. Trends Glycosci Glycotechnol 14, 127–137.

Bard, J.B.L., and Rhee, S.Y. (2004). Ontologies in biology : design, applications and future challenges. Nat Rev Genet 5, 213.

Barsnes, H., Vizcaíno, J.A., Eidhammer, I., and Martens, L. (2009). PRIDE Converter : making proteomics data-sharing easy. Nat Biotechnol 27, 598.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al. (2017). UniProt : the universal protein knowledgebase. Nucleic Acids Res 45, D158–D169.

Battle, R., and Benson, E. (2008). Bridging the semantic Web and Web 2.0 with Representational State Transfer (REST). Web Semant. Sci. Serv. Agents World Wide Web 6, 61–69.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D., and Sayers, E.W. (2018). GenBank. Nucleic Acids Res 46, D41–D47.

Bohne-Lang, A., Lang, E., Förster, T., and von der Lieth, C.W. (2001). LINUCS : linear notation for unique description of carbohydrate sequences. Carbohydr Res 336, 1–11.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29, 365.

Buneman, P., Khanna, S., and Wang-Chiew, T. (2001). Why and Where : A Characterization of Data Provenance. In Lecture Notes in Computer Science, pp. 316–330.

Campbell, M.P., Ranzinger, R., Lütteke, T., Mariethoz, J., Hayes, C.A., Zhang, J., Akune, Y., Aoki-Kinoshita, K.F., Damerell, D., Carta, G., et al. (2014). Toolboxes for a standardised and systematic study of glycans. BMC Bioinformatics 15 Suppl 1, S9.

Cook, C.E., Bergman, M.T., Cochrane, G., Apweiler, R., and Birney, E. (2018). The European Bioinformatics Institute in 2017 : data coordination and integration. Nucleic Acids Res 46, D21–D29.

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

Doubet, S., and Albersheim, P. (1992). CarbBank. Glycobiology 2, 505.

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

Dudhe, A., and Sherekar, S.S. (2014). Performance Analysis of SOAP and RESTful Mobile Web Services in Cloud Environment. IJCA Spec. Issue Recent Trends Inf. Secur. RTINFOSEC, 1–4.

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G.A., and Berthold, M.R. (2017). KNIME for reproducible cross-domain analysis of life science data. J Biotechnol 261, 149–156.

Gaitatzes, A., Johnson, S.H., Smadbeck, J.B., and Vasmatzis, G. (2018). Genome U-Plot : a whole genome visualization. Bioinformatics 34, 1629–1634.

Goble, C., and Stevens, R. (2008). State of the nation in data integration for bioinformatics. J Biomed Inf. 41, 687–693.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics : current and future challenges. BMC Syst Biol 8 Suppl 2, I1.

Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., Moreau, L., and Oinn, T. (2003). Provenance of e-Science Experiments-experience from Bioinformatics. p.

Halevy, A.Y. (2001). Answering queries using views : A survey. VLDB J 10, 270–294.

Hamid, J.S., Hu, P., Roslin, N.M., Ling, V., Greenwood, C.M.T., and Beyene, J. (2009). Data integration in genetics and genomics : methods and challenges. Hum Genomics Proteomics 2009.

Harvey, D.J., Merry, A.H., Royle, L., Campbell, M.P., and Rudd, P.M. (2011). Symbol nomenclature for representing glycan structures : Extension to cover different carbohydrate types. Proteomics 11, 4291–4295.

Herget, S., Ranzinger, R., Maass, K., and Lieth, C.-W.V.D. (2008). GlycoCT-a unifying sequence format for carbohydrates. Carbohydr. Res. 343, 2162–2171.

Hermjakob, H. (2006). The HUPO proteomics standards initiative–overcoming the fragmentation of proteomics data. Proteomics 6 Suppl 2, 34–38.

Hoehndorf, R., Schofield, P.N., and Gkoutos, G.V. (2015). The role of ontologies in biological and biomedical research : a functional perspective. Brief Bioinform 16, 1069–1080.

Hsiao, S.-W., Sun, Y.S., Ao, F.-C., and Chen, M.C. (2011). A Secure Proxy-Based Cross-Domain Communication for Web Mashups. In 2011 IEEE Ninth European Conference on Web Services, p.

Joshi, H.J., von der Lieth, C.-W., Packer, N.H., and Wilkins, M.R. (2010). GlycoViewer : a tool for visual summary and comparative analysis of the glycome. Nucleic Acids Res. 38, W667-670.

Joshi, H.J., Jørgensen, A., Schjoldager, K.T., Halim, A., Dworkin, L.A., Steentoft, C., Wandall, H.H., Clausen, H., and Vakhrushev,

S.Y. (2018). GlycoDomainViewer : a bioinformatics tool for contextual exploration of glycoproteomes. Glycobiology 28, 131–136.

Karim, M.R., Michel, A., Zappa, A., Baranov, P., Sahay, R., and Rebholz-Schuhmann, D. (2017). Improving data workflow systems with cloud services and use of open data for bioinformatics research. Brief Bioinform.

Kelso, J., Hoehndorf, R., and Prüfer, K. (2010). Ontologies in Biology. In Theory and Applications of Ontology : Computer Applications, (Springer, Dordrecht), pp. 347–371.

Khoury, G.A., Baliban, R.C., and Floudas, C.A. (2011). Proteome-wide post-translational modification statistics : frequency analysis and curation of the swiss-prot database. Sci Rep 1, 90.

Kikuchi, N., Kameyama, A., Nakaya, S., Ito, H., Sato, T., Shikanai, T., Takahashi, Y., and Narimatsu, H. (2005). The carbohydrate sequence markup language (CabosML) : an XML description of carbohydrate structures. Bioinformatics 21, 1717–1718.

Kolarich, D., Rapp, E., Struwe, W.B., Haslam, S.M., Zaia, J., McBride, R., Agravat, S., Campbell, M.P., Kato, M., Ranzinger, R., et al. (2013). The minimum information required for a glycomics experiment (MIRAGE) project : improving the standards for reporting mass-spectrometry-based glycoanalytic data. Mol. Cell. Proteomics MCP 12, 991–995.

Komatsoulis, G.A., Warzel, D.B., Hartel, F.W., Shanbhag, K., Chilukuri, R., Fragoso, G., Coronado, S. de, Reeves, D.M., Hadfield, J.B., Ludet, C., et al. (2008). caCORE version 3 : Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inf. 41, 106–123.

Konishi, Y., and Aoki-Kinoshita, K.F. (2012). The GlycomeAtlas tool for visualizing and querying glycome data. Bioinforma. Oxf. Engl. 28, 2849–2850.

Lapatas, V., Stefanidakis, M., Jimenez, R.C., Via, A., and Schneider, M.V. (2015). Data integration in biological research : an overview. J Biol Res 22, 9.

Lauc, G., Pezer, M., Rudan, I., and Campbell, H. (2016). Mechanisms of disease : The human N-glycome. Biochim. Biophys. Acta BBA - Gen. Subj. 1860, 1574–1582.

Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2015). MatrixDB, the extracellular matrix interaction database : updated content, a new navigator and expanded functionalities. Nucleic Acids Res. 43, D321-327.

Le Pendu, J., Nyström, K., and Ruvoën-Clouet, N. (2014). Host–pathogen co-evolution and glycan interactions. Curr Opin Virol 7, 88–94.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2011). The European Nucleotide Archive. Nucleic Acids Res 39, D28–D31.

Leser, U., and Naumann, F. (2007). Informationsintegration : Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen.

von der Lieth, C.-W., Freire, A.A., Blank, D., Campbell, M.P., Ceroni, A., Damerell, D.R., Dell, A., Dwek, R.A., Ernst, B., Fogh, R., et al. (2011). EUROCarbDB : An open-access platform for glycoinformatics. Glycobiology 21, 493–502.

Lisacek, F., Mariethoz, J., Alocci, D., Rudd, P.M., Abrahams, J.L., Campbell, M.P., Packer, N.H., St\a ahle, J., Widmalm, G., Mullen, E., et al. (2016). Databases and Associated Tools for Glycomics and Glycoproteomics. In Methods in Molecular Biology, pp. 235–264.

Liu, Y., McBride, R., Stoll, M., Palma, A.S., Silva, L., Agravat, S., Aoki-Kinoshita, K.F., Campbell, M.P., Costello, C.E., Dell, A., et al. (2017). The minimum information required for a glycomics experiment (MIRAGE) project : improving the standards for reporting glycan microarray-based data. Glycobiology 27, 280–284.

Luscombe, N.M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics ? A proposed definition and overview of the field. Methods Inf Med 40, 346–358.

Lütteke, T. (2008). Web resources for the glycoscientist. Chembiochem 9, 2155–2160.

Lütteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M., and von der Lieth, C.-W. (2006). GLYCOSCIENCES.de : an Internet portal to support glycomics and glycobiology research. Glycobiology 16, 71R-81R.

Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A., and Ferrari, R. (2018). Genome, transcriptome and proteome : the rise of omics data and their integration in biomedical sciences. Brief Bioinform 19, 286–302.

Masseroli, M., Mons, B., Bongcam-Rudloff, E., Ceri, S., Kel, A., Rechenmann, F., Lisacek, F., and Romano, P. (2014). Integrated BioSearch : challenges and trends for the integration, search and comprehensive processing of biological information. BMC Bioinformatics 15, S2.

Matsubara, M., Aoki-Kinoshita, K.F., Aoki, N.P., Yamada, I., and Narimatsu, H. (2017). WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures. J Chem Inf Model 57, 632–637.

Munevar, S. (2017). Unlocking Big Data for better health. Nat Biotechnol 35, 684–686.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., et al. (2004). Taverna : a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20, 3045–3054.

Packer, N.H., Lieth, C.-W. von der, Aoki-Kinoshita, K.F., Lebrilla, C.B., Paulson, J.C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., and York, W.S. (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.

Patwardhan, A. (2017). Trends in the Electron Microscopy Data Bank (EMDB). Acta Crystallogr Struct Biol 73, 503–508.

Pérez, S., Sarkar, A., Rivet, A., Breton, C., and Imberty, A. (2015). Glyco3D : A Portal for Structural Glycosciences. In Glycoinformatics, (Humana Press, New York, NY), pp. 241–258.

Pérez, S., Sarkar, A., Rivet, A., Drouillard, S., Breton, C., and Imberty, A. (2016). Glyco3D : A Suite of Interlinked Databases of 3D Structures of Complex Carbohydrates, Lectins, Antibodies, and Glycosyltransferases. In A Practical Guide to Using Glycomics Databases, pp. 133–161.

Ponniah, P. (2004). Data Warehousing Fundamentals : A Comprehensive Guide for IT Professionals (John Wiley & Sons).

Prasad, T.S.K., Mohanty, A.K., Kumar, M., Sreenivasamurthy, S.K., Dey, G., Nirujogi, R.S., Pinto, S.M., Madugundu, A.K., Patil, A.H., Advani, J., et al. (2017). Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res 27, 133–144.

Quesada-Martínez, M., Duque-Ramos, A., Iniesta-Moreno, M., and Fernández-Breis, J.T. (2017). Preliminary Analysis of the OBO

Foundry Ontologies and Their Evolution Using OQuaRE. Stud Health Technol Inf. 235, 426–430.

Ranzinger, R., Herget, S., Wetter, T., and von der Lieth, C.-W. (2008). GlycomeDB – integration of open-access carbohydrate structure databases. BMC Bioinformatics 9, 384.

Ranzinger, R., Aoki-Kinoshita, K.F., Campbell, M.P., Kawano, S., Lütteke, T., Okuda, S., Shinmachi, D., Shikanai, T., Sawaki, H., Toukach, P., et al. (2015). GlycoRDF : an ontology to standardize glycomics data in RDF. Bioinformatics 31, 919–925.

Sahoo, S.S., Thomas, C., Sheth, A., Henson, C., and York, W.S. (2005). GLYDE-an expressive XML standard for the representation of glycan structure. Carbohydr Res 340, 2802–2807.

Sahoo, S.S., Thomas, C., Sheth, A., York, W.S., and Tartir, S. (2006). Knowledge modeling and its application in life sciences. In Proceedings of the 15th International Conference on World Wide Web - WWW '06, p.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13, 2498–2504.

Shoaib, M., Ansari, A.A., and Ahn, S.-M. (2017). cMapper : gene-centric connectivity mapper for EBI-RDF platform. Bioinformatics 33, 266–271.

Silva, J.C.F., Carvalho, T.F.M., Basso, M.F., Deguchi, M., Pereira, W.A., Sobrinho, R.R., Vidigal, P.M.P., Brustolini, O.J.B., Silva, F.F., Dal-Bianco, M., et al. (2017). Geminivirus data warehouse : a database enriched with machine learning approaches. BMC Bioinformatics 18, 240.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al. (2007). The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255.

Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., and Brass, A. (2000). TAMBIS : transparent access to multiple bioinformatics information sources. Bioinformatics 16, 184–185.

Struwe, W.B., Agravat, S., Aoki-Kinoshita, K.F., Campbell, M.P., Costello, C.E., Dell, A., Feizi, T., Haslam, S.M., Karlsson, N.G., Khoo, K.-H., et al. (2016). The minimum information required for a glycomics experiment (MIRAGE) project : sample preparation guidelines for reliable reporting of glycomics datasets. Glycobiology 26, 907–910.

Tanabe, M., and Kanehisa, M. (2012). Using the KEGG Database Resource. In Current Protocols in Bioinformatics.

Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.-A., Julian, R.K., Jr, Jones, A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., et al. (2007). The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25, 887.

Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R.D., York, W.S., Karlsson, N.G., Lisacek, F., Packer, N.H., Campbell, M.P., Aoki, N.P., et al. (2017). GlyTouCan : an accessible glycan structure repository. Glycobiology 27, 915–919.

Toukach, P.V. (2011). Bacterial carbohydrate structure database 3 : principles and realization. J Chem Inf Model 51, 159–170.

Turk, Ž. (2006). Construction informatics : Definition and ontology. Adv. Eng. Inform. 20, 187–199.

Varki, A., Cummings, R.D., Aebi, M., Packer, N.H., Seeberger, P.H., Esko, J.D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., et al. (2015). Symbol Nomenclature for Graphical Representations of Glycans. Glycobiology 25, 1323–1324.

Vizcaíno, J.A., Csordas, A., del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2016). 2016 update of the PRIDE database and its related tools. Nucleic Acids Res 44, D447–D456.

Weidman, S., and Arrison, T. (2010). Steps Toward Large-Scale Data Integration in the Sciences : Summary of a Workshop (Washington (DC) : National Academies Press (US)).

York, W.S., Agravat, S., Aoki-Kinoshita, K.F., McBride, R., Campbell, M.P., Costello, C.E., Dell, A., Feizi, T., Haslam, S.M., Karlsson, N., et al. (2014). MIRAGE : the minimum information required for a glycomics experiment. Glycobiology 24, 402–406.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res 46, D754–D761.

Zhang, H. (2016). Overview of Sequence Data Formats. Methods Mol Biol 1418, 3–17.

Ziegler, P., and Dittrich, K.R. (2004). Three Decades of Data Integration — all Problems Solved ? In IFIP International Federation for Information Processing, pp. 3–12.

(2010). Essentials of Glycobiology (Cold Spring Harbor (NY) : Cold Spring Harbor Laboratory Press).

W3C Semantic Web Activity Homepage.History of the World Wide Web - Wikipedia.HTML - Wikipedia.webcomponents.org - Discuss & share web components.Yuan TQ, Sun SN, Xu F, Sun RC (2011) Characterization of lignin structures and lignin-carbohydrate complex (LCC) linkages by quantitative C'13 and 2D HSQC NMR spectroscopy. *J Agric Food Chem* **59:** 10604-10614

Zhang N, Li S, Xiong L, Hong H, Chen Y (2015) Cellulose-hemicellulose interaction in wood secondary cell-wall. Modelling Simul. *Mater. Sci. Eng.* **23**: 85010-85024 doi:10.1088/0965-0393/23/8/085010

Zykwinska A, Thibault JF, Ralet MC (2008) Modelling of xyloglucan, pectins and pectic side chains binding onto cellulose microfibrils. *Carbohydr Polym* **74***(1):23–30*