An Investigation of Carbohydrate Databases and The Design Principles That Govern Them

Literature Review

Joseph Sidley
SDLJOS001@myuct.ac.za
University of Cape Town
Cape Town, South Africa

ABSTRACT

This literature review aims to discuss and analyse databases for carbohydrate molecules. The use of these databases, and the content of them, plays a large role in academic fields and industry. The use of glycans in bioinformatics has long been absent compared to that of proteomics. This may be in part due to the scattered and unstandardised nature of the information surrounding these glycans. Strong database systems will help solve this issue and allow for better use of glycomics in the scientific sphere. Through discussing the differences between current carbohydrate databases, issues these databases face, how to deal with these issues as well as trends towards more usable databases with more functionality and utility, ideas about scientific databases were formed. This culminated in the view that although functionality in these databases is generally quite high, usability and aesthetic elements make them unappealing and uninviting to use, which decreases utility.

CSS CONCEPTS

- **Information Systems → Data Management System**

KEYWORDS

Nuclear Magnetic Resonance (NMR), Relational Database Management System, Serotype, SQL, PHP, Usability Testing, Service Oriented Architecture (SOA)

1. INTRODUCTION

Databases of carbohydrate molecules produced by microbes are useful to scientists. Carbohydrates are present on the surface of many pathogens and malignant cells. They can be targeted in drug and vaccine development [21] and thus having a database of these molecules, where they are found, and their structure, is very important for this purpose. There are an increasing number of databases for carbohydrates, all with different strengths, weaknesses and technologies used, as well as slight differences in purpose. Some databases focus on a specific organism such as e. coli and some are more general. Some store 3D models in the repository and some only 2D or chemical structure. The eight databases this paper will discuss are EK3D [11], Carbohydrate Structure Database [9], ECODAB [17], PolySacDB [1], GlyTouCan [19], Glycosciences.de [20], UniCarbKB/Glygen [5][22] and Glycans Structure Database [8].This review aims to survey and compare the currently available databases for carbohydrate molecules, the area of focus of each database, the functionality provided by these databases and the tools used to build the database. This will be achieved through a summary and a table format. The review will also survey best practice in database design for molecular databases, and common issues faced by databases in specific scientific settings. It will also discuss how to best deal with the issues that these databases face and how database design practices are changing to deal with these. A critical analysis of the databases compared to best design practices will then be performed and discussed.

2. CURRENT DATABASES

2.1 Database Summary

Eight databases were analysed in order to find out the technology that they are built upon, the specific search queries that they used, the representations that they outputted and any extra features that the databases had in their specific fields. Any specific uses of the databases were noted. Along with this, my own personal usability notes are added to the summary table, along with if the database had any user testing reported in the research paper of their creation.

EK3D: This e. coli K-antigen database has many of the features needed in any biochemistry database: a search function, a repository, a connection to a 3D modeller and it provides information on the sugar compositions. Through the search function, one can search sugars and it returns the K-antigens which have that sugar fragment in them. It was created by the Indian Institute of Technology. It uses Jsmol under a png format to show the 3D model of the molecular structure.

Carbohydrate Structure Database: This is a large carbohydrate structure database, with many necessary feature for a biochemistry/carbohydrate database, created by the N.D. Zelinsky Institute of Organic Chemistry in Russia. This database has 28634 compounds in it. It can be searched by structure, composition, organism, NMR signals and publication, along with having other statistical and analytical tools incorporated into it. You can search for a structure using CSDB nomenclature and it will return it as a 2D model in two different styles.

ECODAB: ECODAB is similar to EK3D but it looks at e. coli O-antigens rather than K-antigens. It has data on 169 O-antigen entries and 338 glycosyltransferases. It is stripped back and minimalist in terms of design, while still having all the

necessary data for each antigen or glycosyltransferase. It does not deal with the 2D or 3D models of the molecules.

PolySacDB: This database of antigenic polysaccharides from 347 Microbes, has a lot of data, and search queries on each polysaccharide such as: origin, function, antibodies produced, antibody utilities and 13 more. The structure ID redirects to the CSDB [9] earlier described. It is very detailed in the way it can be traversed but has relatively few entries and is thus not an all-encompassing database.

GlyTouCan: This is registry of 160 000 glycans. Users can register new glycans if the structure is viable, and can query glycans already registered in the database. A glycan is given an identification number and its structural representation is added. It is a large repository but it is barebones on extra information.

Glycosciences.de: A database with a particular focus on 3D modelling of the carbohydrates and the storage of those models. It has 28000 entries and 15000 3D models. A user can model their own 3D carbohydrate image and can use premodelled 3D images from the database. The search query options are extensive, with a focus on structure, publication and NMR.

UniCarbKB/Glygen: GlyGen is the successor to UniCarbKB. It integrates results from the original UniCarbKB Database as well as GlyTouCan [5] and the Protein Data Bank [21]. It provides the user with data on number of sugars, glycoproteins, structural composition and monoscopic mass based on a query of ID, organism, structure or enzyme. Note: EuroCarb was non-functional at the time of writing, and is an older database which was a conceptual framework for UniCarbKB/Glygen.

Glycans Structure Database: In this small stripped back database, one can search by sub-structure, molecular weight, IUPAC codes and composition or a combination of these. The user can also search using GlycanBuilder input. It returns chemical structure along with ID, family and source (organism).

Paper/Database	Language(s)	Purpose	Search Functions	Representations and Extra Features	User Testing and usability notes
EK3D [11]	Apache web-server management system and PHP web-server scripting language.	Used for E coli K-antigens.	By sugar, by disaccharide fragment, by keyword, by group serotype.	Multimer modelling. Schematic, chemical and 3D representation	No. Very usable. Intuitive and understandable database
Carbohydrate Structure Database [9]	PHP 5.4 programming language, MySQL 5.5 database engine and DHTML/Javascript for the-web pages	Used for general carbohydrates, with a focus on NMR Prediction	Structural Fragment, ID, Structural composition, Taxonomy, Publication, NMR Signal threshold.	Schematic representation, NMR Prediction	No. Usable, understandable, with good documentation and all necessary information seems to be included
ECODAB [17]	Relational database managed by MySQL, front end coding in PHP	Used for E coli O-antigens	Search by: String, Chemical Shift, GT Name, GENEBANK code.	Chemical Representation, NMR information	No. Easy to use if not a bit basic, but understandable and well put together
PolySacDB [1]	PHP, HTML and CSS used to build the web interface. MySQL works at the backend	Used for general carbohydrates. Seems very all purpose with 17 query options	Very extensive search, add a query and tick one of 17 boxes for it to search in that realm (includes ID, antigenic nature Carbohydrate name, microbe, drugpedia, and structure).	Hyperlinks to CSDB [9] for representation, has information on function, proposed utility, antigenic nature and antibodies	No. Easy to use, has a lot of information on each carbohydrate, but still very intuitive. A bit cluttered.
GlyTouCan [19]	UNKNOWN	Used for general carbohydrates.	Search by 2D schematic representation (GlyanBuilder) and by textual chemical representation (broken as of this moment).	Takes you to GlyCosmos for information on carbohydrate. Seems to be slightly broken – with errors coming when opening some structures	No. Hard to use, confusing UI, sometimes seems to be broken, many hyperlinks to different websites where user experience is different.
Glycosciences.de [20]	UNKNOWN	Used for general carbohydrates.	Search by graphical input (GlycanBuilder), textual input, composition, molecular formula, publication, and NMR.	Supports 3D modelling, PDB analysis and PDB validation	No. Easy to use, full of good necessary information, intuitive design
UniCarbKB/GlyGen [5] [22]	UI built in Scala along with Jquery and Bootstrap JS libraries. Backend built with PostgreSQL 9.2	Used for general carbohydrates.	Search by taxonomy, tissue, protein, accession and disease.	Represented as 2D schematic representation	No. Hard to use, the distinction between Glygen and UniCarbKB is confusing.
Glycans Structure Database [8]	UNKNOWN	Used for general carbohydrates	Search by textual input: composition, molecular formula or structure, can search by structure or by GlycanBuilder input	Links to 3D model in glycam (currently non-functional), shows schematic and chemical representation	No. Easy to use but maybe too basic, and uninviting design

Table 1: Overview of eight carbohydrate databases

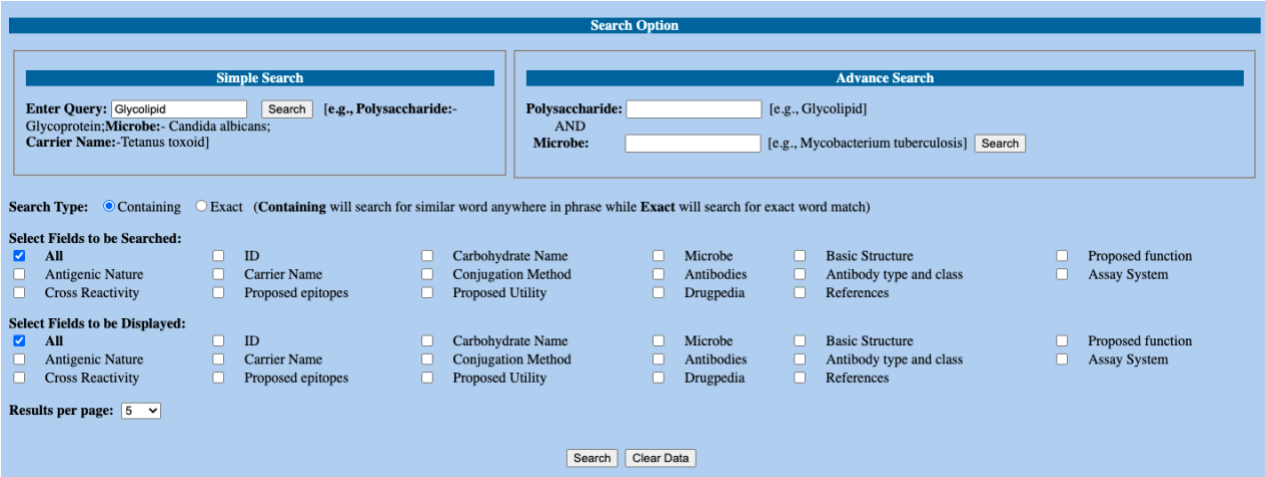

Figure 1: PolySacDB query options

2.2 Findings

The main language used for coding the database is SQL and the most commonly used database management systems are MySQL and PostgreSQL. On the front end, JavaScript, PHP, HTML, CSS and Scala are used to develop the websites used for framing the database. It seems as though PHP is the most widely used.

The query functions vary, but a core of some combination of a type of identification number, textual input of structure, graphical input of structure and taxonomical information seem to be the most common and important.

In general the usability level is relatively high, especially as someone with background in the topic. Some software issues do occur and the aesthetic and design aspects may have not been focused on as much as some other databases, which have graphical ways to enter queries such as the Human Virus Database [21]. There is some cluttered design, and some confusion in some of the databases, but the databases do reach their goals most of the time.

3 SCIENTIFIC DATABASE DESIGN BEST PRACTICES

3.1 Issues

Scientific databases have different uses and different needs compared with databases in other fields and thus different challenges [15]**.** This means they must deal with these challenges in a different way. There are some guidelines that one can follow in order to develop usable scientific software, and in specific macromolecule databases. The need for new types of queries as data becomes better understood means that databases must be built in a more innovative and mutable way. The diversity of types of data and types of queries is a big challenge for the database designer. Depending on the specific application of the database, the volume and size of data stored may be an issue, along with the complexity of the data that is being stored.

Another issue discussed is the non-standardisation of glycan structure formats. There are many different format structures used in these databases such as IUPAC, LinearCode, KEGG Chemical Function, Oxford Format and GlycoCT. This may present an issue for anyone using multiple databases in conjunction as these are often not compatible or interchangeable with each other [13].

Furthermore, a concern in scientific databases is the lack of computer aided data analysis. In the medical sphere for example, analysation systems have been integrated with database systems to great success, as it speeds up the calculation of vital information needed by the user of the database [14]. In specific situations in glycomics, such as NMR prediction, this can be a useful tool in molecular databases as well. It has however, usually been left to the scientist to do this themselves, rather than the database to pre-calculate. Although often not strictly necessary it is a feature that improves usability and speeds up processes for the users of the database.

3.2 The technological options

The use of relational database frameworks such as MySQL in scientific applications is common and widespread, but the challenge is that sometimes highly specific and complex data types are not supported and image storing is harder with these databases compared to object oriented or object-relational database frameworks such as PostgreSQL [15]. All databases surveyed were connected to the internet, and not institutional. Any recent database for widespread use should be internet connected, using any internet framework. In any case, the queries used should be built for intent, speed, and to maximise utility [15].

3.3 How to handle issues:

The IEEE has many technical recommendations about scientific and statistical databases, which strengthen theoretical knowledge of how a database should be built in order to stay up to date with industry needs [18]. The use of compression, and importance of efficiency in manipulation and access of large data pool is paramount. To achieve this, multidimensional

data structures and different file organisation strategies can be used. The ideal database should be one that "provides data to users in a way that maximises their immediate utility". The application of that statement to a specific database is dependent on the level of complexity the users require in that field. Another important condition which much be followed is that of simplicity [4]. This seems to be paradoxical to the previous aims, but a balance between simplicity and ease of use on the front end and staying complex enough on the backend to maximise utility is paramount.

A solution proposed to the non-standardisation of the database formats is the use of GlycanBuilder, which can be used for drawing molecular structures, but also the conversion of structural formats [13].

3.4 Gold Standard

The Protein Data Bank is the gold standard for macromolecule database design, and has been in used for decades [3]. Many of the databases surveyed seem to have been built using a similar design model following on from the PDB's success. This model is very functional, with all of the query options needed for its field and a very large repository. It does not however have design aesthetic elements developed, which may put some people off of using it as it does not seem inviting or easily understandable [6]. The database design seems to have ignored aesthetic and instead is very focused on functionality. Most databases surveyed in this paper seem to follow on from this lack focus on visual components. This is shown very clearly in Figure 1. In Figure 1 the functionality level is very high but it is quite cluttered and is not an attractive system to use. It looks as though it was built with purely functionality in mind. This shows the Aesthetic-Usability effect in action, as the look of Figure 1 makes a user assume it is a less usable system due to its look rather than its actual function [6].

3.5 Maggie

The world of databases has been one with relatively little change, but the Maggie query system was developed in 2019 [16]. This system was an integrated chatbot and natural language processor that tried to understand the intent of the queries rather than having the user understand the functionality of the interface entirely. It was developed as a service oriented architecture, as a way to better the user query experience. It was then usability tested against the Protein Data Bank and other conventional databases. Ease of use, usefulness, ease of learning and satisfaction were all seen as improved in Maggie compared to traditional databases.

The concept of user feedback is something not present in any of the surveyed databases of this paper, and something which is shown to be important in increasing usability [12].

4. DISCUSSION

Most of the current carbohydrate structure database solutions do fulfil most of their goals, however we cannot tell whether the code of the database fulfilled speed and storage best practices set out by the design papers. Most of the databases have very basic design, which is not necessarily a negative as it keeps simplicity, however aesthetic design is not seen as a necessity in many of these databases, which may increase usability if followed. Following from this, usability testing is completely ignored in all surveyed databases. As software moves more towards human centred design, this seems like an oversight and a remnant of a less mature era of software engineering. The databases all seem to have a similar programming structure using PHP and SQL most commonly, and what sets them apart is their query functions and the content of the databases, as well as the outputted molecular models and any extra information or features. The main focus is most of these databases should be, and does seem to be, allowing query types that maximise the utility of the user, and in this regard, more options usually leads to a system that is easier to use.

Newer models of database query systems, such as Maggie [16], have shown that changing a database's design to fit more with modern view on usability, will have better results when tested compared to older database designs. This is once again shown with the Human Virus Database [21], which although not molecule focused, has a design that is more attractive to a user.

5. CONCLUSIONS

There are many databases of macromolecules in general and carbohydrates in specific. These are designs that usually work, and ones that have stood the test of time. They are however basic in aesthetic and user experience, and not usability tested. There are many ways outlined in this paper to better a database's design – but increasing usability and utility is the focus of all of these. This means the speed, the choice of query type options and the design of the database are all areas to focus on.

REFERENCES

[1] Abhijit Aithal, Arun Sharma , Shilpy Joshi et al. 2012. PolysacDB: A Database of Microbial Polysaccharide Antigens and Their Antibodies. PLOS ONE Volume 7, Issue 4, April 2012, 1-4. DOI: https://doi.org/10.1371/journal.pone.0034613

[2] Rena Astronomo, Dennis Burton. 2010. Carbohydrate vaccines: developing sweet solutions to sticky situations?. *Nat Rev Drug Discov,* Volume 9**,** 308–324 (Aprin 2010). DOI: https://doi.org/10.1038/nrd3012

[3] Helen Berman, John Westbrook, Zukang Feng et al. 2000. The Protein Data Bank, *Nucleic Acids Research*, Volume 28, Issue 1, 1 January 2000, 235–242 DOI: https://doi.org/10.1093/nar/28.1.235

[4] Ewan Birney, Michele Clamp. 2005. Biological database design and implementation, *Briefings in Bioinformatics*, Volume 5, Issue 1, March 2004, Pages 31–38, DOI: https://doi.org/10.1093/bib/5.1.31

[5] Matthew Campbell, Robyn Peterson, Julien Mariethoz et al. UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res*. 2014;42(Database issue):D215-D221. DOI: https://doi.org/10.1093/nar/gkt1128

[6] Nick Cawthon and Andrew Moere. 2007. The Effect of Aesthetic on the Usability of Data Visualization, *2007 11th International Conference Information Visualization (IV '07)*, , pp. 637-648, DOI: https://doi.org/10.1109/IV.2007.147

[7] Dunren Che, Yangjun Chen and K. Aberer. 1999. A query system in a biological database. IEEE Proceedings. *Eleventh International Conference on Scientific and Statistical Database Management*, 1999, pp. 158-167, DOI: https://doi.org/10.1109/SSDM.1999.787631

[8] Catherine Cooper , Mathew Harrison, Marc Wilkins, et. al. 2001. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Research*, Volume 29 Issue 1 (2001), 332–335. DOI: https://doi.org/10.1093/nar/29.1.332

[9] Ksenia Egorova, Anna Kondakova, Phyl Toukach. 2015. Carbohydrate Structure Database: tools for statistical analysis of bacterial, plant and fungal glycomes. *Database (Oxford)*. Volume 2015, September 2015. DOI: https://doi.org/10.1093/database/bav073

[10] Martin Kersten, Stratos Idreos, Stefan Manegold et. al. 2011. The researcher's guide to the data deluge: querying a scientific database in just a few seconds. *ACM*, Volume 4, Issue 12 (August 2011), 1474–1477. DOI: https://doi.org/10.14778/3402755.3402799

[11] Bharathi Reddy Kunduru, Sanjana Anilkumar Nair, Thenmalarchelvi Rathinavelan. 2015. EK3D: an *E. coli* K antigen 3-dimensional structure database. *Nucleic Acids Research*, Volume 44, Issue D1, January 2016, D675–D681. DOI: https://doi.org/10.1093/nar/gkv1313

[12] Markus List, Peter Ebert & Felipe Albrecht 2017. Ten Simple Rules for Developing Usable Software in Computational Biology. *PLOS Computational Biology*, Volume 13, Issue 1, DOI: https://doi.org/10.1371/journal.pcbi.1005265

[13] Xing Li, Zhijue Xu, Xiaokun Hong et. al. 2020. Databases and Bioinformatic Tools for Glycobiology and Glycoproteomics. *International journal of molecular sciences*, Volume 21 Issue 18, (2020) 6727. DOI: https://doi.org/10.3390/ijms21186727

[14] Phillip Metnitz, P. Laback, Christian Popow et al. 1995. Computer assisted data analysis in intensive care: the ICDEV project-development of a scientific database system for intensive care. *J Clin Monitor Comput* Volume 12**,** 147–159 (1995). DOI: https://doi.org/10.1007/BF02332689

[15] William Michener and James Brunt. 2009. *Ecological Data*. John Wiley & Sons, Chichester, 2009, 48-58.

[16] Ritzel Paixao-Cortes, Walter & Paixão-Cortes, Vanessa & Ellwanger, Cristiane & Norberto de Souza, Osmar. 2019. Development and Usability Evaluation of a Prototype Conversational Interface for Biological Information Retrieval via Bioinformatics, pp. 575-594, June 2019, DOI: https://doi.org/10.1007/978-3-030-22660-2_43

[17] Miguel Rojas-Macias, Jonas Ståhle, Thomas Lütteke, Göran Widmalm, Development of the ECODAB into a relational database for *Escherichia coli* O-antigens and other bacterial polysaccharides, *Glycobiology*, Volume 25, Issue 3, March 2015, 341–345. DOI: https://doi.org/10.1093/glycob/cwu116

[18] Arie Shoshani and H. K. T. Wong. 1985. "Statistical and Scientific Database Issues," *IEEE Transactions on Software Engineering*, vol. SE-11, no. 10, pp. 1040-1047, Oct. 1985, DOI: https://doi.org/10.1109/TSE.1985.231851

[19] Michael Tiemeyer, Kazuhiro Aoki, James Paulson et. al. 2017. GlyTouCan: an accessible glycan structure repository, *Glycobiology*, Volume 27, Issue 10, October 2017. Pages 915–919. DOI: https://doi.org/10.1093/glycob/cwx066

[20] Philip Toukach, Hiren J Joshi, René Ranzinger, Yuri Knirel, Claus-W. von der Lieth, Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the *Bacterial Carbohydrate Structure DataBase* and *GLYCOSCIENCES.de*, *Nucleic Acids Research*, Volume 35, Issue Suppl 1, 1 January 2007, Pages D280–D286. DOI: https://doi.org/10.1093/nar/gkl883

[21] Sifan Ye, Congyu Lu, Ye Qiu et al. 2022,.An atlas of human viruses provides new insights into diversity and tissue tropism of human viruses, *Bioinformatics*, 2022, DOI: https://doi.org/10.1093/bioinformatics/btac275

[22] William York, Raja Mazumder, Rene Ranzinger, Nathan Edwards, Robel Kahsay et. al. 2020. GlyGen: Computational and Informatics Resources for Glycoscience, *Glycobiology*, Volume 30, Issue 2, February 2020, Pages 72–73, DOI: https://doi.org/10.1093/glycob/cwz080