Glycomics and Lipidopics

Extended Abstract

Glycosciences.DB: an annotated data collection linking glycomics and proteomics data

Michael Böhm

Abstract:

Glycosciences.DB, the glycan structure database of the Glycosciences.de portal, collects various kinds of data on glycan structures, including carbohydrate moieties from worldwide Protein Data Bank structures. This way it forms a bridge between glycomics and proteomics resources. A major update of this database combines a redesigned web interface with a series of new functions. These include separate entry pages not only for glycan structures but also for literature references and entries, improved substructure search options, a newly available keyword search covering all types of entries in one query, and new types of information that is added to glycan structures. These new features are described in detail in this article, and options how users can provide information to the database are discussed as well. Glycosciences.DB Carbohydrates, often referred to as glycans, are one of the four major classes of biomolecules, next to nucleic acids, proteins and lipids. Of these, carbohydrates are the most abundant and also the most complex molecules. Besides their well-known functions as energy storage or structural components they are parts of glycoproteins or glycolipids and cover cell surfaces in the glycocalyx. Here, they serve as recognition sites for cell–cell and cell–matrix interactions but also for pathogens such as viruses, which frequently interact with glycans on the cell surface to enter their host cells. Glycans are also involved in immune responses, inflammation and diseases such as cancer. Carbohydrates are often specifically recognized. For example, human and avian influenza viruses recognize their hosts by specific glycan motifs. Therefore, researchers in glycomics-related projects need to be able to find information on the specific glycans they are interested in. *Glycosciences.DB*, formerly known as *SweetDB* , was one of the first efforts to collect information on carbohydrate structures and make them available online. Initially seeded with data from the discontinued Complex Carbohydrate Structure Database (CCSD, often referred to as *CarbBank*), further information has been added over the years, such as 3D structure models generated by *Sweet-II* , nuclear magnetic resonance (NMR) spectra imported from *SugaBase* or manually entered from the literature, or links to entries of the worldwide Protein Data Bank ( ) that feature carbohydrates. Currently, the is the main source of new data in *Glycosciences.DB*. At the time of writing, *Glycosciences.DB* contains ~25 000 glycan structure entries with 12 500 3D structure models, 20 000 literature references, 3400 ^{1}H or ^{13}C NMR spectra and more than 10 000 references to carbohydrate-containing entries. In 2018, a major update of the *Glycosciences.de* portal has been released, which not only puts a more modern design to the portal but also adds a series of new functionality to *Glycosciences.DB*, including improvements in search features and in display of information. Prior to the 2018 update, only glycans had been considered as entries in *Glycosciences.DB*. All other items such as literature references or structures were only displayed as parts of the glycan entries or in search result lists. Now structures and publications also receive individual entry pages, which display more data than in the former release The three types of entries, i.e. glycans, publications and structures, are cross-linked with each other. For each entry type an individual symbol is used, that is displayed in the header of the entry and also used in cross-links and search results lists, so that users can directly see what kind of entry will be opened in a link. Screenshots of *Glycosciences.DB* glycan structure entry (front, truncated at the dashed line), literature entry (middle) and entry (back). All three entries are linked to each other: The entry contains both the displayed N-glycan core structure entry and the literature reference. No glycan structure is registered yet with the literature entry; the link to the N-glycan core structure entry is assigned via the entry. New entries are added weekly by downloading newly released structures from the and searching them for carbohydrate moieties. This process is mostly automatic. Human intervention is only required in case of potential problems, such as mismatches between the residue name and the residue that is actually present in the 3D structure, or newly introduced residue names for which no definition is stored in pdb2linucs and pdb-care, the tools used to detect and validate the glycans in structures. The primary citation of a entry is also imported from the and stored in *Glycosciences.DB*. This way entries can be automatically linked with both glycan and literature entries. Cross-links between the latter two types of entries cannot be added automatically in a dependable manner, as there is no tool available that can reliably extract information on relevant carbohydrates from a publication. Nevertheless, the

Michael Böhm

Institute for Virology, University of Cologne, Fürst-Pückler-Str. 56, 50935 Cologne, Germany.

primary reference of a entry often also deals with the carbohydrates in that entry, in particular in case of protein–carbohydrate complexes, where the carbohydrate moieties have been added on purpose and thus are usually (but not certainly) also an important topic of the publication. This is not necessarily the case with glycoproteins, where the glycans also might be a major topic of the publication, but often (particularly in case of short, truncated glycans) are just stated as "also detected" or even not mentioned at all. Therefore, cross-links between glycan and literature entries that are assigned via entries are not listed together with manually assigned cross-links, but in a separate section, so that users can identify them easily. Glycan structure entries still form the main part of *Glycosciences.DB* content. The entries collect information on a carbohydrate structure, such as 3D structure models, NMR spectra, literature references, references to entries and information on residue composition, substructure motifs, trivial names and taxonomy data. The 2018 update comes along with some further items. The glycan structure information (monosaccharid sequence and linkage positions) was only given in a 2D annotation in CarbBank format so far. Now we also offer the structure in LInear Notation for Unique description of Carbohydrate Sequences (LINUCS) notation, the notation internally used in the database to store and identify the glycan structures, and, where possible, in GlycoCT_condensed and GlycoCT_xml format . For further information on glycan structure formats, please refer to . In addition to these text formats, Symbol Nomenclature For Glycans (SNFG) graphs have also been added to many glycan entries. At the time of writing, however, not all of the newly defined features of the current SNFG version are incorporated yet. Cross-links to corresponding entries of other databases of the *Glycosciences.de* portal (*GlycoMapsDB* and *GlycoCD*) are also given now where applicable. A feature that is used by many genomics, proteomics or literature databases but to our knowledge not yet by glycomics databases is the option to add keywords to a database entry, which can be used to identify that entry in a database search. This option is implemented now in *Glycosciences.DB*. In analogy to literature entries and entries, titles can now be added to glycan structure entries in *Glycosciences.DB*. It will be hardly possible to add meaningful titles to all entries. Nevertheless, there are various glycans for which trivial names are widely used (e.g. for Lewis-type blood group antigens, human milk oligosaccharides , glycosphingolipids of the ganglio series, etc.), and for many other glycans a brief description such as 'core-fucosylated N-glycan core structure' can be helpful for users who are not yet familiar with glycan structures. These titles can be used in database queries as well, and they are displayed together with the glycan structure in structure query results and in structure lists e.g. in literature entries to help users to identify the displayed glycans. The 3D structure models that are provided with many entries can give researchers a perception of what the glycans look like. However, it can be difficult to read a glycan's 3D structure and find a specific residue within the structure, because the monosaccharide building blocks that form the glycans are very similar to each other. Therefore, we have added an option to color-highlight the residues using the colors of the SNFG symbols, which makes it easier to orient oneself in a glycan 3D structure. Halos or bond colors can be toggled with the check-boxes in the display options next to the 3D structure. So far, colors are set by PDB 3-letter codes for frequently occurring residues. The list of supported 3-letter codes will be further extended to cover more residues in the future. Residue highlighting in a plant N-glycan with core fucosylation and xylose (LinucsID 13934). Without highlighting, the residues are difficult to identify (top). This becomes easier when halos (bottom left) or bond colors (bottom right) are used with colors matching those of the SNFG symbols, even when the structure is oriented in a different way than the SNFG symbols.

Michael Böhm

Institute for Virology, University of Cologne, Fürst-Pückler-Str. 56, 50935 Cologne, Germany.