scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Source files of the Carbohydrate Structure Database: the way to sophisticated analysis of natural glycans

Philip V. Toukach & Ksenia S. Egorova

The Carbohydrate Structure Database (CSDB, http://csdb.glycoscience.ru/) is a free curated repository storing various data on glycans of bacterial, fungal and plant origins. Currently, it maintains a close-to-full coverage on bacterial and fungal carbohydrates up to the year 2020. The CSDB web-interface provides free access to the database content and dedicated tools. Still, the number of these tools and the types of the corresponding analyses is limited, whereas the database itself contains data that can be used in a broader scope of analytical studies. In this paper, we present CSDB source data files and a self-contained SQL dump, and exemplify their possible application in glycan-related studies. By using CSDB in an SQL format, the user can gain access to the chain length distribution or charge distribution (as an example) in a given set of glycans defined according to specific structural, taxonomic, or other parameters, whereas the source text dump files can be imported to any dedicated database with a specific internal architecture differing from that of CSDB.

Background & Summary

Glycoinformatics is a relatively new research branch, which provides the scientists with various means of accessing, processing and handling all sorts of carbohydrate-related data[1]. The broad usage of glycomic databases and associated software tools has been recently reported[2–8]. Similar to other data-related scientific branches, glycoinformatics heavily depends on high-quality data repositories. In the last decades, several such repositories have been developed. They include a historical CCSD project (CarbBank; contained more than 15,000 natural glycans before it was discontinued in 1996; the source of older data for most of the existing carbohydrate databases)[9]; Glycosciences.DB (contains the CCSD data supplemented with NMR spectra, 3D structures and analytical tools)[10,11]; UniCarbKB (contains eukaryotic glycans supplemented with NMR, MS and HPLC data)[12]; KEGG Glycan (glycan-related data from the Kyoto Encyclopedia of Genes and Genomes)[13]; Japan Consortium for Glycobiology and Glycotechnology (JCGG/ACGG collection of databases on glycoproteins and glycome-associated diseases supplemented with analytical data)[14]; and CSDB (the Carbohydrate Structure Database, see below)[15], to name a few.

Successful application of any database depends not only on the quality and completeness of its data, but also on the capabilities and user friendliness of its interface. Thus, most of the chemical and biological databases, including carbohydrate ones, are equipped with a web interface. However, in many cases the source database files contain much more data than those accessible via the Internet, because the frontend interfaces and the backend tools behind them are usually designed to serve only the most popular and demanded user queries.

The Carbohydrate Structure Database (CSDB, http://csdb.glycoscience.ru/) is a free curated repository, which stores various types of data (structural, taxonomical, bibliographical, NMR spectroscopic, etc.) on glycans of bacterial, fungal and plant origins[15]. One of the most significant characteristics of CSDB is its completeness[15]. Currently, it provides a close-to-full coverage on bacterial and fungal carbohydrates up to the year 2020. The fungal coverage has been achieved in 2021 and has not been reported elsewhere.

CSDB is supplied with a web interface, which provides free access to the CSDB content and dedicated data analysis and simulation tools. These tools include coverage statistics, monomeric residue properties,

N.D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Leninsky prospect 47, Moscow, 119991, Russia. e-mail: netbox@toukach.ru; egorova-ks@ioc.ac.ru

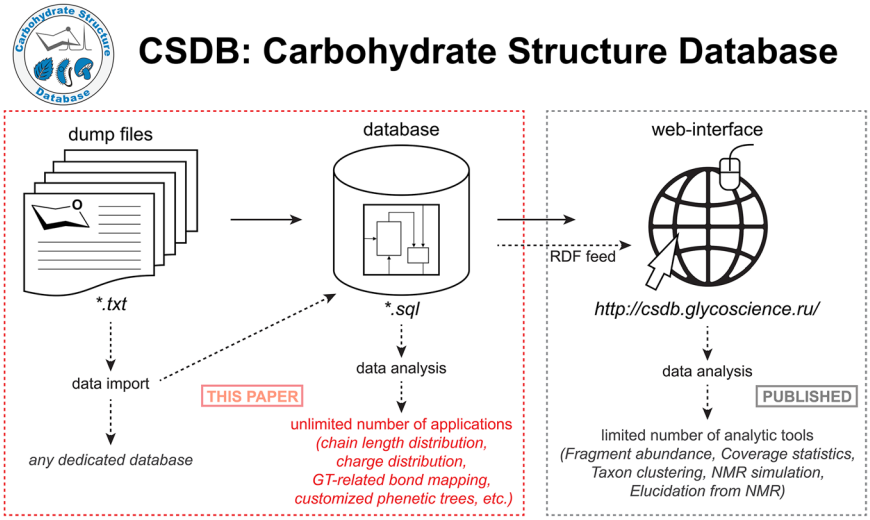

Fig. 1 Source text dump files and SQL files for CSDB are reported in this paper. The CSDB web-interface, associated web tools, and RDF-ized data have been reported elsewhere[15,16,26,37]. Solid arrows represent immanent logic of the database; dashed arrows show inferred data flows upon usage.

multiparametric analysis of distribution of carbohydrate structural elements among taxa[16], simulation of 1D and 2D NMR spectra[17–19], NMR-based structure elucidation[20], and structure translators to various carbohydrate and chemical notations[21] and optimized atomic coordinates[22]. CSDB is integrated with a glycosyltransferase database (CSDB_GT), which currently covers GTs from the three most studied non-animal model species (*Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*)[23–25].

According to user feedback, citing and access log analysis, the above-listed CSDB instruments are most demanded in routine research on natural carbohydrates. However, the number of these tools and the types of the corresponding analyses is limited, whereas the database itself contains data that can be used in a broader scope of analytical studies. For example, by using the existing database in an SQL format, the user can gain access to the chain length distribution or charge distribution in a given set of glycans, which can be defined according to specific structural, taxonomic, or other parameters. In their turn, the source text dump files can be potentially imported to any dedicated database with a desired internal architecture differing from that of CSDB (Fig. 1). In this paper, we present CSDB source text files (called dump files) and a self-contained SQL backup, and exemplify their possible application in glycan-related studies. By using these dump files, scientists can build dedicated databases suited for their particular scientific needs. The CSDB data can also be downloaded as an RDF feed generated within the GlycoRDF ontology[26] for further import to an external triplestore.

Methods

Database architecture. CSDB stores data in a MySQL relational database. For structures, the connection table approach is used, where nodes are monosaccharide and other residues, and vertices are bonds with elimination of water. Relationships between the data from scientific publications and their indices are visualized in Fig. 2. The data are stored in database tables (see caption to Fig. 2). The following data categories are used (as reflected by the color of the table headers): molecular structure (violet); compound as a whole (cyan); bibliography (red); NMR spectra (pink); taxonomy (green); glycosyltransferases (olive); simulated conformations (grey); and main relations (yellow). The SET data type means a term from a controlled vocabulary, including large lists, such as monomer names, species, journal names, etc. Where not explained in the figure (marked with *), the following controlled vocabularies are implied:

- organisms.tax_group - bacteria, archaea, fungi, plant, protista, animal, mammal, human, etc.
- main_link.tax_group - the same as above (denormalized).
- conformations.methods - MM3-2000, MMFF-94, GLYCAM, AMBER, CHARMM36, OPLS-AA, PDB.
- conformations.solvent - none, GB, STIL, TIP3P, etc.
- compounds.unit_type - chem, biol, sbiol, oligo, mono, homo, cyclo, fragment, motif.
- link_types.link_type - glycosidic, amidic, amine, diester, carbon-carbon, etc.
- disease.attr_name (attributes) - ICD code, Life stage, Sex.
- gtr.molecule role - O-antigen, CPS, EPS,core, lipid A, GPI, N-glycan, O-glycan, C-glycoside, etc.
- gtr.confirmed - *in vivo*, indirect, semi-direct, *in vitro*, *in silico*, suggested.
- publication_specific.synthesis - chemical, enzymatic, fragmentary, biosynthesis, etc.
- nmr_solvents.unit - %, vol %, M, mkM, etc.
- external_resources.resource - CA, PubChem, GlycomeDB, CCSD, US patent, GlyTouCan, etc.

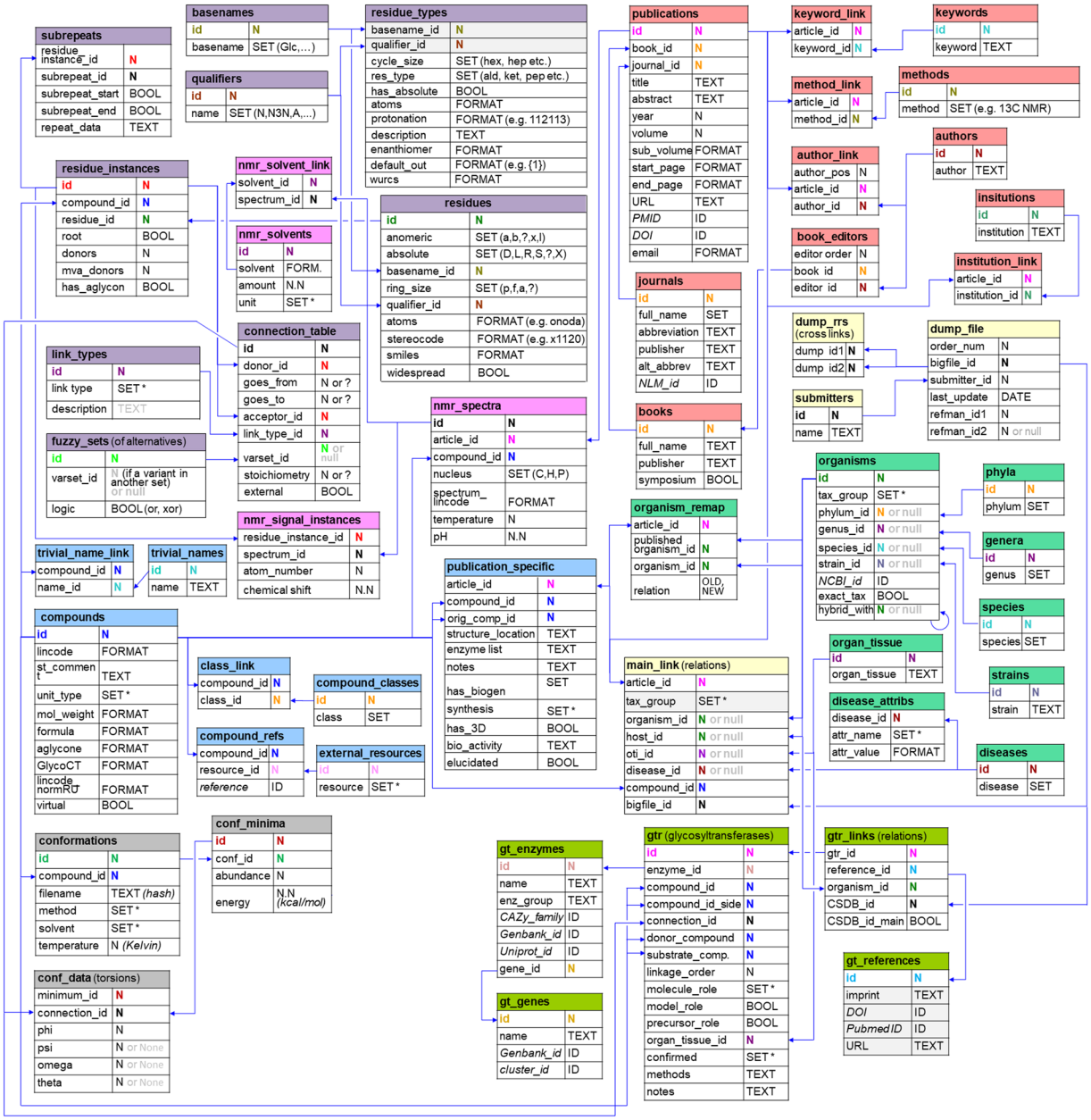

Fig. 2 CSDB entity relationship scheme. In each table, the first column corresponds to the field, and the second column – to the data type: N, integer (the symbols of the same color correspond to the same indices connected by arrows in different tables); N.N, float; TEXT, text; FORMAT, formatted text, SET, controlled vocabulary term (see the main text); BOOL, boolean switch; ID, identifier in the external database. The color of the table headers reflects the data category: violet = molecular structure; cyan = compound as a whole; red = bibliography; pink = NMR data; green = taxonomy; olive = glycosyltransferases; grey = simulated conformations; yellow = main relations. The table meaning is explained in parentheses, where unclear from the table name (shown in bold). Blue arrows show one-to-many relations between the fields. Links to external resources are shown in italic; denormalized data are greyed.

The data are imported from main text dumps (see below), with a few exceptions, the detailed description of which are beyond the scope of this paper.

The conformation map subdatabase[27] (grey headers in Fig. 2) is imported from a set of molecular dynamics files (XML, one file per molecule, described at http://csdb.glycoscience.ru/jsmol/confmap_data/processed_trajectory_format.txt) generated automatically by a dedicated postprocessor of molecular dynamics trajectories simulated by the CAT software (Conformation Analysis Tools[28]). Generation of these files is automatized and implemented at the CSDB calculation server; *ca.* 20–30 new files are completed monthly. Currently, there are 2597 data files available for download at http://csdb.glycoscience.ru/jsmol/confmap_data/minima/, including those imported from the GlycomapsDB[29] database.

The Glycosyltransferase subdatabase (olive headers in Fig. 2) is imported from a separate set of UTF-8 text dumps that are exported from Microsoft Excel spreadsheets filled by another team of annotators. A detailed description of the glycosyltransferase dumps is beyond the scope of this paper.

Averaged chemical shifts and glycosylation effects used in the empirical NMR spectrum simulation together with a database-driven approach are stored in a set of text files, cached to memory, and used directly upon NMR simulation.

A vocabulary of supported monomeric residues, their atomic properties, stereo codes, and their records in WURCS[30] and SMILES[31] notations are imported from separate text files (http://csdb.glycoscience.ru/database/core/residues.txt and http://csdb.glycoscience.ru/database/core/smiles.txt).

Annotation rules. The CSDB database is supplemented with data by means of retrospective analysis and annotation of scientific literature. The annotation procedure includes the following steps:

- Retrieval of abstracts and meta-data from the acknowledged bibliographic databases (Web of Science (Clarivate Analytics), Scopus (Elsevier), and NCBI PubMed) by using dedicated search queries (*performed by a human expert*);
- Preliminary examination of the retrieved abstracts and selection of candidate articles for annotation (*performed by a human expert*);
- Acquisition of full texts of the selected publications and secondary examination (after this stage, *ca*. 10% of the initially found papers are left for further processing) (*performed by a human expert*);
- Selection of publications containing the carbohydrate or derivative structures that match the database scope (see the criteria below) (*performed by a human expert*);
- Retrieval of the relevant information from the published data (*performed by a human annotator*);
- Encoding of the information in the strict format in a text dump file (see below) (*performed by a human annotator*);
- Various error detection routines, correction of annotation errors, and tracking of errors in publications (*performed by machine means, experts in glycobiology, and information scientists*);
- Temporary upload of the resulting dump into a service shadow of the database and subsequent checking for errors detectable in the database context only (e.g. invalid internal cross-links) (*performed by machine means and expert analysis of warnings*);
- Manual validation of the annotated data by a human curator (25–100% entries are checked);
- Approval of the dump file and its merging with the main dump, which serves as a backup of the database;
- Update of the database content from the main dump (performed annually).

To match the database scope, an article must contain at least one explicit or implicit molecular structure that meets any of the following criteria:

- The structure contains at least one carbohydrate residue (except nucleic acids studied in genomic or transcriptomic context);
- The glyco moiety of the structure is established in this or previous publications with the degree of unambiguity sufficient to derive most of its monomeric composition and at least a half of its linkages, and residue configurations;
- The structure is associated with an unambiguously specified biological source (taxon), and this taxon belongs to prokaryotes, plants, fungi or protista.

The carbohydrate structure can be published explicitly (as a figure, scheme, IUPAC name, etc.) or implicitly (as a trivial name or even a free-text description by the authors). The structure is considered present in a publication if any of the following conditions is met:

- The primary structure or its conformation is elucidated;
- A motif of the structure is suggested;
- Various properties of the structure, including its biological activity, are studied;
- Synthesis or modeling of the structure is described;
- The structure is reassigned to another taxon;
- The biological role or other properties of the structure are referenced or reviewed.

The association of a given structure with a biological source (taxon) implies any of the following:

- The structure was extracted from a biological source (i.e. the structure is natural);
- The structure is a part of a larger natural molecule, and this part is discussed separately (e.g. O-glycan moiety of a glycoprotein);
- The structure is synthetic and is identical to a natural structure (or differs from it only by an aglycon);
- The structure was obtained as a sample by modification or degradation of a natural structure, e.g. as a result of the analytical procedures;
- The structure was produced outside the organism by an enzyme from this organism, and: (i) was reported elsewhere to be present in this taxon; or (ii) its precursor was reported elsewhere to be present in this taxon or to be consumed by this taxon; or (iii) its precursor was reported elsewhere to be present in the host organism infected by this taxon.

Text dump format. Except for derived content, such as oligosaccharide conformation maps, the CSDB is imported from human-readable text files called "dump files". The main CSDB dump file is manually filled and appended by a team of annotators, who perform the search for matching scientific literature and its analysis. Before import, the dumps undergo automated syntactic validation and manual data quality control by another team of curators. Data correction and content updates are performed on the main dump file.

The main CSDB dump file is a UTF-8 text backup of the database and a reference file for all the database content. The dump file contains records separated by two blank lines; the main dump contains all the CSDB records. Lines starting with the symbol # are comments for annotators and are not processed. Every record is a unique combination of a molecular structure and a paper, in which this structure is discussed. These data are appended by other annotations, such as biological context, etc. The record consists of 44–47 lines, one line per field. The line starts with the field name followed by colon (**:**), after which the field content is provided. Line breaks inside the field are not allowed. The detailed explanation of the fields is given in Online-only Table 1.

As an example, we provide a step-by-step description of the annotation procedure for one of the CSDB records (ID 4676; http://csdb.glycoscience.ru/database/core/search_id.php?id_list=4676; see Online-only Table 2). This record was added to the database upon annotating the papers on the structures of carbohydrates from the bacterial genus *Proteus* that were published in the years 1996–2000. The corresponding papers were selected via the Web of Science database. The following search query can be used in the current version of the WoS (in the Advance Search mode): **(TS = (carbohydrate* OR *saccharide) AND TS = (Proteus) AND PY = (1996–2000))**. The paper itself[32] is open-access available at the publisher web-site (https://febs.onlinelibrary.wiley.com/doi/full/10.1046/j.1432-1327.2000.01041.x).

As stated in the *Annotation rules* section, only the papers containing at least one explicit or implicit molecular structure that meets any of the above-mentioned criteria concerning the structure and its biological source are used for filling up the database. In this example, the polysaccharide structure was given in the abstract and text of the paper, and the source of this structure was unambiguously stated as *Proteus penneri* strain 63. Thus, the data from the paper were added to the CSDB database.

The annotation of this paper included the following steps:

1. A template for a new record was created in the text dump file. This template contained all the mandatory and optional fields (see Online-only Table 1). A CSDB ID was assigned to the record in accordance with the previously assigned IDs.
2. The bibliographical fields of the annotation form were completed: AU (authors), TI (title), JN (journal), PY (publication year), VL (volume and issue), PG (pages), and RL (bibliographical identifiers). The additional fields EA (corresponding author e-mail), AD (author affiliations), AB (abstract) and KW (keywords) were also completed. (See the corresponding fields in Online-only Table 2.) The data added to the text dump file were retrieved from page 601 of the paper.
3. The structure-related fields ST1 (the carbohydrate structure in the CSDB Linear encoding, according to the paper; the rules of the CSDB Linear encoding are beyond the scope of this paper paper and were published elsewhere[33]), ST2 (type of the structure; in this case – CHEM, chemical repeating unit of a polymer, because the exact polymerization frame, i.e. a biological repeating unit, was not reported in the paper), SL (structure location in the paper, in this case – abstract or Fig. 3 in the paper) and CC (compound classes/roles) were completed. The fields ST3 (polymerization degree), AG (aglycon information), MF (molecular formula), and 3D (3D structure and conformational data) were irrelevant or unknown and were therefore left empty. (See the corresponding fields in Online-only Table 2.)
4. The fields related to the biological source of the structure were completed: SO (the organism, from which the structure was extracted; in this case - *Proteus penneri* 63), KD (taxonomic domain/taxonomic phylum; in this case - bacteria/Proteobacteria), and TAX (identifier from the NCBI Taxonomy database; in this case - (102862), which refers to the Proteus penneri species; it is given in parentheses, because there is no separate record for strain 63 in the NCBI Taxonomy, but a TaxID exists for a higher rank on the tree of life, namely species). The additional field DSS (disease of the host organism associated with the structure or its biological source) was also completed in accordance with the International Classification of Diseases, version 11.
5. The fields related to the elucidation of the structure were completed: MT (methods, in accordance with the Materials and Methods section of the paper), NMRH (^{1}H NMR assignment), NMRC (^{13}C NMR assignment), NMRS (solvent, in which NMR spectra were recorded, as stated the Materials and Methods section), and NMRT (temperature, at which NMR spectra were recorded, in Kelvins, as stated in the Materials and Methods section). (See the corresponding fields in Online-only Table 2.) Templates for the fields NMRH and NMRC can be generated by using the "generate NMR template" link at the "Submit record" page in the "Extras" section at the CSDB web-site. Note that to generate an NMR template, the ST1 field must be completed.
6. In accordance with the methods given in the previous step, the TH field was completed. In this case, it contains "1", because the paper describes the elucidation of the structure being annotated.
7. Finally, the annotator and service fields were completed: U1 (annotator's last name), U2 (record submission date), U3 (ID of the paper in a local database of the Carbohydrate Chemistry Lab, N.D. Zelinsky Institute of Organic Chemistry, RAS), RR (IDs of related CSDB records), and DB (references to other structure databases; in this case – ID of the structure in the GlyTouCan database). (See the corresponding fields in Online-only Table 2.)

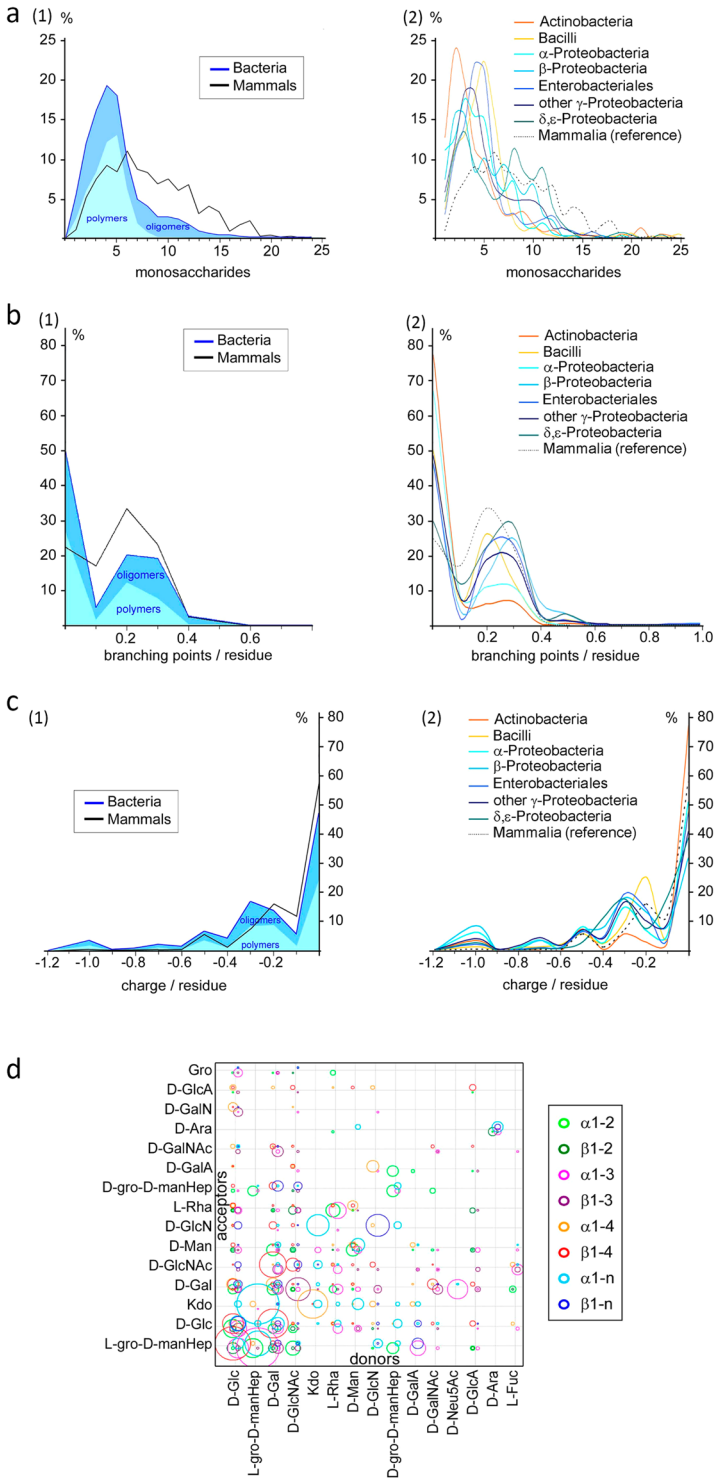

Fig. 3 Examples of analytical studies carried out earlier directly on CSDB and Glycosciences.de content. (**a**) Size distribution of carbohydrate sequence units, (**b**) branching index distribution, and (**c**) mean charge density distribution in two taxonomic domains. (**d**) Glycosidic linkage distribution in bacterial oligomeric glycans. Reproduced with permission from[36].

Note that in this record, several fields remain empty. Thus, the paper provided no information on the host organism and organ/tissue, from which the structure was extracted (fields HO and OTI), on the enzymes processing the structure (EI), on its biological activity (BA), biosynthesis (BG), and chemical synthesis (SY). There is also no trivial name of the compound (NC). Online-only Table 2 shows the final CSDB record 4676, as present in the CSDB text dump file.

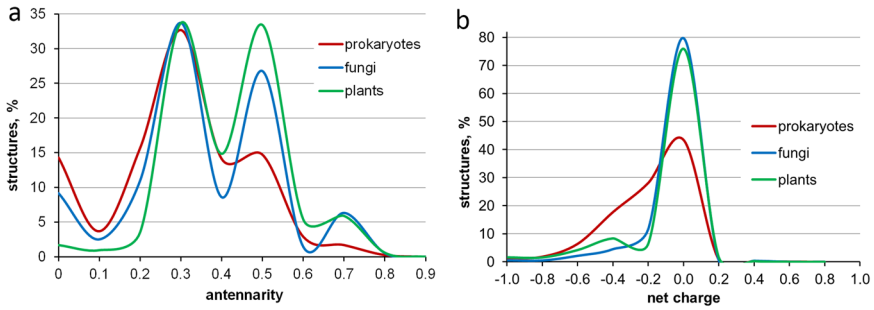

Fig. 4 Distribution of structures in CSDB according to their antennarity (**a**) and net charge (**b**) in the organisms from the three kingdoms represented in CSDB. Prokaryotes and fungi have a complete coverage on published carbohydrate structures (up to the year 2020), while plants are covered up to 1997 only.

For annotation, we use full texts of articles from publisher web sites (open-access papers), the Zelinsky Institute subscriptions and library, and requests to the authors if allowed by a publisher license. We do not provide the access to full texts themselves.

Data Records

The following data described in this paper are publicly available[34]:

Text dump for prokaryotic carbohydrates: bcsdb_2021dec06.txt;

Text dump for fungal carbohydrates: fcsdb_2021dec06.txt;

Text dump for plant carbohydrates: pcsdb_2021dec06.txt;

Self-contained CSDB backup (for import using MySQL): CSDB_2021dec23_full.sql;

Supplementary data (Tables S1 and S2) used to prepare plots in Fig. 4;

Supplementary code used for generation of data in Tables S1 and S2 (needs a running instance of the CSDB database).

The text dump files are subject to update biennially.

Technical Validation

The dump files are subject to machine error checking upon import. The quality of the CSDB data is maintained by automatic detection of *ca.* 100 types of data errors and suspicious data combinations. Manual verification of the data is performed by human experts; it allows revealing logical and factual errors that cannot be detected automatically[33]. The most widespread type of errors found in CSDB are those imported from other databases, e.g. CCSD (which, according to a retrospective analysis, contains errors in *ca.* 35% entries[35]).

According to manual error checking of the CSDB dumps, *ca.* 2000 errors imported from other databases (primarily CCSD) and *ca.* 350 errors in structures and NMR spectrum assignment in original publications were found. In the latter case, when the errors could be corrected without additional experimental studies, corrections were suggested. At that, the original erroneous structure was stored in the ST1ORIG field. This field is also provided when a particular structure is revised in later publications. Users can send an error correction request via a dedicated from (a link to this form is available for each database entry at the CSDB web site).

Usage Notes

The CSDB content has been directly used for analyzing distributions of carbohydrate structures from bacteria, protista, archaea, fungi, and plants according to various criteria. Such analysis cannot be performed by means of the CSDB web interface. For example, a comparison of bacterial and mammalian carbohydrates from the viewpoint of their characteristics and diversity, in particular, structure size, branching index and mean charge density distributions (Fig. 3a–c), formalized the differences in basic features of the carbohydrate architecture between bacteria and mammals[36]. A distribution of glycosidic linkages in oligomeric (Fig. 3d) and polymeric prokariotic glycans was also analyzed. These data, in turn, were purposed for further revealing of the immunogenic patterns of pathogenic bacteria.

For illustrative purposes, we carried out an analysis of the current distribution of carbohydrate structures in CSDB in accordance with their antennarity (Fig. 4a) and net charge (Fig. 4b). In total, *ca.* 25400 structures were considered. In this work, antennarity is the ratio of the number of non-reducing termini to the number of residues in a given structure. The net charge of a molecule is a ratio of the formal integer charge of a structural unit (an oligoglycan or a repeating unit of a polymer) to the size of the structural unit. It allows estimation of the density of charged groups (such as $-NH_3^+$, $-COO^-$, $-PO_4^{3-}$, and $-SO_4^{2-}$) in a glycan. A detailed description of the sampling and deriving of the statistic data is provided in Tables S1 and S2 (see the file *analysis_on_raw_DB.xls* in a dataset[34]). These two examples visually demonstrate differences and similarities between prokaryotic and eukaryotic organisms in terms of their carbohydrate architecture. Due to a close-to-complete coverage on published carbohydrate structures from bacteria and fungi, the presented distributions provide valid scientific information on the studied glycans from these organisms.

Of note, the Data Records include a self-contained CSDB image; however, we would like to note that the usage of this image is less flexible for utilizing the data being reported since it implies the same database format as the one already implemented in CSDB.

Code availability

The SQL and PHP code used for producing the exemplary distributions shown in Fig. 4 is provided in the file *supplementary_code_for_article.zip* and is publicly available[34].

Received: 21 September 2021; Accepted: 3 February 2022;
Published online: 30 March 2022

References

1. Egorova, K. S. & Toukach, P. V. Glycoinformatics: Bridging isolated islands in the sea of data. *Angew. Chem. Int. Ed.* **57**, 14986–14990 (2018).
2. Aoki-Kinoshita, K. F. Using databases and web resources for glycomics research. *Molecular & Cellular Proteomics* **12**, 1036–1045 (2013).
3. Abrahams, J. L. *et al*. Recent advances in glycoinformatic platforms for glycomics and glycoproteomics. *Curr. Opin. Struct. Biol.* **62**, 56–69 (2020).
4. Copoiu, L. & Malhotra, S. The current structural glycome landscape and emerging technologies. *Curr. Opin. Struct. Biol.* **62**, 132–139 (2020).
5. Li, X., Xu, Z., Hong, X., Zhang, Y. & Zou, X. Databases and bioinformatic tools for glycobiology and glycoproteomics. *Int. J. Mol. Sci.* **21**, 6727 (2020).
6. Scherbinina, S. I. & Toukach, P. V. Three-dimensional structures of carbohydrates and where to find them. *Int. J. Mol. Sci.* **21**, 7702 (2020).
7. *Glycoinformatics*. (Humana Press, 2015).
8. *A Practical Guide to Using Glycomics Databases*. (Springer, 2017).
9. Doubet, S. & Albersheim, P. Letter to the Glyco-Forum. *Glycobiology* **2**, 505–505 (1992).
10. Lütteke, T. *et al*. GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* **16**, 71R–81R (2006).
11. Böhm, M. *et al*. Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update). *Nucleic Acids Res.* **47**, D1195–D1201 (2019).
12. Campbell, M. P. *et al*. UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.* **42**, D215–D221 (2014).
13. Aoki-Kinoshita, K. F. & Kanehisa, M. In *Glycoinformatics* Vol. 1273 (eds Lutteke, T. & Frank, M.) 97–107 (Springer, 2015).
14. Maeda, M. *et al*. in *Glycoinformatics* Vol. 1273 (eds Lutteke, T. & Frank, M.) 161–179 (Springer, 2015).
15. Toukach, P. V. & Egorova, K. S. Carbohydrate Structure Database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Res.* **44**, D1229–D1236 (2016).
16. Egorova, K. S., Kondakova, A. N. & Toukach, P. V. Carbohydrate Structure Database: tools for statistical analysis of bacterial, plant and fungal glycomes. *Database* **2015**, bav073 (2015).
17. Kapaev, R. R., Egorova, K. S. & Toukach, P. V. Carbohydrate structure generalization scheme for database-driven simulation of experimental observables, such as NMR chemical shifts. *J. Chem. Inf. Model.* **54**, 2594–2611 (2014).
18. Kapaev, R. R. & Toukach, P. V. Improved carbohydrate structure generalization scheme for ^{1}H and ^{13}C NMR simulations. *Anal. Chem.* **87**, 7006–7010 (2015).
19. Kapaev, R. R. & Toukach, P. V. Simulation of 2D NMR spectra of carbohydrates using GODESS software. *J. Chem. Inf. Model.* **56**, 1100–1104 (2016).
20. Kapaev, R. R. & Toukach, P. V. GRASS: semi-automated NMR-based structure elucidation of saccharides. *Bioinformatics* **34**, 957–963 (2018).
21. Bochkov, A. Y. & Toukach, P. V. CSDB/SNFG structure editor: An online glycan builder with 2D and 3D structure visualization. *J. Chem. Inf. Model.* **61**, 4940–4948 (2021).
22. Chernyshov, I. Y. & Toukach, P. V. REStLESS: automated translation of glycan sequences from residue-based notation to SMILES and atomic coordinates. *Bioinformatics* **34**, 2679–2681 (2018).
23. Egorova, K. S. & Toukach, P. V. CSDB_GT: a new curated database on glycosyltransferases. *Glycobiology* **27**, 285–290 (2017).
24. Egorova, K. S., Knirel, Y. A. & Toukach, P. V. Expanding CSDB_GT glycosyltransferase database with *Escherichia coli*. *Glycobiology* **29**, 285–287 (2019).
25. Egorova, K. S., Smirnova, N. S. & Toukach, P. V. CSDB_GT, a curated glycosyltransferase database with close-to-full coverage on three most studied non-animal species. *Glycobiology* **2020**, cwaa107 (2020).
26. Ranzinger, R. *et al*. GlycoRDF: an ontology to standardize glycomics data in RDF. *Bioinformatics* **31**, 919–925 (2015).
27. Scherbinina, S. I., Frank M. & Toukach P. V. Carbohydrate structure database (CSDB) oligosaccharide conformation tool. *Glycobiology*, https://doi.org/10.1093/glycob/cwac011 ePub ahead of print (2022).
28. Frank, M. In *Glycoinformatics* Vol. 1273 (eds Lütteke, T. & Frank, M.) 359–377 (Humana Press, 2015).
29. Frank, M., Lutteke, T. & von der Lieth, C. W. GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res.* **35**, 287–290 (2007).
30. Matsubara, M., Aoki-Kinoshita, K. F., Aoki, N. P., Yamada, I. & Narimatsu, H. WURCS 2.0 update To encapsulate ambiguous carbohydrate structures. *J. Chem. Inf. Model.* **57**, 632–637 (2017).
31. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
32. Shashkov, A. S. *et al*. Structure of a 2-aminoethyl phosphate-containing O-specific polysaccharide of *Proteus penneri* 63 from a new serogroup O68. *Eur. J. Biochem.* **267**, 601–605 (2000).
33. Toukach, P. V. & Egorova, K. S. New features of Carbohydrate Structure Database notation (CSDB Linear), as compared to other carbohydrate notations. *J. Chem. Inf. Model.* **60**, 1276–1289 (2020).
34. Toukach, P. V. & Egorova, K. S. Source files of the Carbohydrate Structure Database: the way to sophisticated analysis of natural glycans. *OSF* https://doi.org/10.17605/OSF.IO/P6DHG (2021).
35. Egorova, K. S. & Toukach, P. V. Critical analysis of CCSD data quality. *J. Chem. Inf. Model.* **52**, 2812–2814 (2012).
36. Herget, S. *et al*. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct. Biol.* **8**, 35 (2008).
37. Toukach, P. V. & Egorova, K. S. In *Glycoinformatics* Vol. 1273 (eds Lütteke, T. & Frank, M.) 55–85 (Humana Press, 2015).
38. Wiggins, E. V. The NLM current catalog. *Bull. Med. Libr. Assoc.* **57**, 36–40 (1969).
39. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2011).
40. The Lancet Infectious, D. ICD-11: in praise of good data. *Lancet Infect. Dis.* **18**, 813 (2018).

Acknowledgements

This work was funded by Russian Science Foundation grant 18-14-00098-P.

Author contributions

P.V.T. elaborated the CSDB concept and design, programmed and supervised the whole project. K.S.E. designed the CSDB_GT concept and annotated and verified the data. Both authors prepared the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.V.T. or K.S.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.