JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

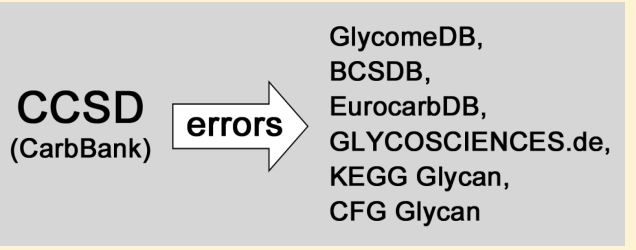
Critical Analysis of CCSD Data Quality

K. S. Egorova and Ph. V. Toukach*

N.D. Zelinsky Institute of Organic Chemistry, Leninsky prospekt 47, 119991 Moscow, Russian Federation

Supporting Information

ABSTRACT: Systematization and classification of carbohydrates contribute greatly to development of modern biomedical sciences. CCSD (CarbBank) data constitute the significant part of nearly all existing carbohydrate databases. However, these data have not been verified from their original deposit. During the expansion of Bacterial Carbohydrate Structure Database (BCSDB) project, we checked CCSD data quality and found that about 35% of records contained errors. The CCSD data cannot be used without manual verification, while CCSD errors migrate from database to database.



INTRODUCTION

The role of natural carbohydrates in cellular and molecular biology can hardly be overestimated.[1] They are involved in the pathology of cancer,[2] HIV,[3] and almost all bacterial infections.[4] Glycans are used as diagnostic and therapeutic targets,[5] therapeutic agents,[6,7] and vaccines,[8] whereas synthetic saccharides are employed as probes in biological research.[9] Glycomics aims at systematization and classification of known carbohydrates, as well as elucidation of their role in human health and disease.[10] The amount of information on natural carbohydrates accumulates rapidly; therefore, freely available and regularly updated databases are demanded. To date, several carbohydrate databases have been developed: CCSD (CarbBank, ceased in 1996),[11,12] GLYCOSCIENCES.de,[13] GlycoSuiteDB,[14] Consortium of Functional Glycomics (CFG) Glycan Database,[15] KEGG-Glycan (being a part of Kyoto Encyclopedia of Genes and Genomes),[16] Bacterial Carbohydrate Structure database (BCSDB),[17] GlycoBase (Dublin),[18] GlycoBase (Lille),[19] ECODAB,[20] Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB), design study EurocarbDB,[21] meta-database GlycomeDB,[22] and others. Unlike to CCSD before 1995, none of these projects, except BCSDB, claims to provide complete coverage of published data. All carbohydrate databases almost completely lack fungal and plant structures published after 1996, whereas erroneous entries initially deposited in CCSD have migrated through many databases only a few of which possess efficient error control.[17]

The Complex Carbohydrate Structure Database (CCSD, The CarbBank Project) was developed by Complex Carbohydrate Research Center of the University of Georgia (United States) as a computerized database for cataloging all published complex carbohydrates structures of oligosaccharides and glycoconjugates with three or more glycosyl residues, excluding synthetic intermediates.[12] Each CCSD entry includes the primary structure of a carbohydrate, the citation, taxonomy, and other information. Since the database contained many complex structures with aglycon substituents, a software called Carbbank was developed to manage the CCSD file.[12] CCSD collected 14 887 carbohydrate structures in approximately 50 000 entries,[17] which presented almost full coverage up to 1995. Although funding of CCSD stopped in 1996, its data are utilized in nearly all active projects, including CFG Glycan (O- and N-glycans from CCSD),[15] GlycoSCIENCES.de,[13] KEGG Glycan,[16] EurocarbDB,[21] GlycomeDB,[22] and BCSDB (bacterial glycans).[17] Due to this, conventions introduced by CCSD, such as structure encoding, residue naming, types and format of data, etc., are still important.

Recently we have started a new project, the main goal of which is to expand BCSDB, the Bacterial Carbohydrate Structure Database, devised for provision of structural, bibliographic, taxonomic, NMR spectroscopic, and other related information on bacterial carbohydrate structures. Up to 2011, the BCSDB had included structures found in bacteria, or their derivatives (bacteria-associated structures from CCSD plus structures manually extracted from publications indexed in NCBI PubMed).[17] In 2012−2014 BCSDB is to be expanded to the Carbohydrate Structure Database (CSDB) containing not only bacterial but also fungal and plant carbohydrates.

RESULTS AND DISCUSSION

The first step of the BCSDB expansion was the import of the fungal and plant part of CCSD (see the Experimental Section for details). In the process of collecting data on plant and fungal carbohydrates deposited in CCSD, we met the necessity to check the CCSD entries against the original publications. Therefore, the following record fields were verified: the primary structure, the biological source of a carbohydrate, and the presence of a given structure in the corresponding publication. Here, 857 and 4281 entries were obtained from CCSD for fungal and plant carbohydrates, correspondingly. Of them, we have checked 857 fungal and 464 plant entries, and additionally

Received: June 18, 2012

   dx.doi.org/10.1021/ci3002815 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Errors Found in CCSD Entries for Fungal, Bacterial, and Plant Carbohydrates[a]

taxonomical domain	fungi			bacteria			plants		
total entries in CCSD for this domain	absolute	% checked	example[b]	absolute	% checked	example[b]	absolute	% checked	example[b]
total entries processed	857			301			464		
entries with checked publications	498			223			263		
entries with errors	299	60.0		148	29.7		86	17.3	
incorrect strain	186	37.3	10711	77	15.5	39779	10	2.0	1126
structure not found	14	2.8	7020	9	1.8	34238	1	0.2	10187
incorrect structure	18	3.6	27713	30	6.0	35	21	4.2	11253
supposed structure (elucidated elsewhere)	11 + 7[c]	2.2 + 1.4	47089	6	1.2	50002	21	4.2	11181
incorrect organism	19	3.8	49943	26	5.2	23513	44	8.8	10742
structures missing from CCSD (missing entries/number of checked papers)	60/107			33/68			191/68		

[a]"Absolute" is the number of records, and "% checked" is this number related to the total number of records checked against original publications, in percent. [b]One CCSD ID per error type is provided as an example for each domain. Complete information on the found CCSD errors can be retrieved from field U5 in a BCSDB dump. [c]Seven fungal entries have not been published as fully determined structures, and some structural properties were assumed in the CCSD notation.

301 bacterial entries selected randomly from a bacterial CCSD subdump file (5657 records). The found errors were divided into six major categories:

- incorrect strain (the strain designation is present in the paper but absent or wrong in CCSD);
- structure not found (the structure cannot be found in the cited paper);
- incorrect structure (the structure from CCSD differs from that in the paper or is not as complete as in the paper);
- supposed structure (elucidated elsewhere) (the paper contains only a trivial name or an abbreviated symbolic formula of a carbohydrate, whereas its primary structure has been elucidated elsewhere, or the structure deposited in CCSD is elucidated only partially in the paper);
- incorrect organism (the organism is assigned incorrectly in CCSD);
- missing structure (the structure is present in the paper but missing from CCSD; frequency of this error was not calculated, and the number of such entries is not included into the total number of the processed entries).

Errors in the CCSD fields other than structural and taxonomic ones (bibliography, methods, cross-references, etc.) have not been checked within this study. The cumulative error frequency data are shown in Table 1. In total, we found errors in 60% of fungal, 30% of bacterial, and 17% of plant entries checked. In the case of bacterial and fungal carbohydrates, the most frequent error found in 15.5% and 37.3% of entries, respectively, was incorrect strain assignment, whereas in the case of plant carbohydrates, the most frequent error (8.8% of entries) was incorrect organism assignment.

Certain entries also contained erroneous carbohydrate structures (3.6% of fungal, 6.0% of bacterial, and 4.2% of plant carbohydrates). The entries with incorrectly recorded structures, which could nevertheless be correctly deduced using a scientific common sense, have not been included in the calculation. Most of these cases were improper naming of residues and their absolute configurations, e.g. "2-deoxy-D-glucopyranose" instead of "2-deoxy-D-arabino-hexopyranose".

Moreover, a significant number of structures present in the cited papers was missing from CCSD: 60 structures from 107 papers, 33 structures from 69 papers, and 191 structures from 68 papers for fungal, bacterial, and plant domains, respectively. About 10% of these missing entries comprised disaccharides which were deliberately not included into CCSD.[12] However, it should be noted that CCSD still contains some mono- and disaccharide entries. Several of the checked entries have not been originally published in the cited papers as fully determined structures, and some structural properties were assumed in the CCSD notation (error type "supposed structure, elucidated elsewhere"). This is rather bibliographic inconsistency than an error, since the structures in CCSD are correct but cannot be tracked.

It should be noted that CCSD records related to mammalian glycans, which were not included in our study, are expected to contain less errors as they possess lower structural diversity and usually lack thorough taxonomical annotations, such as strains or serogroups.

CONCLUSION

It must be stated that the CCSD data cannot be used and referenced without verification. Regretfully, these types of errors can be neither detected nor corrected automatically, except for a few special cases of "incorrect structure" errors, and the verification process requires an expert analysis of original publications. As for autodetectable and autocorrectable errors, to date only GlycomeDB and BCSDB include automatic verification mechanisms applied to all deposited carbohydrate structures,[17,22] whereas BCSDB is the only carbohydrate database with manually verified structural, bibliographic, and taxonomic annotations.

We suggest that the existing carbohydrate databases should either undergo retrospective publication analysis by human experts or integrate contents of the databases for which such verification has been performed. Inconsistent taxonomical annotations, the most frequent error found in CCSD records, seriously hamper the outcome of the databases related to microorganisms, as one of prevailing types of user queries is a search for structures specific for a certain serogroup (data from statistical analysis of requests to BCSDB). Therefore, a mutual database integration and development of data exchange standards are of extreme importance.[23] Glycoscientists, in turn, should always check database-retrieved data originally published before 1996, since these entries were most likely imported from CCSD.

dx.doi.org/10.1021/ci3002815 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

The CCSD data corrected and/or approved within this study are exportable as a part of the CSDB dump using the "Export" feature at the CSDB Web site, and available as an RDF feed produced by the CSDB engine (see online CSDB documentation for details).

EXPERIMENTAL SECTION

Data were processed using scripts written in PHP5. Interaction with GlycomeDB utilized PostgreSQL 8.1 requests to a local copy of the database provided by the developers. Interaction with the NCBI Taxonomy database utilized "E-utilities" web-services provided online by NCBI. Statistical processing and derivation of the cumulative data were performed in Microsoft Excel 2010.

The CCSD dump file was split into subdumps corresponding to different taxonomic domains (bacteria, plant, fungi, animals, other metazoa). The splitting routine was realized as a PHP script querying GlycomeDB and NCBI taxonomy and producing separated flat subdumps. This script utilized taxonomic annotations retrieved for CCSD-derived structures deposited in GlycomeDB[22] and information on the relationships between taxons deposited in NCBI Taxonomy database.[24] The structures from CCSD not deposited in GlycomeDB have not been processed, since their absence suggests that the carbohydrate moiety could not be parsed due to erroneous or ambiguous notation. Therefore, the calculated percentage of errors in the corresponding taxonomical domains of Carbbank can be underestimated but not overestimated. To exclude errors possibly introduced by the GlycomeDB engine, we used a flat CCSD dump file downloaded from the GlycomeDB Web site, for data retrieval, while the GlycomeDB-dependent splitting routine was utilized to generate the list of CCSD IDs only.

ASSOCIATED CONTENT

Supporting Information

GlycomeDB-based CCSD dump splitting routine. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: netbox@toukach.ru.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was performed in the framework of development of Plant and Fungal Carbohydrate Structure Database funded by Russian Foundation for Basic Research, grant N 12-04-00324.

REFERENCES

(1) Jones, C. J.; Larive, C. K. Carbohydrates: cracking the glycan sequence code. *Nature Chem. Biol.* **2011**, *7* (11), 758−759.

(2) Slovin, S. F; Ragupathi, G.; Fernandez, C.; Jefferson, M. P.; Diani, M.; Wilton, A. S.; Powell, S.; Spassova, M.; Reis, C.; Clausen, H.; Danishefsky, S.; Livingston, P.; Scher, H. I. A bivalent conjugate vaccine in the treatment of biochemically relapsed prostate cancer: a study of glycosylated MUC-2-KLH and Globo H-KLH conjugate vaccines given with the new semi-synthetic saponin immunological adjuvant GPI-0100 OR QS-21. *Vaccine* **2005**, *23*, 3114−3122.

(3) Pantophlet, R.; Wilson, I. A.; Burton, D. R. Hyperglycosylated mutants of human immunodeficiency virus (HIV) type 1 monomeric gp120 as novel antigens for HIV vaccine design. *J. Virol.* **2003**, *77*, 5889−5901.

(4) Jones, C. Vaccines based on the cell surface carbohydrates of pathogenic bacteria. *An. Acad. Bras. Cienc.* **2005**, *77*, 293−324.

(5) Gornik, O.; Dumić, J.; Flogel, M.; Lauc, G. Glycoscience - a new frontier in rational drug design. *Acta Pharm.* **2006**, *56*, 19−30.

(6) Shriver, Z.; Raguram, S.; Sasisekharan, K. Glycomics: a pathway to a class of new and improved therapeutics. *Nat. Rev. Drug Discovery* **2004**, *3*, 863−873.

(7) Ernst, B.; Magnani, J. L. From carbohydrate leads to glycomimetic drugs. *Nat. Rev. Drug Discovery* **2009**, *8*, 661−677.

(8) Astronomo, R. D.; Burton, D. R. Carbohydrate vaccines: developing sweet solutions to sticky situations? *Nat. Rev. Drug Discovery* **2010**, *9*, 308−324.

(9) Boltje, T. J.; Buskas, T.; Boons, G.-J. Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. *Nat. Chem.* **2009**, *1*, 611−622.

(10) von der Lieth, C.-W.; Lütteke, T.; Frank, M. The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochim. Biophys. Acta* **2006**, *1760*, 568−577.

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

(12) Doubet, S.; Albersheim, P. Carbbank. *Glycobiology* **1992**, *2*, 505.

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

(14) Cooper, C. A.; Joshi, H. J.; Harrison, M. J.; Wilkins, M. R.; Packer, N. H. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.* **2003**, *31*, 511−513.

(15) Raman, R.; Venkataraman, M.; Ramakrishnan, S.; Lang, W.; Raguram, S.; Sasisekharan, R. Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology* **2006**, *16*, 82R−90R.

(16) Hashimoto, K.; Goto, S.; Kawano, S.; Aoki-Kinoshita, K. F.; Ueda, N.; Hamajima, M.; Kawasaki, T.; Kanehisa, M. KEGG as a glycome informatics resource. *Glycobiology* **2006**, *16*, 63R−70R.

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

(18) Campbell, M. P.; Royle, L.; Radcliffe, C. M.; Dwek, R. A.; Rudd, P. M. GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* **2008**, *24*, 1214−1216.

(19) Maes, E.; Bonachera, F.; Strecker, G.; Guerardel, Y. SOACS index: an easy NMR-based query for glycan retrieval. *Carbohydr. Res.* **2009**, *344*, 322−330.

(20) Stenutz, R.; Weintraub, A.; Widmalm, G. The structures of *Escherichia coli* O-polysaccharide antigens. *FEMS Microbiol. Rev.* **2006**, *30*, 382−403.

(21) von der Lieth, C.-W.; Freir,e, A. A.; Blank, D.; Campbell, M. P.; Ceroni, A.; Damerell, D. R.; Dell, A.; Dwek, R. A.; Ernst, B.; Fogh, R.; Frank, M.; Geyer, H.; Geyer, R.; Harrison, M. J.; Henrick, K.; Herget, S.; Hull, W. E.; Ionides, J.; Joshi, H. J.; Kamerling, J. P.; Leeflang, B. R.; Lütteke, T.; Lundborg, M.; Maass, K.; Merry, A.; Ranzinger, R.; Rosen, J.; Royle, L.; Rudd, P. M.; Schloissnig, S.; Stenutz, R.; Vranken, W. F.; Widmalm, G.; Haslam, S. M. EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology* **2011**, *21*, 493−502.

(22) Ranzinger, R.; Herget, S.; von der Lieth, C.-W.; Frank, M. GlycomeDB − a unified database for carbohydrate structures. *Nucleic Acids Res.* **2011**, *39*, D373−376.

(23) Toukach, P.; Joshi, H. J.; Ranzinger, R.; Knirel, Y.; von der Lieth, C. W. Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res.* **2007**, *35*, D280−D286.

(24) Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **2012**, *40*, D136−D143.