Bacterial Carbohydrate Structure Database 3: Principles and Realization

Philip V. Toukach

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

Received April 18, 2010

Bacterial carbohydrate structure database (BCSDB) is an open-access project that collects primary publication data on carbohydrate structures originating from bacteria, their biological properties, bibliographic and taxonomic annotations, NMR spectra, etc. Almost complete coverage and outstanding data consistency are achieved. BCSDB version 3 and the principles lying behind it, including glycan description language, are reported.

INTRODUCTION

Bacterial Glycoinformatics. Natural carbohydrates play an important role in living systems by regulating life processes and information transfer between cells. Glycans take part in the pathology of infectious diseases and cancer and are used in diagnostics and therapy. In contrast to genomics and proteomics, universal computer-assisted tools in glycomics are still in the making. Whereas the amount of information on natural carbohydrates increases rapidly, scientists in glycomics lack databases providing a full variety of services.

Glycomics has been aimed at the systematization and the classification of all known carbohydrates and the revelation of their influence on human health and pathology.[1,2] To solve this task, a huge amount of structure- and property-related data has been acquired, especially after the development of instrumental analytical methods, such as NMR and mass spectrometry.[3,4] The main task of glycomics nowadays is processing the data already obtained experimentally. According to the taxonomical annotation search in GlycomeDB metadatabase[5] this implies more than 15 000 distinct natural structures, many of which were published several times in association with various properties or organisms. Freely available and regularly updated databases are demanded to navigate in this ocean of information.

The first project of this sort was the complex carbohydrate structure database (CCSDB, CarbBank),[6,7] which gathered carbohydrate-related data published before 1995. Since maintenance of this database was ceased, there has been at least a three-fold growth of information in natural glycomics, which has not been reflected in the CCSDB. To close this gap, a number of projects have been developed (see below), however most of them were oriented to mammalian and synthetic glycans. At the same time, antigens of many pathogenic microorganisms have carbohydrate origin, and recognition of bacteria by the host immune system is determined by the structure of these compounds. Immunological role of bacterial carbohydrates can hardly be overestimated and is confirmed by the appearance of numerous carbohydrate-based vaccines in the recent decade.[8]

It must be noted that collecting structural information alone is not enough to build a useful tool for glycome researchers. Taxonomical annotations are required for studies on the distribution of structural peculiarities of carbohydrates among various classes of pathogens.[9] Besides their medical applications, these data open the way to creating a glycan pool for the automatic synthesis of glycans limited to certain biological groups.[10] "Design studies related to the development of distributed, Web-based European carbohydrate databases (EuroCarbDB)" were recently performed.[11] They demonstrated that taxonomic, bibliographic, and other annotations as well as NMR and MS experimental data are desirable for a successful carbohydrate database. However, the results of the design study have not been implemented in a universal database.

Other Carbohydrate Databases. This section gives a brief overview of existing projects aimed at provision of similar services as the reported database:

- CCSDB (CarbBank)[6,7] was developed by Complex Carbohydrate Research Center of the University of Georgia (United States) and collected 14 887 carbohydrate structures in approximately 50 000 records, which presented almost full coverage up to 1995. Although funding of this project stopped in 1996, CCSDB data are included in nearly all active projects, and the conventions introduced by CCSDB are still important.
- GLYCOSCIENCES.de portal[12] includes most of the sequences from CCSDB and Sugabase[13] (CCSDB entries with NMR data). Additionally, the database has been updated with manually selected structures and NMR spectra.
- GlycoSuiteDB[14] aims to cover all published, structurally complete mammalian O- and N-glycan structures.
- Consortium of Functional Glycomics (CFG) Glycan Database[15] was built up from mammalian structures in CCSDB and curated structures obtained from a private database developed by Glycominds Ltd., Lod, Israel.
- Kyoto Encyclopedia of Genes and Genomes runs the KEGG-Glycan[16] database built on sequences from the CCSDB and their own curation efforts, containing ca. 11 000 structures.
- GlycoBase (Dublin)[17] contains around 350 N-linked glycan structures elucidated by mass spectrometry.

Address correspondence to never@vilvit.net.

10.1021/ci100150d
Published on Web 12/14/2010

- The GlycoBase (Lille) contains carbohydrate sequences from amphibia species, annotation data, and associated experimental NMR data, which can be accessed via SOACS and SOACS-ol indexes.[18]
- ECODAB[19] provides O-antigen structural information for *E. coli* strains.

None of the listed projects provides access to all types of data, and none of them provides complete coverage. Only six initiatives have taxonomical annotations assigned (CCSDB, GLYCOSCIENCES.de, CFG, KEGG, GlycoBase (Lille), and GlycoSuiteDB), and these annotations usually lack strain information. According to our investigation, a lot of CCSDB data deposited in various databases, including CCSDB itself, contain errors (see Error Tracking Section for details). Bacterial, fungal, and plant carbohydrate structures published after 1996 are almost completely missing from all projects.

To provide access to the data from all these projects within a single user interface, a metadatabase Glycome-DB[5] has been developed by the German Cancer Research Center. Glycome-DB does not collect or store its own data but imports it from original databases, performs limited error checking, links data together, and reports errors to the original database holders. However, it lacks bibliographic references, which causes difficulties in its direct usage in scientific research.

Bacterial Carbohydrate Structure Database. Bacterial Carbohydrate Structure Database (BCSDB) is aimed at provision of structural, bibliographic, taxonomic, NMR spectroscopic, and other related information on bacterial carbohydrate structures. Three key points distinguishing this service from the other carbohydrate database projects are: (1) better coverage (above 90% in the scope of bacterial carbohydrates published up to 2009; this means that a *negative* search answer is still valuable scientific information); (2) higher data consistence achieved by manual data verification (above 90% of error-free records, accordingly to manual curation); and (3) presence of manually curated bibliographic, NMR spectroscopic, and taxonomic annotations.

The coverage value of 90% was related to the raw estimation of the total number of bacterial carbohydrate structures reported. This number includes bacteria-associated structures in CCSDB plus structures manually extracted from publications indexed in NCBI PubMed, which have a carbohydrate-related term and a bacterial taxon name in keywords, title, or abstract.

BCSDB includes structures that have been found in bacteria or obtained by modification of those found in bacteria. In this project, "carbohydrate" means a structure composed of any residues linked by glycosidic, ester, amide, phospho- or sulphodiester, and other bonds, in which at least one residue is a sugar or its derivative. BCSDB is freely available on the Internet as a Web service at http://www.glyco.ac.ru/bcsdb3/. In this paper we report version 3 of BCSDB and the principles lying behind it, including glycan description language.

Glycan Description Languages. The exchange of data between glyco-related databases and their efficient cross-referencing is seriously hampered by the lack of generally accepted data exchange formats and structure description standards.[20] Whereas medically oriented scientists tend to use descriptions like composition and symbolic representations, biochemically oriented groups prefer a residue-based description (mainly one of the various IUPAC depictions).

Carbohydrate sequences contain special informatic challenges caused by the property of branching and the very high diversity of monomers and their chemical modifications.[21] They can be probably best described in computational terms as graphs with the monomer residues as nodes and the linkages as edges. Since glycosidic bonds display a preferred direction, these graphs are directed. The existence of potential multiple connections between two nodes degenerates the graph to a multigraph. The cyclization of carbohydrate structures leads to cyclic graphs. To avoid combinatorial expansion of identical substructures, repeating units of polysaccharides are frequently encoded as special entities. Limited analytical techniques resulting in partial structure elucidation produce uncertainties in sequences. Some residues or their secondary modifications are present nonstoichiometrically. Partial structure elucidation results in alternative residue names and superclasses at certain nodes.

The mentioned peculiarities define the demands, which a glycan description language should follow:

- The ability to describe all features that can be present in carbohydrate molecules, including uncertain and incompletely defined structural moieties. All existing languages cope well with description of widespread structures and differ in how many special cases they can treat. *Completeness* of the language (a property reflecting how many different structural peculiarities of glycans a language can handle) is often limited by a monomer name vocabulary.
- A glycan description should have a bijection with the structure; it should be unambiguous and computer parseable. IUPAC-based languages lack this property.
- Since all raw data and all databases contain errors, the structural description should be readable by humans, otherwise it would produce multiple problems in error-tracking that are very difficult in detection and correction. Languages based on a connection table approach usually lack this property and need special visualization and tracking procedures.

Table 1 summarizes existing glycan description languages as well as the BCSDB glycan description language developed within this work (*emphasized*) to fit all the mentioned conditions. We evaluated the listed parameters using four marks: poor, acceptable, good, and best. *Pseudographics* mean text residue names are arranged in multiple lines to reflect the topology. These names imply application of suffixes, prefixes, and other special rules to describe the chemical structure. *Tree* is a residue connection schema (a graph), which has a root node and utilizes a parent−child relationship between nodes (every nonroot node must have a single parent; every node may or may not have multiple children). Such a tree can be recorded as text or XML using rules that vary in different projects. Some rare structures, e.g., cyclic oligomers, cannot be encoded within this approach without the introduction of the additional rules. *Connection table* is a data structure describing connectivity between all nodes (to which other residues every residue is connected and how). This approach has no topological limitations.

Table 1. Glycan Description Languages

language, project	approach	completeness	unambiguity	human readability	parseability	uncertain descriptions
IUPAC[22]	tree (text)	poor	acceptable	good	poor	acceptable
IUPAC extended - CCSDB	pseudographics	acceptable[a]	acceptable	best	poor	acceptable
GLYDE 1[23]	tree (XML)	good	good	poor	best	acceptable
CabosML[24]	tree (XML)	?	good	poor	best	poor
GlycoCT - Glycome-DB,[25] GlydeII[b]	connection table	best (carbohydrate residues only)	best	poor	best	good
LinearCode™ - CFG[26]	tree (text)	poor	good	acceptable	best	good
LINUCS - GLYCOSCIENCES.de[27]	tree (text)	poor	best	good	best	good
KFC - KEGG-Glycan[28]	connection table	acceptable	best	poor	best	acceptable
BCSDB	**tree (text)**	**good**	**best**	**good**	**best**	**good**

[a] Although, in principle, the CCSDB rules are able to provide a unique description of a monosaccharide, unfortunately, they have not been consistently applied to all CCSDB entries. [b] GlydeII, which is similar to GlycoCT, has been agreed by scientists at the Deutsches Krebsforschungszentrum (DKFZ), the Consortium for Functional Glycomics (CFG), the Kyoto Encyclopedia for Genes and Genomes (KEGG), and the Complex Carbohydrate Research Center (CCRC) as carbohydrate sequence exchange format,[29] however its specification has not been published.

RESULTS AND DISCUSSION

Glycan Description Language. The BCSDB glycan description language (see Figure 1) developed within this project is based on a tree encoded as a single text line. It utilizes unambiguous and easily parseable but nevertheless human-readable encoding with controlled vocabulary of monomer names. The language is capable of describing any polymeric or oligomeric structures built of residues interlinked by glycosidic, amide, phospho- and sulphodiester, and other bonds. Multiple linkages between residues, nonstoichiometric entities, superclasses, alternative substructures, and other structural uncertainties are supported. The limitations of the language include the maximal nesting level of alternative side chains (= 2) and the unsupported combination of repeating and nonrepeating units within a single molecule.

If an aglycone name is beyond the monomer name vocabulary, it is stored in a separate field in the database. The type of repeating unit (oligomer, monomer, homopolymer, chemical repeating unit, biological repeating unit, and repeating unit of a cyclic structure) and the number of repeating units in a polymer also utilize separate fields.

This section provides a brief declaration of BCSDB glycan description language features and syntax. For more detailed specification, please refer to http://www.glyco.ac.ru/bcsdb3/help/rules.html. Knowledge of this language is desirable but not required for usage of the database Web front-end.

Topology. Residues are described by the sequence of terms like `<residue name>(<outlink>-<inlink>)`, where `outlink` denotes a position (carbon number) by which this residue substitutes another residue (usually 1 or 2), and `inlink` denotes to which position the linked residue is substituted. Both `outlink` and `inlink` can be presented by question mark (?) if unknown. In the case of a reducing end residue, the expression in parentheses is not needed, e.g., `A(1-3)B(1-4)C`. Here and below, the uppercase letters stand for residue names. If the structure is polymeric, the leftmost and rightmost residues should have open linkages, e.g., `-2)A(1-3)B(1-4)C(1-`.

If there are branching points, one chain is always considered the main one, and others are the side chains. The side chains are enclosed in square brackets together with the parentheses indicating their linkage, e.g., `A(1-3)[B(1-4)]C`. Several side chains attached to one residue are separated by commas. Side chains may also be linear or branched, and all combinations of nesting square brackets are allowed (see example on Figure 1B).

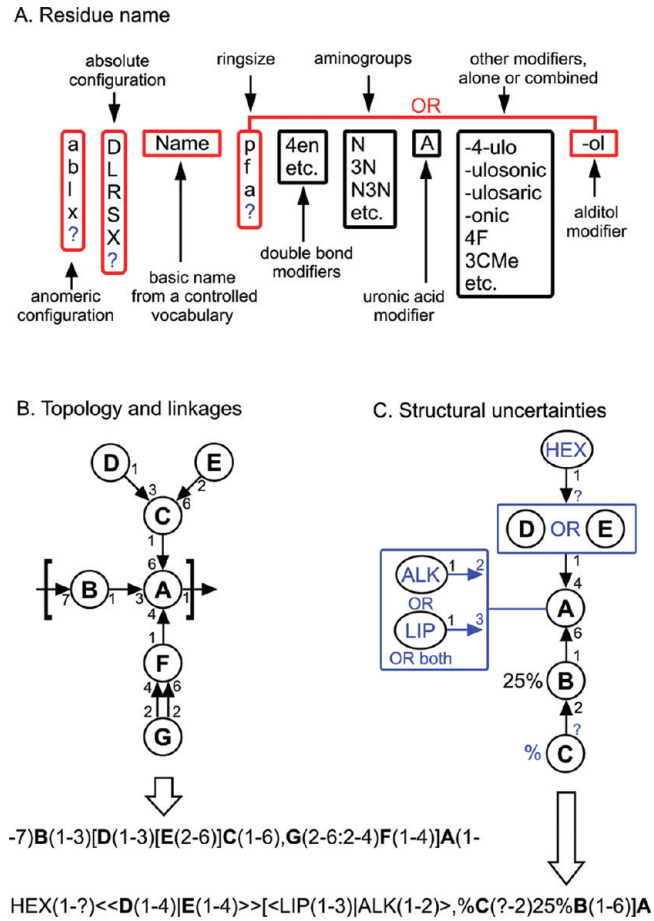

Figure 1. Basic features of glycan description language. Residues are represented by capital letters in bold. Underdetermined structural entities are given in blue. (A) Residue name components. Obligatory components are in red squares. (B) Example of topology and linkage encoding: residues **A** and **B** form the polymer backbone; residue **A** and **C** are branching points; residues **E**, **D**, and **G** are terminal; and residue **G** is dually linked to residue **F**. (C) Encoding example of an underdetermined or uncertain structure: unknown hexose is linked to an unknown position of either residue **D** or residue **E**, which is (1−4)-linked to residue **A**. In 25% of molecules, residue **A** is (1−6) substituted by residue **B**, which is partially (in unknown part of molecules) substituted at position 2 by an unknown position of residue **C**; residue **A** is substituted by alkyl at position 2 or by acyl at position 3 or by both of them.

If a residue forms multiple outgoing linkages to its acceptor (pyruvates, biphosphates, etc.), then a colon is used to separate the linkages, e.g., xRPyr(2−6:2−4)aDGal. The higher position of the acceptor always goes first. Biphosphates and bisulphates have 0 for their <outlink> index.

Names of Residues. Each residue name is composed of several fields following each other without separators (Figure 1A):

- Anomeric configuration (a = alpha, b = beta, l = a lipid residue, x = this residue has no anomeric forms or is a mixture of anomers, and ? = unknown). Monovalent residues do not require this field.
- Absolute configuration (D, L, R, S; X = this residue is not optically active, and ? = unknown). X is specified if absolute configuration is implied by a residue name (e.g., Kdo is always D) or if it is a part of a residue name (e.g., LDManHep for L-gro-D-mannoheptose already encodes absolute configuration within its name). Monovalent residues do not require this field.
- Residue name, including deoxygenation information if any. Together with subsequent modifiers (except the ring size), these names form a controlled vocabulary of 339 monomers that is updated when new residues are published. The currently included residues are monosaccharides, amino acids, nucleotides, fatty acids, alcohols, and others. Atomic descriptions of residues are stored in the database and used for error tracking.
- Ring-size modifier (p = pyranose, f = furanose, a = open chain, and ? = unknown or in any form).
- Double-bond modifier (Xen, where X is less carbon number of those forming a double bond), if present.
- Aminogroup modifiers (0 or more uppercase N characters). If the position of aminogroup is other than 2, it should be specified before N, e.g., aLRha4N.
- Uppercase A is used if a residue is an uronic acid (the last carbon forms a carboxyl group).
- All other modifiers (-ulosonic, -ulosaric, -onic, -X-ulo, XF, XCMe, XHMe, etc., where X is a modifier position) in alphanumeric order, if any. Multiple modifiers are allowed, e.g., aLDmanHepp4enN3NA6F.
- The -ol modifier is used if a residue is an alditol (which is not implied by residue basic name as in, e.g., Gro). This modifier is incompatible with the ring size and with some other modifiers.

Examples: aD6dTalA, aXKdo, xLGro, ?DManN-ol, Ac, xRPyr, and bDFucN3N. The complete monomer namespace is documented at http://www.glyco.ac.ru/residues.php.

Lipid Base Names. The naming system for lipid residues matches the general naming system described above; l should be used for anomeric configuration. For most lipids, there are reserved names like Pam, Ole, Vac, etc. If there is no reserved name, then the following rules may be used to construct a new term to become a candidate for uploading to the monomeric name vocabulary:

- The skeleton of the name is C<n>, e.g., C17, where <n> is the total number of carbons in all subchains of the residue.
- The skeleton may be appended with prefixes and postfixes in alphanumeric order, saving that hydroxy-prefix should be the closest to the skeleton.
- The postfix for double bond is the following: = {<comma-separated ascending list of double-bond positions>}, e.g., C17 = {9,11}. If the double-bond configuration is known, it may be specified as t for E-configuration (trans-) and c for Z-configuration (cis-), e.g., C24 = {t9,c11,17}.
- The postfix for carbon cyclization is the following: c{<comma-separated cyclization positions>}, e.g., C17c{9,11}.
- The hydroxy functions are encoded with the <n>HO prefix, where <n> is the attachment position. If there are several hydroxy functions, attachment positions are separated with commas, e.g. 3,15HOC17.
- The branching sites are indicated by one or more of the following prefixes: i for iso-branching (at prelast carbon), ai for anteiso-branching (at prepre-last carbon), b for branching at unknown position, and <X>b<Y> for other types of branching, where X is the branching position, and Y is the length of the side subchain, e.g., 13b1 means methyl group at C13. The number after capital C still remains the total number of carbons, including side subchains.

Structural Uncertainties. Unknown anomeric or absolute configurations and ring sizes are encoded at the level of a residue name (see above). For unknown linkage positions, a question mark is used, e.g., Subst1(?−?)bLFucp or xXEtN(1−P−?)aXKdop. For residues that are not fully elucidated, superclasses are used (see below).

Fuzziness on the topological level is described by one of two syntactic constructions: <<A(n−m)|B(p−q)>> for an exclusive combination (logical XOR) and <A(n−m)|B(p−q)> for an inclusive combination (logical OR). For example, <<A(1−3)|B(1−4)>>C is a disaccharide, in which either C3 of the residue C is substituted by the residue A or C4 of the residue C is substituted by the residue B but not both at once.

A residue in angle brackets can be substituted itself, e.g., D(1−2)<<A(1−3)|B(1−4)>>C means the structure is either D(1−2)A(1−3)C or D(1−2)B(1−4)C. If a variant inside a substituted uncertain block is longer that one residue, it is assumed that the residue on its reducing end is substituted, i.e., A<BC|D>E is interpreted as ADE OR B[A]CE, rather than ADE OR ABCE. More than two variants inside a fuzzy block are allowed, e.g., <A|B|C>.

Angle brackets can be nested up to one level, e.g., <<D|<<A|B>>|C>> or <<<A|B>>|C>, etc. Fuzzy residues on the reducing end or in the rightmost position in a polymer repeating unit are not supported. (Use ...[<<lXDco(2−1)|lXLin(2−1)>>]aDGalp, instead of ... aDGalp(1−2)<<lXDco|lXLin>>.) Encoding of an example structure with several uncertainties is depicted on Figure 1C.

Monovalent and Inorganic Acid Residues. All monovalent substituents (Ac, Me, Et, Fo, and other unsubstitutable residues from the controlled vocabulary) should be described as separate residues, e.g., aDGal(1−3)bDGlcNAc should be recorded as aDGal(1−3)[Ac(1−2)]bDGlcN. If a monovalent residue is an aglycone at the reducing end, then the following syntax is used: aDGlc(1−Me.

Except for a bisubstitution, phosphates and sulphates should be included into the linkage parentheses like this:

aDGlc(1−P−4)bLFuc (in the chain) or P−4)bLFuc (at the nonreducing end) or aDGlc(1−P (at the reducing end). Longer chains (di-, tri-, etc.) are allowed: aDGlc(1−P−P−4)bLFuc.

Distinguishing between Main and Side Chains; the Side Chain Order. The chain is called normal if it is not secondary. The chain is secondary if it starts with any of the following residues (has it at the reducing end or consists only of it):

- Monovalent residues (Ac, Me, etc.).
- Phosphates (P) and sulphates (S).
- Subst and SubstN alias (see below).

To keep this encoding unambiguous, the chains are ranked as the following:

- In the case of polymers, a chain that forms the polymer backbone is always main.
- In oligomers and fragments of structure, if one chain is normal and others are secondary, then the former chain is taken for main.
- If all chains are either normal or secondary, then the chain substituting a position with smaller number is taken for main.
- If there are several side chains attached to one residue, then the order in which they follow inside square brackets is substitution position descending.
- Variants inside angle brackets are sorted the same way as side chains (substitution position descending).

When comparing substitution positions, the following special cases exist:

- A question mark (?) is always greater than any numeric position.
- If a donor is not fully elucidated, then this may result in a fuzzy substitution position in the acceptor. Such position has the same rank as question mark, e.g., <...−5)|...−7)> is greater than ...−8).
- If substitution positions are the same (e.g., both are ? or result from a construction like <A(1−5)|B(1−5)>), then alphanumeric comparison of the names of the reducing-end residues is applied.

Nonstoichiometric Linkages. A linkage is nonstoichiometric if its donating residue is present in nonstoichiometric amounts in the polymer repeating units or if the structure represents a mixture of oligomers. In this case the residue name should be preceded with the stoichiometry degree in percents (e.g., 40%bDGlc). A single percent sign without a number (e.g., %Ac) means that the residue is present in nonstoichiometric amount but its exact amount is unknown. Phosphate and sulfate residues can be preceded by percentage as well, e.g., xDRib-ol(1−50%P−4)bDGalp.

The percentage is applied only to the outgoing linkage of the residue, which means that if the residue is substituted, then all its children chains are nonstoichiometric too, e.g., 40%A(n−m)40%B(p−q)C means that 40% of the residues C are substituted with the residue B, of which 40% is substituted with the residue A (a moiety of B in the whole structure is 40%, and a moiety of A is $40 \times 40\% = 16\%$).

Aliases and Superclasses. If the exact residue at a certain location of the structure is unknown, a superclass name can be used instead of a residue name. Superclasses do not require anomeric and absolute configurations and ring size. The following superclasses are supported: TET (any tetrose), PEN (any pentose), HEX (any hexose), HEP (any heptose), OCT (any octose), NON (any nonose), SUG (any monosaccharide), ALK (any alkyl chain), LIP (any acyl chain), CER (any N-acylated sphingoid), and SPH (any sphingoid).

Aliases are used if a residue has no clear chemical definition or is missing from the vocabulary or if there is a structural feature that cannot be encoded with the language. Alias types allowed are: Sug (some new sugar), Subst, and SubstN, where N is a number (other substituent).

Sug alias should have an anomeric and absolute configuration and a ring size (which may be ?=unknown). All aliases should be explained in the comment section of the encoding after two slashes (//). Several alias explanations are separated by semicolons, e.g., Subst1(?−6)aDGlc-(1−?)Subst2 // Subst1 = acyl or polyglycerol; Subst2 = unelucidated Lipid A.

If Subst or SubstN alias stands for aglycone (a moiety on the reducing end that cannot be encoded using the rules listed above) and is attached to unknown or default (C1 of aldoses and C2 of ketoses) position, then it should be encoded in the aglycone field rather than in the structure.

Database Architecture. During the development of BCSDB schema, the following requirements were taken into account: (1) The database is importable from and exportable to a human-readable dump; (2) there are as few free-text fields as possible; (3) the internal representation of structures is based on a connection table approach, and descriptions in BSCDB linear notation are regenerated for verification; and (4) there is as much indexed data as possible. As a result, a relational database was created. The relationships between entities are summarized in Figure 2 as intertable connections (an arrow means a one-to-many relationship). The more detailed database schema containing tables filled with sample content for clarity is available at http://www.glyco.ac.ru/bcsdb3/help/bcsdb_schema.pdf.

The header colors indicate six major data domains: structure (cyan), compound properties (blue), bibliography (red), taxonomy (green), NMR (pink), and interconnections and service information (yellow). The first column in each table lists data fields (primary keys in bold), and the second column contains data types (N stands for an integer number, ENUM is for terms from a list, TEXT is a pregenerated or free text, BOOL is a Boolean value (yes/no), and FORMAT is for formatted data, e.g., date or disease code). Integer indices have different colors to track where else they appear.

There are four exportable indices: a compound ID identifies a unique molecule, an article ID identifies a unique publication, an organism ID identifies a unique organism (strain), and a record ID says under which entry in a dump file three former indices come together.

The following sections explain which data are stored in each domain:

Compound (Blue). Tables of this domain associate the compound ID with compound-related data: type of the structural unit (oligo, poly, etc.), molecular weight and formula, aglycone information, BCSDB linear code generated from the structure for error tracking, trivial name, compound class, and references to external resources (Chemical Abstracts Service registry numbers, patent

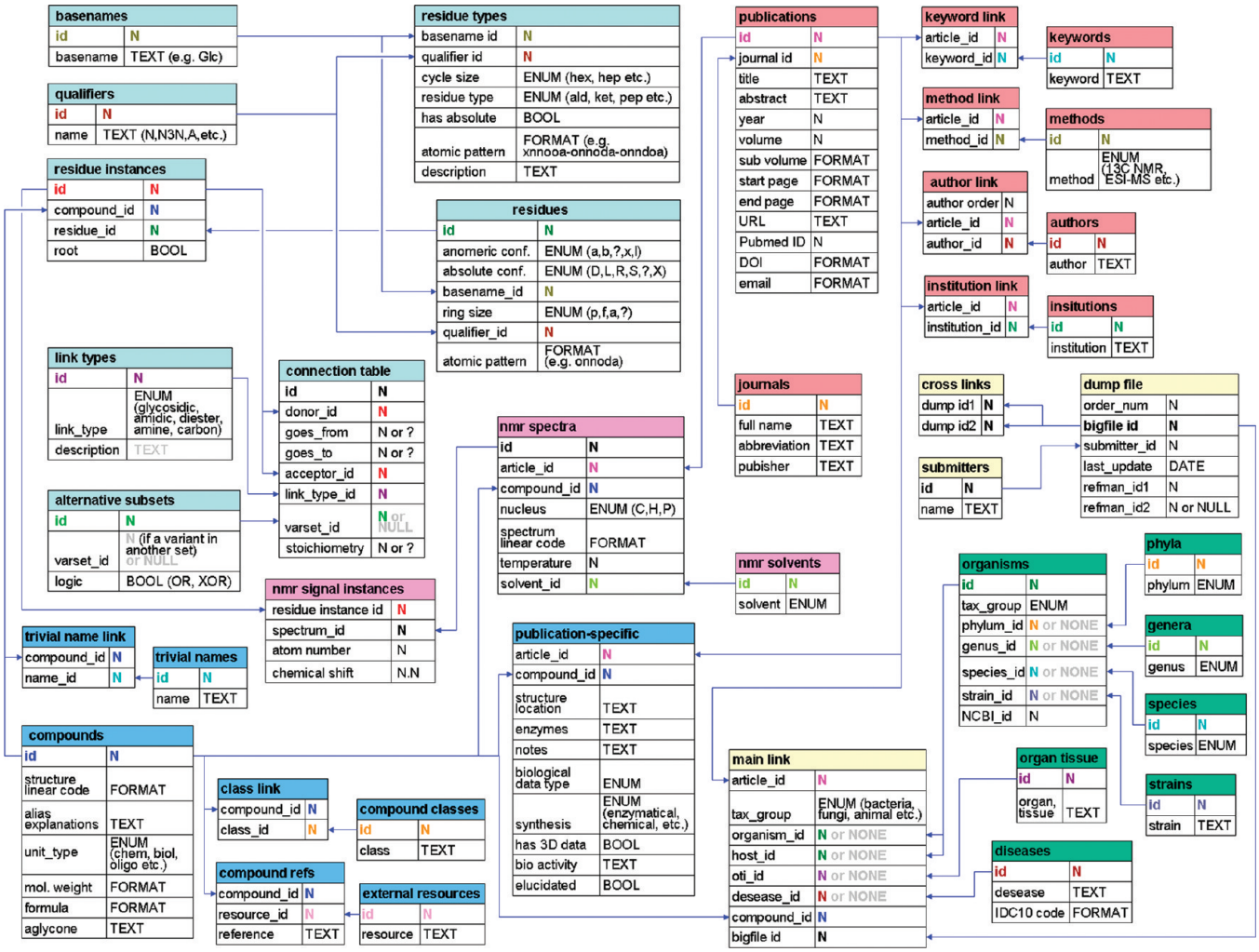

Figure 2. Entity relationships of BCSDB (database schema). See explanations in the text.

numbers, etc.). The last three data types are indexed. Publication specific data (the flag identifying whether the structure of the compound was elucidated within this publication, location of the compound reference inside the publication, information on synthesis, biological activity, biosynthetic pathways, and conformation) are stored in a separate table that links them to both compound and article identifications (IDs).

Structure (Cyan). The domain stores information for all residue instances in all compounds in the database and the connection table that links these residue instances together. Every instance of every residue is associated with residue information imported from the residue vocabulary, absolute, anomeric and ring size configurations and with the compound ID. The residue-related information, including atomic pattern, can be accessed via a residue subdatabase (upper part of the cyan domain in Figure 2). Besides regular linkages, the connection table format supports multiple and nonstoichiometric linkages and linkages with alternative substructures. The connection table approach significantly increases search performance and allows automatic tracking of many errors on import.

Bibliography (Red). The domain links article IDs with issue data, publication abstract, author contact information, external references (PubMed ID, DOI, and Web URL) and indexed data which include journal information, authors and their order, institutions involved, keywords, and analytical methods utilized in the publication.

Biology (Green). The domain is proposed to store taxonomical annotations for organism IDs and biological annotations of structures. These annotations (all indexed) include phylum, genus, species, strain or serogroup, reference to the NCBI Taxonomy database, organ or tissue, and associated disease. The coverage of phylogenetic information is complete; thus every microorganism or host organism can be located in the tree of life.

NMR (Pink). The domain stores assigned ^{13}C and ^{1}H NMR spectra, if they are available in publications, and related information (solvent and temperature) in accordance with compound ID, article ID, and residue instances. Human-readable descriptions of NMR assignment tables are generated from the spectra for error control.

Interconnections (Yellow). The domain links all other domains together and tracks location of data in a human-readable dump file which is used by BCSDB staff for database updates and validation.

Coverage and Data Sources. BCSDB includes information referenced in scientific publications only. Currently it contains data on 9021 distinct carbohydrate structures, 3749 publications, and 4034 bacterial strains. This number covers nearly all bacterial carbohydrate structures published up to 2009 inclusively. About 600 more records, mainly those from

publications in 2010, are in the process of validation. The average content growth is about 600 structures and 300 publications per year. New data appear in BCSDB approximately a year after their publication.

A majority of data published before 1996 (approximately 3,000 structures and 2,000 publications) were imported from CCSDB (CarbBank).[6] Other data came from retrospective literature analysis (structures published after 1995) and manual data posting.

Unfortunately, carbohydrate journals do not demand obligatory posting of published carbohydrate structures to the repository, thus, submission of user's own data to the database remains arbitrary. Nevertheless, we deployed tools for the whole process of user data submission, including data upload via the project Web site, automatic prevalidation, and interaction with database curators.

Error Tracking. During the import of CCSDB data to BSCDB we tracked all structures that could not be automatically parsed, and if the proper structure could not be deduced from the context, then we compared them to the original publication. The CCSDB biological annotations, especially strain information, have been checked for every record. According to our calculation of occurrences of improperly formatted and/or erroneous data in CCSDB dump, more than 30% of CCSDB entries associated with bacteria domain contain errors. This number includes errors inherited from original publications, such as defected pseudographical representation of structures, which spoils the identification of branching points. Other databases that utilize CCSDB data, except Glycome-DB, have not been reported to possess extensive error checking procedures, which allows one to assume that the quality of data imported from CCSDB did not increase significantly. In Glycome-DB, many types of erroneous structures are detected and prevented from import, thus they are not deposited until manual correction in the primary database occurs.

Import of BCSDB, including its CCSDB part, and other databases into Glycome-DB attempted in 2008[30] and subsequent data quality evaluation showed that the quality of data is one of the major problems in carbohydrate databases. In many cases, errors can only be detected by re-examining the original publications, which makes the error correction a highly time-consuming process that cannot be fully automated. There are three types of content errors:

- Errors that can be corrected automatically, e.g., spelling of journal names, missing genus for a specified species, redundant stereochemical information in residue names, in some cases missing cross-references to other databases, etc.
- Errors that can be detected automatically but require acquisition of additional information for correction, e.g., partial structural information, chemically impossible substitution, and incomplete stereochemical information in residue names. This group includes some syntactical errors in primary sources, since free-text information often cannot be analyzed automatically. To discover errors of this group, we implemented routines called on every act of data import. They are proposed to check the syntax and the chemical admissibility of structures, the existence of monomeric names extracted from parsed structures, and the self-consistency of bibliographic and taxonomic information. Errors that do not imply loss or corruption of data are corrected automatically, and those that do are annotated and enclosed to the data for manual correction and approval. If manual validation reveals structural errors present in an original publication, then the structures are corrected by experts in carbohydrate chemistry, if it is clear what the authors meant, and a special annotation is added. Accordingly to our studies, ca. 10% of publications contain such errors, and for those data imported into CCSDB, processing by the import engine increased this value to ca. 20%.
- Errors that cannot be detected automatically, e.g., a mistyped linkage position that still remains chemically and biologically possible. Such errors can be corrected only during the systematic comparison of data with original publications, which was performed for a part of the records only.

Table 2 summarizes widespread errors in data published in literature or imported from CCSDB. The last column indicates whether BCSDB can handle the error without human expert participation. Only errors that can be detected by BCSDB engine are listed.

Interface and Usage. BCSDB has a Web front-end with freely available user part and password-protected administrative part. Both usage of the database and its mechanism are documented on the project Web site. User operations are listed in the main menu of the project (on left of the lower screenshot on Figure 3).

Users can search the database using IDs, fragments of structure, taxonomical, bibliographical, and NMR spectroscopic data. Search requests of different types can be combined using logical AND ("search in the results of the previous query"), OR ("combine with the results of the previous query"), and NOT ("negate search") operations as many times as desired.

The database answers are grouped accordingly to the search type:

- After a search for a structural fragment, answers are grouped by compounds and contain major compound-related data (compound ID, structural formula in IUPAC extended format, structure type, aglycone, molecular weight, chemical formula, trivial name, compound class, and references to other structural databases) and the list of publications in which the compound is described. Every publication in this sublist is accompanied with a link to BCSDB record ID and a list of associated organisms within the record. The example of such output is depicted on Figure 3 (lower screenshot).
- After a search for bibliographic data, answers are grouped by publications and contain major publication-related data (article ID, authors, title, issue data, keywords, publisher, involved institutions, corresponding author email, used methods, and references to other bibliographic databases) and the list of compounds that are described in the publication. Every compound in this sublist is accompanied with a link to BCSDB record ID and a list of associated organisms within the record.
- After a search for taxonomic data, answers are grouped by organisms and contain major organism-related data

Table 2. Widespread Content Errors Detectable by BCSDB

error description	example of erroneous notation	suggested proper BCSDB notation	auto-correctable
violation of residue naming conventions	4-deoxy-Xyl (for 4-deoxy-threopentose)	??4dthrPen?	sometimes
missing position identifiers for modification	anhydro-Kdo	?X2,7anhKdo?	no
type error	aDGcl*p*N	aDGlc*p*N or aDGal*p*N	sometimes
a special residue notation exists but is not used	6d-Gal (instead of Fuc)	??Fuc?	yes
impossible or missing stereochemistry	D-GlcHep (for D-glycero-D-glucoheptose)	?XDDglcHep?, ?XLDglcHep?, etc.	no
unexplained entity	Sug(1-3)aDGl*cp*	??Sug?(1-3)aDGl*cp* // Sug=...	no
redundant stereochemical information	D-Kdo	?XKdo?	yes
contradictory modifications	2,3-anhydro-Glc*p*N	??3,4anhGlc*p*N, ??2,3anhGlc*p,* etc.	no
anomeric or absolute configuration or ring size is specified for a residue that cannot have it	a-Rib-ol	x?Rib-ol	yes
impossible ring size	D-Gal*p*5N	?DGal*p*4N, ?DGal*f*5N, etc.	sometimes
missing ring size or anomeric or absolute configuration	Glc	??Glc?, bDGlc*p*, etc.	no
incomplete or missing linkage information	a-D-Man*p*-*P*-D-Rib-ol	aDMan(1-P-?)xDRib-ol	no
substitution at chemically impossible position (or C-C bond is required)	aDGlc*p*(1-6)aDFuc*p*	e.g. aDGlc*p*(1-4)aDFuc*p*	no
substitution beyond the carbon skeleton size	Ac(1-8)aDDmanHep*p*	e.g. Ac(1-7)aDDmanHep*p*	no
multiple donors substitute the same position in the acceptor	aDGlc*p*(1-2)aDGlc*p*NAc	e.g. aDGlc*p*(1-3)[Ac(1-2)]aDGlc*p*	no
aglyca encoded as free text rather than as a part of structure	aDGlcp(1-methyl	aDGlcp(1-1)Me	yes, except a few cases
incomplete or contradictory sequence information (often due to wrong location of side chains in pseudo graphical 2D IUPAC figures in publications)	Ac(1 | 2 aDMan*p*(1-4)bDGal*p*	Ac(1-2)aDMan*p*(1-4)bDGal*p* or aDMan*p*(1-4)[Ac(1-2)]bDGal*p*	no
unit type does not correspond to the structure, or odd number of open termini	-5)aDGlc*p*(1-3)bDGal*p*	-5)aDGlc*p*(1-3)bDGal*p*(1-	sometimes
multiple records provide different molecular properties for the same structures			no
different notation for methods, compound classes etc.	13C NMR, NMR 13C, ESI-MS, electrospray MS	NMR 13C, ESI-MS	yes
contradictory issue data (a volume does not match a year, a journal did not exist in that year, page number violation etc.)	Carbohydr. Res., 2001, v. 100, pp. 345-340	Carbohydr. Res., 2001, v. 331, pp. 345-350	sometimes
different notations for national characters	Mueller A, Müller B	Mueller A, Mueller B	yes
improper abbreviation of journal names	Carb. Research	Carbohydr. Res.	no
multiple records have different publication authors or titles for the same issue data			no
missing domain, phylum, genus, species or contradictory taxonomical information	*Escherichia sapiens*	*Escherichia coli, Homo sapiens*	yes, except a few cases
too much data for a serogroup (often due to missing species separator)	E. coli R14 P. penneri 34	*Escherichia coli* R14, *Proteus penneri* 34	no
internal cross-references are orphaned or not bijective			yes
format of an external reference is improper for the specified resource	PubMed ID: 100.10/2	PMID: 100102, DOI: 100.10/2	no
NCBI Taxonomy reference does not match specified genus or species			yes
number, naming or location of residues in the NMR assignment table does not match the structure			no
the number of NMR signals specified for the residue does not match the number of atoms			no
NMR chemical shifts are beyond the characteristic range for the specified nucleus			no

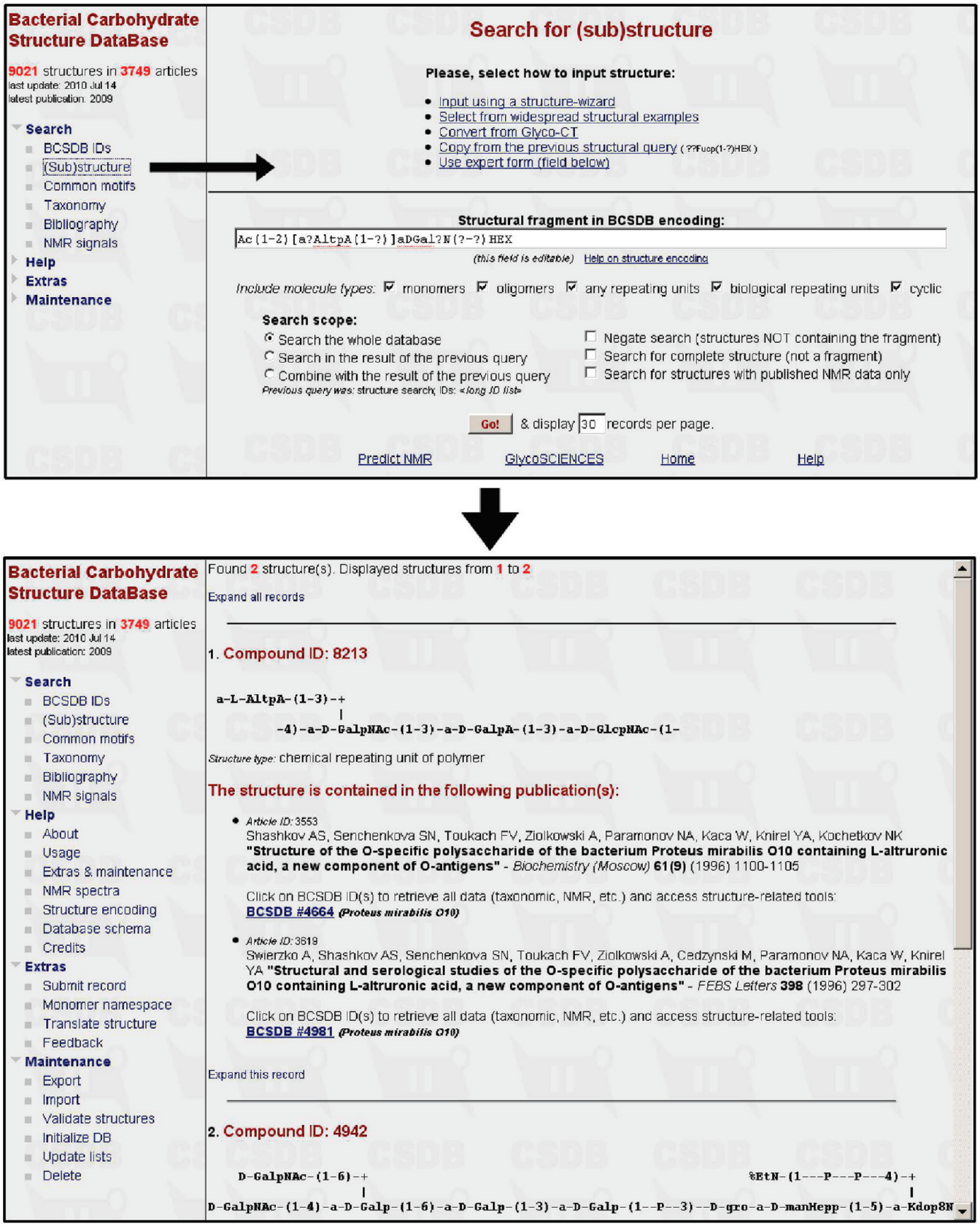

Figure 3. An example query and database output. The upper screenshot presents an example of a substructure search top page with sample structure entered. The lower screenshot presents a part of the database output in collapsed form. The project menu items on the left of the lower screenshot are fully opened to reflect all top-level user operations.

(organism ID, systematic name, taxonomic domain, phylum, and references to other taxonomic databases) and the list of compounds that are associated with the organism. Every compound in this sublist is accompanied with a link to BCSDB record ID and associated publication within the record.

- After a search for BCSDB record ID or following links to the BCSDB record IDs in three previous

cases, answers are grouped by records displaying all data originating from the paper.

Carbohydrate structures are output to users in IUPAC extended format and can be converted into other representations on-the-fly. This format was used as a display format in CCSDB and Sweet-DB[13] (ancestor of GLYCOSCIENCES.de) and is widely accepted by carbohydrate researchers. The output format was extended to adopt all supported structural features (alternative substructures, incompletely elucidated structures, nonstoichiometric entities, dual linkages, etc.).

Navigation through the multiple output pages and output resorting tools are available.

Below is a brief description of the search query interface.

ID Search. Search::BCSDB IDs on the Web page is a direct way to access BCSDB data using record, article, compound or organism ID, or ID range (e.g., "1−10,15, 20−22").

Structural Search. (Search::(sub)structure on the Web page) Form allows specification of a structural fragment (or a whole structure) that should be present in all returned compounds. The search can be limited to a certain unit type (e.g., oligomeric structures only) or to structures with NMR assignment only. The top page of the structure search is depicted on Figure 3 (upper screenshot) together with an example query. There are several ways to input the structure: copy from previous queries and modify, select from a list of widespread carbohydrate fragments and modify, convert from GlycoCT encoding, enter directly in BCSDB linear format, or input using a wizard. The structural wizard allows "assembly" of the structure by visual operations with selectors and dropdown lists. Usage of the wizard does not require any special knowledge except nomenclature of monosaccharides. It has some limitations (no more than four carbohydrate residues, no dual linkages, and no alternative substructures), but its output can be edited directly. More details are available at http://www.glyco.ac.ru/bcsdb3/help/usage.html#wizard.

Bibliographic Search. (Search::bibliography on the Web page) Form is intended to retrieve data using partial or complete bibliographic information: authors, terms contained in title and/or abstract, keywords, journal name, year or range of years, volume number, and page numbers. The query language supporting logical operations, term grouping, and wildcards can be used to construct complex queries. Author names are indexed to avoid spelling errors. The search can be limited to publication with primary structure elucidation only. More details are available at http://www.glyco.ac.ru/bcsdb3/help/usage.html#bibl.

Taxonomic Search. (Search::taxonomy on the Web page) Form allows selection of taxonomic identifiers that should fit returned organisms. As soon as a genus is selected, the lists of species and strains are regenerated. The search can be limited to certain taxonomic domains or to host organisms only. If only genus is specified, all its taxonomic children (strains with or without species information) are also processed, unless specified otherwise. The form allows usage of NCBI TaxIDs as search criteria and external processing of taxonomical information by the NCBI taxonomy database. More details are available at http://www.glyco.ac.ru/bcsdb3/help/usage.html#bac.

NMR Search. (Search::NMR signals on the Web page) Form processes chemical shifts and returns compounds with matching NMR spectrum assigned. The chemical shift set can be limited to a single residue. To compare spectra, the BCSDB engine forms all possible subspectra of the stored NMR spectrum, with the primary limitation of the subspectrum size to the number of signals in the user input. The best-fitting subspectrum is used to calculate the similarity value. If the user input has less signals than the ideal experimental subspectrum, which may occur due to signal overlap, then it will result in a slightly different set of subspectra and, thus, lower but still nonzero similarity. The similarity is a reverse-logarithmic estimation of average difference between signals. A value of 0 stands for no similarity at all, and a value of 1000 is for full similarity (exact match of chemical shifts). If a compound has more than one spectrum (e.g., in different conditions or in different publications), then the similarity for this compound is calculated as an average of the similarity values of all spectra assigned to this compound for the given nuclei. The output compounds are sorted in a similarity-descending order, and only the spectra possessing higher similarity values than specified are displayed. The chemical shifts close to those from the search term are highlighted. More details are available at http://www.glyco.ac.ru/bcsdb3/help/usage.html#nmr.

Besides search functions, other glyco-related features are available:

Data Submission. This form leads users through all steps of data upload, including error checking and communication with the BCSDB team. User structures, NMR assignment tables, and other data can be checked for integrity independently of the submission process.

Translation Tool. This allows conversion of structures from GlycoCT{condensed} and to GlycoCT{condensed}, GlycoCT{XML}, GlydeII, and LINUCS.

NMR Prediction. This tool uses the extended BIOPSEL approach[31] to predict ^{13}C NMR assignment table for a given structure. Within BCSDB, the BIOPSEL algorithm was adapted to oligomeric structures, improved to treat ketosugars and other 'special cases', and equipped with a Web front-end.

The prediction engine utilizes the spectroscopic subdatabase and substitution effect subdatabase that contain averaged literature data on chemical shifts, glycosylation, and phosphorylation effects. The approximate coverage is 80 residues, 2500 dimers and trimers, and 150 effects; data are averaged for D_2O solutions at 318 K. The following structural peculiarities are taken into account when searching for particular chemical shifts or substitution effects:

- Which residue is substituted.
- Its anomeric configuration, if it exists.
- The substitution position.
- The type of substituent: pyranose, furanose, alditol, phosphate, and noncarbohydrate.
- Substituent anomeric configuration, if it exists.
- Additional groups attached to C1 of a substituent (none, carboxyl group, and C-chain).
- The type of group at C2 of a substituent (−OH/−NH2/deoxy).

- The orientation of proton at C2 of a substituent (axial/equatorial/unoriented or both).
- The combination of absolute configurations of the substituent and the residue itself (same or different).

The NMR predictor iterates through all residues in the structure and searches the spectroscopic subdatabase for chemical shifts characteristic of this residue in given structural surroundings. If these data are not found, then the subspectrum of the residue is calculated from the spectrum of the free residue and the substitution effects.

If the desired effect is missing from the database, then the type and the orientation of the substituent C2 position are varied until the effect is found. If the effect is found for none of the variants, then the residue being predicted is temporarily replaced with a common residue with the same basic configuration (e.g., Gal instead of FucNAc). If still no effect is found, then it is simulated as +6.0 on α-carbon, −1.0 on β-carbons, +6.0 on C1 of the substituent (for O-linked residues) or as −3.0 on alpha-carbon, +1.0 on β-carbons, and +3.0 on C1 of the substituent (for N-linked residues). If a residue subspectrum is calculated using substitution effects, then chemical shifts of C2 and C5 of C1-linked pyranoses are modified accordingly to known substitution effects for these atoms.[32] The glycosylation effects for three widespread sugar configurations (glc, gal, and man) are represented most completely, usually making the effect prediction for these basic types more accurate.

For every act of prediction, all made assumptions are recorded, and the consistency value is calculated to reflect the degree of trustworthiness of the result. More details are available at http://www.glyco.ac.ru/bcsdb3/help/nmr.html.

Integration with Other Projects. Most of the databases offer no routine or strategy for automated data access, so interested researchers are forced to extract information directly from a dump or even HTML pages. The automated programming interface that renders data from BCSDB in a standardized way has not yet been ported from the previous version (BCSDB 2), where it was realized using XML-wrapped pipelines based on small objects access protocol.[33] However BCSDB is capable of generating data in formats accepted by a number of projects, particularly bibliographic data that can be automatically processed by NCBI PubMed,[34] making cross-requests to NCBI taxonomy[31] database and GLYCOSCIENCES.de, and serving requested data within its own interface. Structure translator to other glycan description languages (GlydeII, GlycoCT, and IUPAC condensed) has a feature of formatted answer to POST requests. All types of BCSDB IDs (records, records, structures, publications, and organisms) can be used to reference BCSDB data from the other projects using Web links. Particularly, BCSDB records are referenced from Glycome-DB pages containing corresponding glycan structures that has been integrated in Glycome-DB.

TECHNICAL DETAILS

BCSDB utilizes MySQL 4.21 database engine. The search functionality and the other code are implemented in PHP 5.2. The interface is realized in Web form using server-generated DHTML and JavaScript. The browser compatibility was tested in Microsoft Internet Explorer 7 and Mozilla Firefox 3. BCSDB runs on a dedicated Linux Red Hat-driven server and is freely available at http://www.glyco.ac.ru/bcsdb3/.

CONCLUSIONS

Bacterial carbohydrate structure database (BCSDB) version 3 unites ideology and features of many carbohydrate databases to provide unique coverage and data consistency. This open-access database simplifies interpretation of information on bacterial carbohydrates acquired by the scientific community. It extends the opportunities of researchers studying structure and biological functions of glycans, teichoic acids, glycoproteins, glycolipids, and related compounds.

ACKNOWLEDGMENT

The pilot version of BCSDB was developed within the framework of the International Science and Technology Center Partner Project grant no. 1197, supported by the Cooperative Threat Reduction Program of the United States Department of Defense. The part of data came from the CCSDB database created in the University of Georgia (Athens, GA) and the bibliographic database of the Laboratory of Carbohydrate Chemistry of N.D. Zelinsky Institute of Organic Chemistry (Moscow, Russia). Later, the project was supported by the Russian Foundation for Basic Research grant N05-07-90099, the Russian Federation President program grant MK-1700.2005.4, and the Deutsches Krebsforschungszentrum (Heidelberg, Germany) guest scientistship stipendia in 2006−2009.

The author thanks Prof. Y.A. Knirel (Zelinsky Institute) for useful collaboration. Complete list of people involved in BCSDB development and their roles is published at http://www.glyco.ac.ru/bcsdb3/help/credits.html.

REFERENCES AND NOTES

(1) Lowe, J.; Marth, J. A genetic approach to mammalian glycan function. *Annu. Rev. Biochem.* **2003**, *72*, 643–691.
(2) 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* (4), 568–577.
(3) Harvey, D. Proteomic analysis of glycosylation: structural determination of N- and O-linked glycans by mass spectrometry. *Expert Rev. Proteomics* **2005**, *2*, 87–101.
(4) Guerardel, Y.; Chang, L.; Maes, E.; Huang, C.; Khoo, K. Glycomic survey mapping of zebrafish identifies unique sialylation pattern. *Glycobiology* **2006**, *16*, 244–257.
(5) Ranzinger, R.; Frank, M.; von der Lieth, C.-W.; Herget, S. Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences. *Glycobiology* **2009**, *19* (12), 1563–1567.
(6) Doubet, S.; Bock, K.; Smith, D.; Darvill, A.; Albersheim, P. The Complex Carbohydrate Structure Database. *Trends Biochem. Sci.* **1989**, *14* (12), 475–477.
(7) Doubet, S.; Albersheim, P. Carbbank. *Glycobiology* **1992**, *2*, 505.
(8) Jones, C. Vaccines based on the cell surface carbohydrates of pathogenic bacteria. *An. Acad. Bras. Cienc.* **2005**, *77*, 293–324.
(9) Herget, S.; Toukach, P.; Ranzinger, R.; Hull, W. E.; Knirel, Y.; von der Lieth, C.-W. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct. Biol.* **2008**, *8*, Article 35; http://www.biomedcentral.com/1472-6807/8/35. Accessed August, 2010.
(10) Seeberger, P. H. Automated carbohydrate synthesis as platform to address fundamental aspects of glycobiology - current status and future challenges. *Carbohydr. Res.* **2008**, *343* (12), 1889–1896.
(11) EuroCarbDB Reports; DKFZ: Heidelberg, Germany; http://www.eurocarbdb.org/about/reports. Accessed September 30, 2010.
(12) Lütteke, T.; Bohne-Lang, A.; Loss, A.; Götz, T.; Frank, M.; von der Lieth, C.-W. GLYCOCIENCES.de: An Internet Portal to Support

Glycomics and Glycobiology Research. *Glycobiology* **2006**, *16* (5), 71R–81R.
(13) Loss, A.; Bunsmann, P.; Bohne, A.; Loss, A.; Schwarzer, E.; Lang, E.; von der Lieth, C.-W. Sweet-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res.* **2002**, *30* (1), 405–408.
(14) Cooper, C.; Joshi, H.; Harrison, M.; Wilkins, M.; Packer, N. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.* **2003**, *31* (1), 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* (5), 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* (5), 63R–70R.
(17) 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* (9), 1214–1216.
(18) Maes, E.; Bonachera, F.; Strecker, G.; Guerardel, Y. SOACS index: an easy NMR-based query for glycan retrieval. *Carbohydr. Res.* **2009**, *344* (3), 322–330.
(19) Stenutz, R.; Weintraub, A.; Widmalm, G. The structures of *Escherichia coli* O-polysaccharide antigens. *FEMS Microbiol. Rev.* **2006**, *30* (3), 382–403.
(20) von der Lieth, C.-W. An Endorsement to Create Open Databases for Analytical Data of Complex Carbohydrates. *J. Carbohydr. Chem.* **2004**, *23*, 277–297.
(21) Herget, S. Glycobiology at the Interface to Informatics. In *Development of Databases, Bioinformatic Procedures and Analytics for Glycobiology (PhD thesis)*, 1st ed.; Krauth-Siegel, L., Ed.; Universität Heidelberg: Heidelberg, Germany, 2009; pp 22−33.
(22) McNaught, A. D. Nomenclature of carbohydrates (recommendations 1996). *Adv. Carbohydr. Chem. Biochem.* **1997**, *52*, 43–177.
(23) Sahoo, S. S.; Thomas, C.; Sheth, A.; Henson, C.; York, W. S. GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydr. Res.* **2005**, *340* (18), 2802–2807.
(24) Kikuchi, N.; Kameyama, A.; Nakaya, S.; Ito, H.; Sato, T.; Shikanai, T.; Takahashi, Y.; Narimatsu, H. The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics* **2005**, *21* (8), 1717–1718.
(25) Herget, S.; Ranzinger, R.; Maass, K.; von der Lieth, C.-W. GlycoCT - a unifying sequence format for carbohydrates. *Carbohydr. Res.* **2008**, *343* (12), 2162–2171.
(26) Banin, E.; Neuberger, Y.; Altshuler, Y.; Halevi, A.; Inbar, O.; Dotan, N.; Dukler, A. A novel linear code nomenclature for complex carbohydrates. *Trends Glycosci. Glycotechnol.* **2002**, *14* (77), 127–137.
(27) Bohne-Lang, A.; Lang, E.; Förster, T.; von der Lieth, C.-W. LINUCS: Linear notation for unique description of carbohydrate sequences. *Carbohydr. Res.* **2001**, *336* (1), 1–11.
(28) Satoh, H.; Koshino, H.; Funatsu, K.; Nakata, T. Representation of molecular configurations by cast coding method. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1106–1112.
(29) Packer, N. H.; von der Lieth, C.-W.; Aoki-Kinoshita, K. F.; Lebrilla, C. B.; Paulson, J. C.; Raman, R.; Rudd, P.; Sasisekharan, R.; Taniguchi, N.; York, W. S. In *Proteomics; 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.NIH:Bethesda,MD, 2006.
(30) Ranzinger, R.; Herget, S.; Wetter, T.; von der Lieth, C.-W. GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinformatics* **2008**, *9*, Article 384; http://www.biomedcentral.com/1471-2105/9/384. Accessed September 30, 2010.
(31) Toukach, P. V.; Shashkov, A. S. Computer-assisted structural analysis of regular glycopolymers on the basis of 13C NMR data. *Carbohydr. Res.* **2001**, *335* (2), 101–114.
(32) Lipkind, G. M.; Shashkov, A. S.; Knirel, Y. A.; Vinogradov, E. V.; Kochetkov, N. K. A computer-assisted structural analysis of regular polysaccharides on the basis of 13C-n.m.r. data. *Carbohydr. Res.* **1988**, *175* (1), 59–75.
(33) Toukach, P.; Joshi, H.; 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., Database Issue* **2007**, *35*, D280–D286.
(34) Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Mizrachi, I.; Ostell, J.; Panchenko, A.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; John Wilbur, W.; Yaschenko, E.; Ye, J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res., Database issue* **2010**, *38*, D5–D16.

CI100150D