Nucleic Acids Research, 2002, Vol. 30, No. 1 405–408

SWEET-DB: an attempt to create annotated data collections for carbohydrates

Alexander Loß, Peter Bunsmann, Andreas Bohne[1], Annika Loß, Eberhard Schwarzer, Elke Lang[2] and Claus-W. von der Lieth[1,*]

University Hildesheim, Institute of Physics and Technical Informatics, Marienburger Platz 22, 31141 Hildesheim, Germany, [1]German Cancer Research Center, Spectroscopic Department, Im Neuenheimer Feld 240, 69120 Heidelberg, Germany and [2]University of Applied Sciences Darmstadt, Department for Information and Knowledge Management, Schöfferstraße 1-3, D-64295 Darmstadt, Germany

Received August 20, 2001; Revised and Accepted October 10, 2001

ABSTRACT

Complex carbohydrates are known as mediators of complex cellular events. Concerning their structural diversity, their potential of information content is several orders of magnitude higher in a short sequence than any other biological macromolecule. SWEET-DB (http://www.dkfz.de/spec2/sweetdb/) is an attempt to use modern web techniques to annotate and/or cross-reference carbohydrate-related data collections which allow glycoscientists to find important data for compounds of interest in a compact and well-structured representation. Currently, reference data taken from three data sources can be retrieved for a given carbohydrate (sub)structure. The sources are CarbBank structures and literature references (linked to NCBI PubMed service), NMR data taken from SugaBase and 3D co-ordinates generated with SWEET-II. The main purpose of SWEET-DB is to enable an easy access to all data stored for one carbohydrate structure entering a complete sequence or parts thereof. Access to SWEET-DB contents is provided with the help of separate input spreadsheets for (sub)structures, bibliographic data, general structural data like molecular weight, NMR spectra and biological data. A detailed online tutorial is available at http://www.dkfz.de/spec2/sweetdb/nar/.

INTRODUCTION

The human genome seems to encode not >30 000–40 000 proteins. This relatively small number of human genes compared to the genome of other species has been one of the big surprises coming out of the analysis of the human genome project (1). A major challenge is to understand how post-translational events, such as glycosylation, affect the activities and functions of these proteins in health and disease. Glycosylated proteins are ubiquitous components of extracellular matrices and cellular surfaces, where their oligosaccharide moieties are implicated in a wide range of cell–cell and cell–matrix recognition events (2–4). Carbohydrate modifications of proteins and lipids are key factors in modulating their structure and function within cells. In the extracellular milieu, they exert effects on cellular recognition in infection, cancer and immune response, but details of the specific mechanisms are most often still rather rudimentary.

The use of proteomics databases has become indispensable for daily work of the molecular biologist, but this situation has not yet been achieved for carbohydrate applications. Even if one takes into account that the number of scientists working on various topics in the glycosciences is considerably smaller than the number of molecular biologists working with proteins and nucleic acids (5), it is obvious that the acceptance of carbohydrate-related data collections is considerably lower in the community of glycoscientists than the acceptance that various proteomics data collections and tools receive by molecular biologists. Moreover, the opposite seemed to become true. The CarbBank project (6–8), the largest collection of carbohydrate-related references that had been built up during the 1980s and 1990s, entered shutdown mode in 1999 due to lack of funding.

Recently, new attempts have been described aiming to create tools which link available information on glycans from various sources. GlycoSuite and BOLD (9,10) are databases which are currently cross-linked with MEDLINE and SWISS-PROT/TrEMBL and contain annotated information extracted from scientific literature on glycoprotein-derived glycan structures. The company GlycoMind (http://www.glycominds.com) currently builds up a database that compiles information about glyco-conjugated molecules, including their structures, functions and interactions with other molecules.

Cross-linking for proteomics tools is mainly achieved on the basis of identity or similarity of gene or protein sequences. Sequences for complex carbohydrates differ significantly from the simple linear form which describes genes and proteins: the number of naturally occurring residues is much larger for carbohydrates, each pair of monosaccharide residues can be

*To whom correspondence should be addressed. Tel: +49 6221 42 4541; Fax: +49 6221 42 4554; Email: w.vonderlieth@dkfz.de

Present addresses:
Alexander Loß and Annika Loß, Gebrüder Gerstenberg GmbH & Co, 31134 Hildesheim, Germany
Peter Bunsmann, Bosch-Blaupunkt, 31134 Hildesheim, Germany

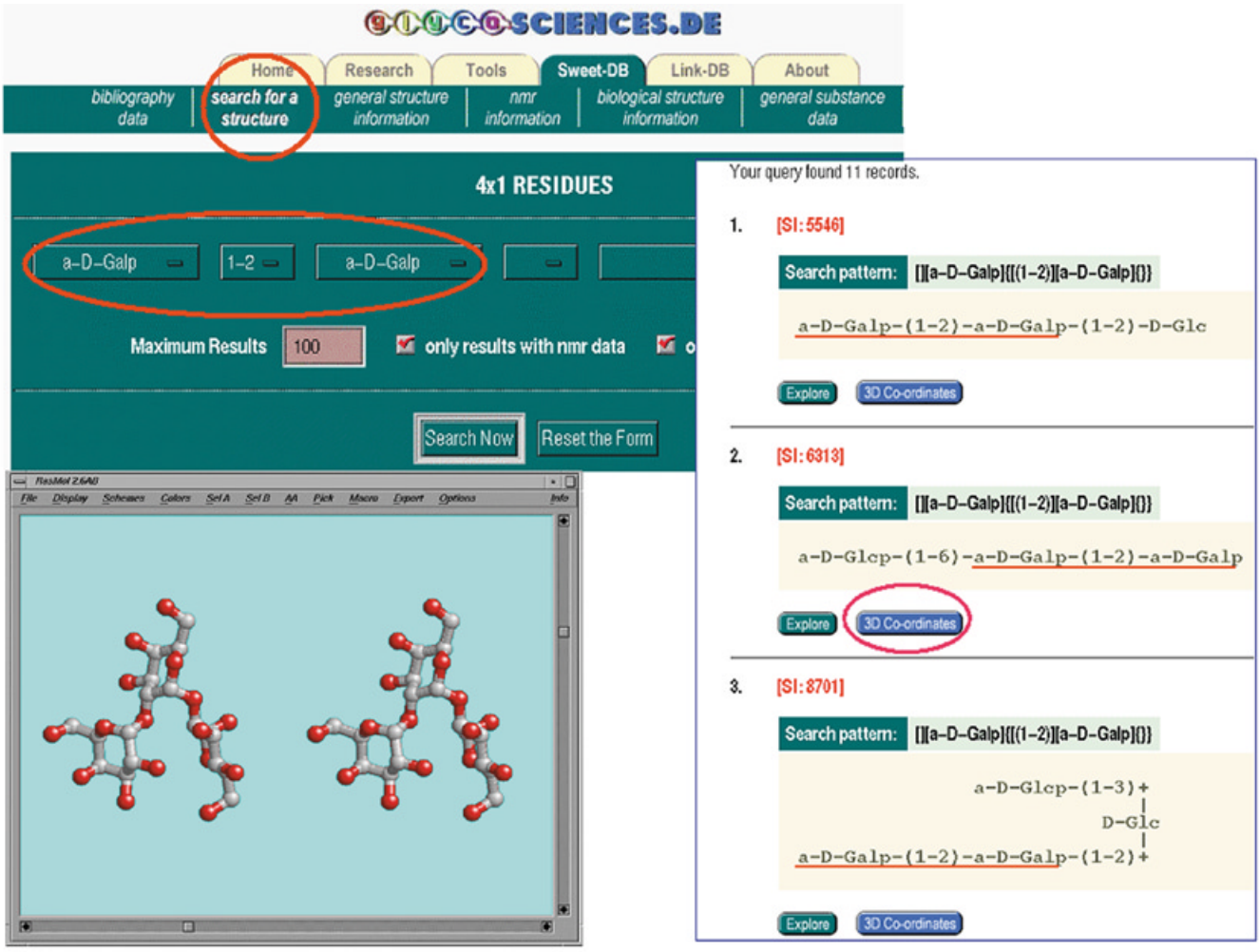


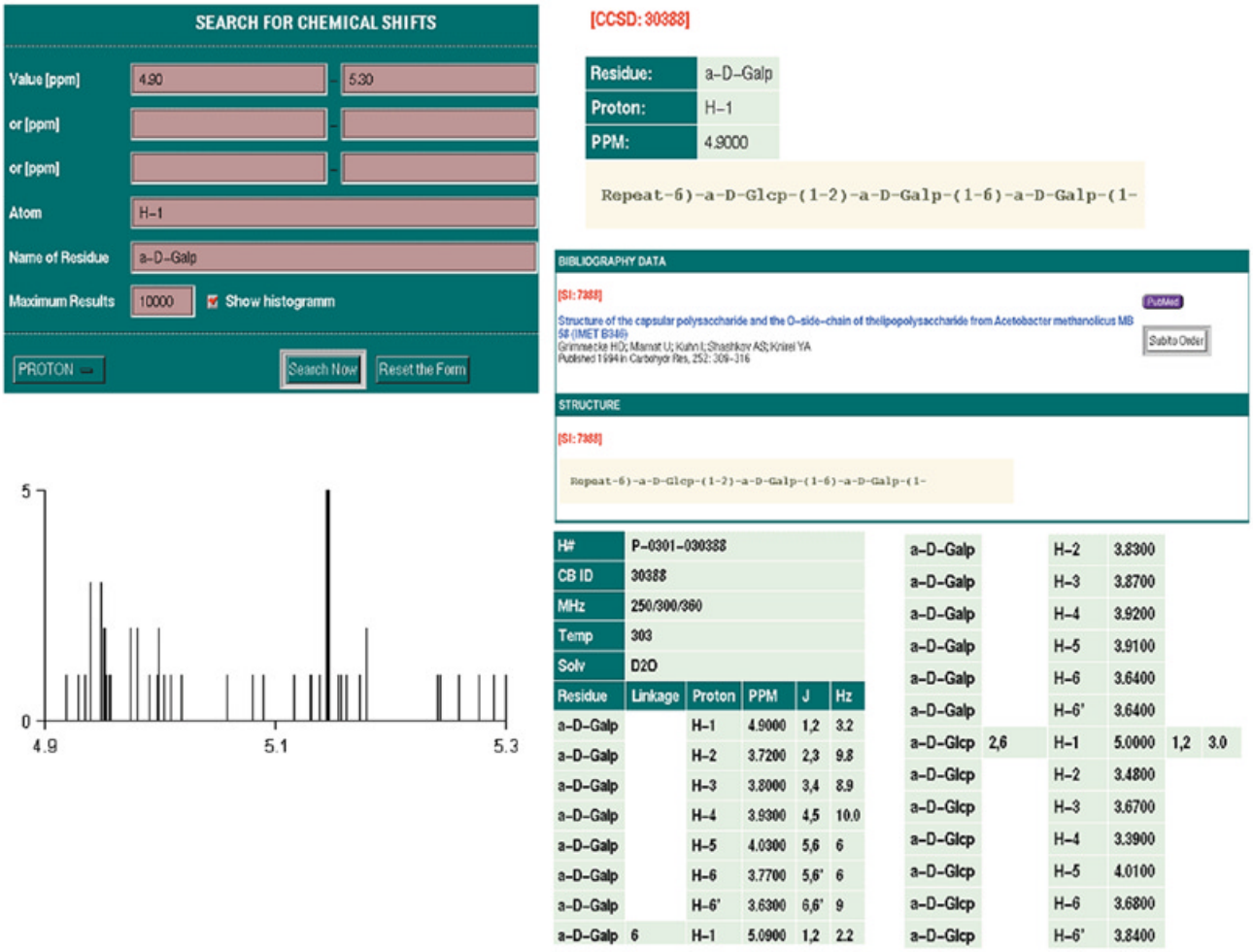
Figure 1. Structure-oriented retrieval of (sub)structure. Input of the topology information is accomplished using 1 × 4 input matrix (top right). Monosaccharides and linkage information can be input using pull-down menus. Eleven structures were found containing the α-D-Galp-(1–2)-α-D-Galp substructure. Topology of first three hits is displayed on the left. Activating the '3D Co-ordinates' button invokes the transfer of a file containing co-ordinates which can be visualized using an appropriate plugin or helper application. Here RASMOL as external helper application is used to display a stereo model of the trisaccharide. Thus, the user has the possibility to look at the structure in different orientations. Activating the 'Explore' button will display all data stored for that sequence.

linked in several ways, and one residue can be connected to three or four others (branching). Thus, a carbohydrate structure database must use more elaborate encoding schemes to be able to describe identity of such structures as well as similarities. Carbohydrates potentially contain information content that is several orders of magnitude higher in a short sequence than any other biological macromolecule (11). Typical N-glycan structures exhibit 9 to ~20 residues. The average sequence length in CarbBank is 6 residues.

SWEET-DB CONTENT

SWEET-DB is an attempt to use modern web techniques to annotate and/or cross-reference carbohydrate-related data collections which allow glycoscientists to find important data for compounds of interest in a compact and well-structured representation. Currently, reference data taken from CarbBank (linked to NCBI PubMed service), NMR data taken from SugaBase (12,13) and 3D co-ordinates generated with SWEET-II (14) can be retrieved for a given carbohydrate (sub)structure.

About 50 000 CarbBank (6–8), entries and 1600 ^{1}H and ^{13}C-NMR spectra taken from SugaBase (12,13) constitute the database for our SWEET-DB implementation. Both collections can be linked using the Linear Notation for Unique Description of Carbohydrate Sequences (LINUCS) (15), a description of the carbohydrate structure which is close to IUPAC-IUBMB nomenclature recommendations (http://www.chem.qmw.ac.uk/iupac/2carb/) (16). Spatial representations were generated with SWEET-II (14) and subsequently optimised using the MM3 force field as implemented in the TINKER package (17). An automatic link to NCBI PubMed (18) service was established based on the reference information provided by the original CarbBank fields.



H#	P-0301-030388
CB ID	30388
MHz	250/300/360
Temp	303
Solv	D2O

Residue	Linkage	Proton	PPM	J	Hz
a-D-Galp		H-1	4.9000	1,2	3.2
a-D-Galp		H-2	3.7200	2,3	9.8
a-D-Galp		H-3	3.8000	3,4	8.9
a-D-Galp		H-4	3.9300	4,5	10.0
a-D-Galp		H-5	4.0300	5,6	6
a-D-Galp		H-6	3.7700	5,6'	6
a-D-Galp		H-6'	3.6300	6,6'	9
a-D-Galp	6	H-1	5.0900	1,2	2.2
a-D-Galp		H-2	3.8300		
a-D-Galp		H-3	3.8700		
a-D-Galp		H-4	3.9200		
a-D-Galp		H-5	3.9100		
a-D-Galp		H-6	3.6400		
a-D-Galp		H-6'	3.6400		
a-D-Glcp	2,6	H-1	5.0000	1,2	3.0
a-D-Glcp		H-2	3.4800		
a-D-Glcp		H-3	3.6700		
a-D-Glcp		H-4	3.3900		
a-D-Glcp		H-5	4.0100		
a-D-Glcp		H-6	3.6800		
a-D-Glcp		H-6'	3.8400		

Figure 2. Using the spreadsheet (top left), all entries having a specific atom (e.g. H-1) in a specific residue (e.g. α-D-Galp) can be retrieved and displayed as frequency-shift histogram (bottom left). Additionally, the shift range of interest can be specified. A list of all entries fulfilling the query is provided. Here only one example is displayed (top right). The complete NMR spectra for this sequence can be recalled. The NMR data are provided as a list of assigned shifts and coupling constants (bottom right).

SWEET-DB ACCESS

The main page of SWEET-DB (http://www.dkfz.de/spec2/sweetdb/) provides several levels of access to the contents of SWEET-DB (Fig. 1, top border). Activating bibliographic search mode produces a spreadsheet which allows the search for authors, catchwords contained in the title, journal name and year of publication. The search for data associated with a complete glycan sequence or parts thereof is definitively the most frequently used way to access SWEET-DB. Depending on the size of the structure one wants to retrieve, different input spreadsheets are provided helping the user to specify correctly the required format for monosaccharide units and linkage information (Fig. 1). The 4×1 input matrix is thought to help the novice user. Frequently occurring monosaccharides are predefined and can be input using pull-down menus. The 4×3 and 6×5 input matrices are suitable for the input of larger, branched sequences. Here the user has to input the monosaccharide unit and linkage type to enable the retrieval of all possible glycan structures.

The matched sequences are displayed using a simple ASCII representation (similar to CarbBank notation). The user has the option to access all data (references, NMR data, molecular weight, glycan composition) stored for one sequence. Additionally a 3D model is available which can be displayed using public domain programs like RasMol, Chime, etc. Access to sequences specifying the range of molecular weight, frequency of atoms, content of monosaccharide components, specific residues and number of residues and branches is provided under the 'general structure information' retrieval option.

If the user wants to find spectroscopic data stored for a certain (sub)structure, the 'search for a structure' option can be used. The input spreadsheet provides the option that only entries containing NMR data will be displayed (Fig. 1). The output of matching entries is accomplished in two steps. In the first step, the sequence of all matched glycans is displayed. In the second step, NMR data are provided as a list of assigned shifts and coupling constants (Fig. 2, bottom) for user-selected entries. In case all glycan sequences shall be retrieved matching certain spectroscopic data (e.g. ^{1}H-NMR shifts), SWEET-DB offers two different retrieval options. (i) Up to

10 NMR shifts (with a certain tolerance) can be input. A list of sequences containing matching spectra is presented in descending order of a score factor. The score factor takes into account the number of matched shifts and their deviation from the input shift. (ii) All NMR shifts assigned to a certain atom within a given monosaccharide unit (for example H1 in α-D-Galp) can be visualised as shift frequency histogram recalled (Fig. 2). Additionally, the shift range of interest can be specified. In such a way the chemical surrounding of each individual NMR shift can be analysed in detail. ^{1}H- as well as ^{13}C-NMR data can be retrieved.

A demonstration of SWEET-DB is provided as a tutorial at http://www.dkfz.de/spec2/sweetdb/nar/.

IMPLEMENTATION

In order to better encode and retrieve the complexity of carbohydrate structures, SWEET-DB has been implemented as a relational database rather than as a flat file. A LAMP (Linux, Apache, MySQL, PHP) system is used to store, retrieve and output the data.

FUTURE DEVELOPMENTS

SWEET-DB is intended as a regular service relational carbohydrate sequence database which strives to provide a high level of annotation, a minimal level of redundancy and high level of integration with other databases. Additionally, we will try to implement AUTO-SWEET-DB as a computer-annotated supplement to SWEET-DB using essentially the same format for both databases. AUTO-SWEET-DB is an attempt to find out how far it is possible to build up and update continuously a data collection (similar to TrEMBL database as an extension to SWISS-PROT for proteins) by scanning and extracting all electronically available resources like abstracts, publications and web pages containing information relevant to glycosciences.

Currently we are working to establish input facilities for SWEET-DB based on a user-friendly web-based interface. A prototype to input new carbohydrate sequences, references and annotations as well as NMR spectra (http://www.dkfz.de/spec2/nmr_eingabe/) is already available. Integration with other online databases is planned.

REFERENCES

1. Venter,J. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
2. Helenius,A. and Aebi,M. (2001) Intracellular functions of N-linked glycans. *Science*, **291**, 2364–2369.
3. Rudd,P., Elliott,T.,Cresswell,P.,Wilson,I. and Dwek,R. (2001) Glycosylation and the immune system. *Science*, **291**, 2370–2376.
4. Wells,L., Vosseller,K. and Hart,G. (2001) Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc. *Science*, **291**, 2376–2378.
5. Hardy,B. and Wilson,I. (1996) Virtual resource development in the glycosciences. *Glycoconj. J.*, **13**, 865–872.
6. Albersheim,P. (1991) Complex carbohydrate structural database. *Glycobiology*, **113**, 113.
7. Doubet,S., Bock,K., Smith,D., Darvill,A. and Albersheim,P. (1989) The complex carbohydrate structure database. *Trends Biochem. Sci.*, **14**, 475–477.
8. Doubet,S. and Albersheim,P. (1992) CarbBank. *Glycobiology*, **2**, 505.
9. Cooper,C., Wilkins,M., Williams,K. and Packer,N. (1999) BOLD—a biological O-linked glycan database. *Electrophoresis*, **20**, 3589–3598.
10. Cooper,C., Harrison,M., Wilkins,M. and Packer,N. (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.*, **29**, 332–335.
11. Laine,R. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yield 1.05 × 1012 structures for a reducing hexasaccharide: the Isomer Barrier to development of single method saccharide sequencing or synthesis system. *Glycobiology*, **4**, 759–767.
12. van Kuik,J. and Vliegenthart,J.F. (1992) Databases of complex carbohydrates. *Trends Biotechnol.*, **10**, 182–185.
13. van Kuik,J., Hard,K. and Vliegenthart,J.F. (1992) A ^{1}H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr. Res.*, **235**, 53–68.
14. Bohne,A., Lang,E. and von der Lieth,C.-W. (1998) W3-SWEET: carbohydrate modeling by Internet. *J. Mol. Model.*, **4**, 33–43.
15. Bohne,A., Lang,E., Förster,T. and von der Lieth,C.-W. (2001) LINUCS: LInear Notation for Unique Description of Carbohydrate Sequences. *Carbohydr Res.*, **336**, 1–11.
16. McNaught,A. (1997) International Union of Biochemistry and Molecular Biology. Joint Commission on Biochemical Nomenclature. Nomenclature of carbohydrates. *Carbohydr. Res.*, **297**, 1–92.
17. Pappu,R., Hart,R. and Ponder,J. (1998) Analysis and application of potential energy smoothing for global optimization. *J. Phys. Chem. B*, **102**, 9725–9742.
18. Wheeler,D., Church,D., Lash,A., Leipe,D., Madden,T., Pontius,J., Schuler,G., Schriml,L., Tatusova,T., Wagner,L. *et al.* (2001) Database resources of the National Center for Biotechnology Information *Nucleic Acids Res.*, **29**, 11–16. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 13–16.