21 Schieberle, P. and Grosch, W. (1991) *Z. Lebensm. Unters. Forsch.* 192, 130–135
22 Gasser, U. and Grosch, W. (1988) *Z. Lebensm. Unters. Forsch.* 186, 489–494
23 Grosch, W., Zeiler-Hilgart, G., Cerny, C. and Guth, H. in *Progress in Flavour Precursor Studies* (Schreier, P. and Winterhalter, P., eds), pp. 329–342, Allured Publishers, Wheaton, IL, USA (in press)
24 Guth, H. and Grosch, W. *Lebensm. Wiss. Technol.* (in press)
25 Holscher, W., Vitzthum, O.G. and Steinhart, H. (1990) *Café, Cacao, Thé* 34, 205–212
26 Schieberle, P. (1991) *J. Agric. Food Chem.* 39, 1141–1144
27 Widder. S., Sen, A. and Grosch, W. (1991) *Z. Lebensm. Unters. Forsch.* 193, 32–35
28 Guth, H. and Grosch, W. (1991) *Food Sci. Technol.* 93, 335–339
29 Blank, I., Fischer, K-H. and Grosch, W. (1989) *Z. Lebensm. Unters. Forsch.* 189, 426–433
30 Blank, I., Grosch, W., Eisenreich, W., Bacher, A. and Firl, J. (1990) *Helv. Chim. Acta* 73, 1250–1257
31 Schieberle, P. (1991) *Z. Lebensm. Unters. Forsch.* 193, 558–565
32 Fischer, K-H. and Grosch, W. (1987) *Lebensm. Wiss. Technol.* 20, 233–236
33 Schieberle, P., Ofner, S. and Grosch, W. (1990) *J. Food Sci.* 55, 193–195
34 Jung, H-P., Sen, A. and Grosch, W. (1992) *Lebensm. Wiss. Technol.* 25, 55–60
35 Blank, I. and Grosch, W. (1991) *J. Food Sci.* 56, 63–67
36 Blank, I., Sen, A. and Grosch, W. (1992) *Food Chem.* 43, 337–343
37 Fischer, N. and Hammerschmidt, F-J. (1992) *Chem. Mikrobiol. Technol. Lebensm.* 14, 129–133
38 Grosch, W., Konopka, U.C. and Guth, H. (1992) in *Lipid Oxidation in Food (ACS Symp. Ser. 500)* (St Angelo, A.J., ed.), pp. 266–278, American Chemical Society
39 Guth, H. and Grosch, W. (1990) *Lebensm. Wiss. Technol.* 23, 59–65
40 Guth, H. and Grosch, W. (1990) *Lebensm. Wiss. Technol.* 23, 513–522
41 Grosch, W., Widder, S. and Sen, A. (1992) in *Aroma Production and Application* (Rothe, M. and Kruse, H-P., eds), pp. 147–154, Deutsches Institut für Ernährungsforschung
42 Konopka, U.C. and Grosch, W. (1991) *Z. Lebensm. Unters. Forsch.* 193, 123–125
43 Guth, H. and Grosch, W. (1993) *Z. Lebensm. Unters. Forsch.* 196, 22–28
44 Schieberle, P. and Grosch, W. (1989) *Z. Lebensm. Unters. Forsch.* 189, 26–31
45 Schieberle, P. and Grosch, W. (1987) *J. Agric. Food Chem.* 35, 252–257
46 Grosch, W. and Zeiler-Hilgart, G. (1992) in *Flavour Precursors. Thermal and Enzymatic Conversions (ACS Symp. Ser. 490)* (Teranishi, R., Takeoka, G.R. and Güntert, M., eds), pp. 183–192, American Chemical Society
47 Sen, A., Laskawy, G., Schieberle, P. and Grosch, W. (1991) *J. Agric. Food Chem.* 39, 757–759
48 Cerny, C. and Grosch, W. *Z. Lebensm. Unters. Forsch.* (in press)
49 Blank, I., Schieberle, P. and Grosch, W. in *Progress in Flavour Precursor Studies* (Schreier, P. and Winterhalter, P., eds), pp. 103–109, Allured Publishers, Wheaton, IL, USA (in press)
50 Sen, A., Schieberle, P. and Grosch, W. (1991) *Lebensm. Wiss. Technol.* 24, 364–369

Review

Databases of complex carbohydrates

Ready access to the wealth of data being amassed on the physical properties of complex carbohydrates is of particular interest to those involved in the study and manipulation of food carbohydrate structure and functionality. This article describes the main databases offering on-line, CD-ROM or diskette retrieval of such information.

J. Albert van Kuik and Johannes F.G. Vliegenthart

The current interest in complex carbohydrates[1–6] ranges from studying the physical properties of polysaccharides to elucidating the involvement of glycoconjugates in biological recognition. The latter feature is extremely important, since the carbohydrate moieties of glycoconjugates such as glycoproteins and glycolipids play key roles as recognition determinants in protein targeting and cell–cell interactions, and as cell-surface receptors. The carbohydrate moieties of glycoproteins are determined by the origin and type of the cell in which the proteins are expressed and, to some extent, by the structure of the protein. For the biotechnological production of recombinant glycoproteins in heterologous cell types, it is important to understand the influence that sugar chains have on the characteristics of these compounds.

The enormous structural diversity of oligosaccharides arises from the wide variety of component monosaccharide residues, each of which, in turn, can have different anomeric configurations and ring forms. These residues are connected via inter-glycosidic linkages at different positions, and can also contain non-carbohydrate substituents. The primary structures of thousands of naturally occurring carbohydrate chains (isolated from many different sources) have been reported, together with information on their biological and/or physical properties (e.g. their ability to bind to lectins or to antibodies, and conformational data).

This is an updated version of a paper first published in *Trends in Biotechnology*, Vol. 10, pp. 182–185.

J.A. van Kuik and **J.F.G. Vliegenthart** are at the Bijvoet Center, Department of Bio-Organic Chemistry, Utrecht University, PO Box 80.075, NL-3508 TB Utrecht, The Netherlands.

A database of complex carbohydrate structures

The Complex Carbohydrate Structure Database (CCSD; Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA) was created[7] to facilitate access to the enormous number of published carbohydrate structures. The CCSD aims to include the primary structures of all naturally occurring carbohydrates that are larger than disaccharides, together with references to the papers in which these structures were originally reported. The format is that of a flat-file database: each CCSD record contains the primary structure of one carbohydrate chain, one bibliographic citation, and associated text (Fig. 1). The CCSD now contains 22 300 records and 13 500 unique carbohydrate chains, but is not yet up to date in representing all the currently available information, and is still being compiled. The corresponding database-management program, CarbBank, runs on IBM-compatible microcomputers under the MS-DOS operating system. Program and data together require 40 megabytes (Mb) of disk space, and updates are released every six months. (See Box 1 for contact addresses for further information about the computer databases mentioned in this article.)

Carbohydrate structures entered into the CCSD are acquired from the literature by specialist curators, and by cooperation with Chemical Abstract Services (CAS). Data obtained from CAS have to be rigorously processed before being entered into the CCSD. This processing is done, in the first instance, by a computer program that translates and merges the CAS bibliographic data and structural data to create CCSD records. The output is then verified by the curators, who compare the records with the original literature. The large number of corrections that are made by the curators demonstrates that this verification procedure is essential in producing a high-quality database.

The CarbBank program is specifically designed to enable fast searches for branched carbohydrate

```
; start of record
; Database= CCSD5.C22
; Record #= 2866
CC: CCSD:5170
AU: Conradt HS; Ausmeier M; Dittmar KEJ; Hauser H; Lindenmaier W
TI: Secretion of glycosylated human interleukin-2 by recombinant
mammalian cell lines
JL: Carbohydr. Res. (1986) 149: 443-450
FC: 132efc31
NT: lymphokines and cytokines
BS: Human Interleukin-2, IL-2 N2
BS: recombinant Human Interleukin-2, IL-2 N2
BS: mammalian cells
KW: glycosidation
KW: Interleukin-2, IL-2 N2
VS: verified; Paulsen H
DA: 27-11-1990
CA: 105:22816j
PR: PIR1:ICHU2
---------------
structure:

                α-D-Neup5Ac-(2→6)┐
                                 │
                              D-GalNAc
                                 │
 α-D-Neup5Ac-(2→3)-β-D-Galp-(1→3)┘
=================end of record
```

Fig. 1

An example of a single CCSD record. Each record contains a carbohydrate structure, bibliographic data and additional information.

Box 1. Contact addresses, access and costs for carbohydrate databases

CarbBank and CCSD

Contact:	Scott Federhen NCBI Data Repository, National Library of Medicine, Building 38A, Room 8N-803, Bethesda, MD 20894, USA
Access:	Was distributed on diskette, and now distributed on CD-ROM
Costs:	Currently unknown, due to the transition to CD-ROM

NMR database and program

Contact:	J.A. van Kuik Department of Bio-Organic Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, PO Box 80.075, NL-3508 TB Utrecht, The Netherlands
Access:	Distributed on diskette
Costs:	Dfl. 300.00

CAS ONLINE

Contacts:

In Denmark, Finland, Ireland, The Netherlands, Norway, Sweden, UK:	The Royal Society of Chemistry, Nottingham, The University, UK NG7 2RD
In Austria, Switzerland and Germany:	Fachinformationszentrum Chemie GmbH, Postfach 12 60 50, Steinplatz 2, D-1000 Berlin 12, Germany
In Japan:	The Japan Association for International Chemical Information, Gakkai Center Building, 2-4-16 Yayoi, Bunkyo-ku, Tokyo 113, Japan
In France:	Centre National de l'Information Chimique, La Maison de la Chimie, 28 rue Saint Dominique, 75007 Paris, France
In all other countries:	Chemical Abstracts Service, Marketing Department 30586, 2540 Olentangy River Road, PO Box 3012, Columbus, OH 43210, USA
Access:	On-line
Costs:	Depends on connection time

structures to be made, taking into account that it would be extremely difficult to build a satisfying search profile from scratch (Fig. 2). To enter a carbohydrate structure that is a reliable search profile for CarbBank requires experience and a thorough understanding of the carbohydrate nomenclature. Furthermore, typing errors are easily made when a branched structure is entered: the D or L configuration, the anomeric configuration, the ring form and the non-carbohydrate substituents must be correct and complete.

The residue-complex concept

To make it feasible for inexperienced users to search for branched carbohydrate structures, CarbBank uses the 'residue-complex' concept. A residue-complex is a structural element that comprises a monosaccharide residue together with all residues that are attached to it. The program searches for all carbohydrate structures that contain this residue-complex element. The user is guided step by step in building the residue-complex: for each step, the user is prompted to choose from a list of relevant items, which is derived from all data in the database. In this way a search profile is built from only those items that really exist in the CCSD. The first selection has to be made from a general class of monosaccharides (e.g. Man). Next, a complete residue has to be selected from a second list (e.g. β-D-Man*p*). (For carbohydrate nomenclature abbreviations, see Box 2.) Subsequently, this residue is extended with linkages, from a list of linkage patterns (e.g. 1,3-linked), and finally a residue-complex can be selected, such as α-D-Man*p*-(1-3)-β-D-Man*p*-(1-4)-β-D-Glc*p*N (a simple, non-branched example). With this residue-complex, CarbBank produces a list that holds all the records from the CCSD that contain the selected structural element (a 'hitlist'). If the hitlist holds only a few structures, the easiest way to find the desired structure is to browse through it, but if the number of 'hits' is too large, supplementary searches can be made that narrow the number of branches or monosaccharide residues, or that put other constraints on the list by using different search profiles. If a structure is selected from the hitlist, a homology search can give all the citations that contain exactly that structure or structures that are closely related, such as structures that have other non-carbohydrate substituents. In a similar way, guided searches can be made for other items, such as authors' names, journal names, year of publication, words in the title of an article, words in all text fields, molecular formulae and nominal molecular mass.

CCSD records contain two other categories of information in addition to the carbohydrate structure and the bibliographical data. The first category summarizes additional details that may be present in the article, such as analytical methods, binding of the carbohydrate chain to lectins or antibodies, the position of attachment to the protein, the biological source and the biological activity. The second category provides links to other databases by giving patent numbers and accession numbers [e.g. to CAS, the International Protein Sequence Database (PIR) or the Protein Data Bank (PDB)].

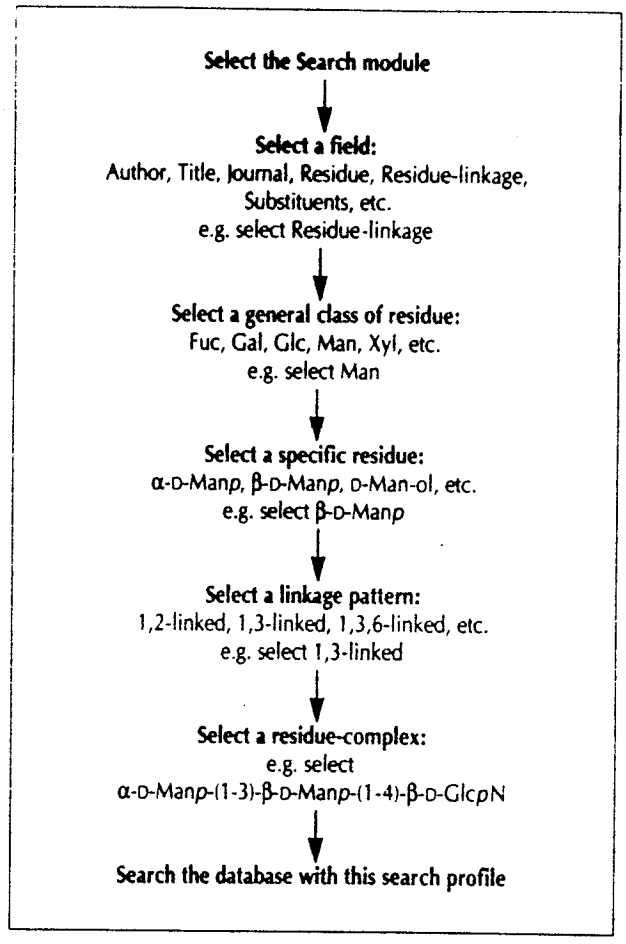


Fig. 2
The steps required to build a CarbBank search profile.

The CCSD has grown to more than 22 000 records and requires more than 45 Mb of disk space. (For this reason it is now distributed on CD-ROM by the US National Library of Medicine/National Center for Biotechnology Information).

Box 2. Abbreviations for carbohydrate nomenclature

Fuc: fucose
Gal: galactose
GalNAc: *N*-acetylgalactosamine
Glc: glucose
GlcNAc: *N*-acetylglucosamine
Man: mannose
NeuAc: *N*-acetylneuraminic acid
Xyl: xylose

α-L-Fuc*p*: α-L-fucopyranose
β-D-Gal*p*: β-D-galactopyranose
β-D-Gal*p*NAc: 2-*N*-acetyl-β-D-galactopyranose
β-D-Glc*p*: β-D-glucopyranose
β-D-Glc*p*NAc: 2-*N*-acetyl-β-D-glucopyranose
α-D-Man*p*: α-D-mannopyranose
α-D-Neu*p*5Ac: 5-*N*-acetyl-α-D-neuraminic acid
β-D-Xyl*p*: β-D-xylopyranose

***N*-linked:** Carbohydrate chain attached to a protein via the HN of Asn
***O*-linked:** Carbohydrate chain attached to a protein via the HO of Ser or Thr

An NMR spectroscopy database

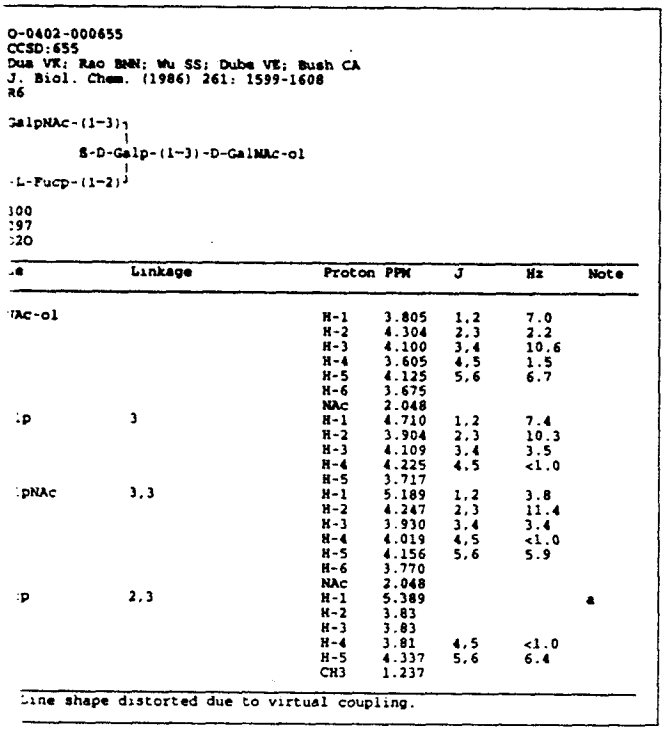
The CCSD, as a database of primary structures of carbohydrate chains, can be used to build other databases that permit access to extra information on these structures. One can imagine carbohydrate databases including fast-atom bombardment mass spectrometry (FAB–MS) or nuclear magnetic resonance (NMR) spectroscopy data, or databases containing three-dimensional information obtained from X-ray crystallography, NMR–NOE (nuclear Overhauser effect) measurements or molecular dynamics calculations. Indeed, a database that combines the structures of carbohydrates with NMR spectroscopic data has already been constructed[8] (Bijvoet Center, Department of Bio-Organic Chemistry, University of Utrecht, NL-3584 CH Utrecht, The Netherlands).

To determine the primary structure of complex carbohydrate chains, tables of ^{1}H NMR and ^{13}C NMR chemical shifts are used extensively. These tables are usually acquired from the literature[9–11]. Although review articles provide easy access to the data, they usually cover only a selected part of all the NMR data available, are neither corrected nor updated, and have to be surveyed manually. These are good reasons to store NMR tables in a computer database, and to develop a program for easy manipulation of the data. The database that has emerged uses a relational model to combine complex carbohydrate structures, bibliographic data, ^{1}H NMR tables and ^{13}C NMR tables. The carbohydrate structures and the bibliographic data are taken predominantly from the CCSD, whereas the NMR tables are collected from the literature. The database-management program runs on IBM-compatible personal computers under MS-DOS, and database and program together currently require 5 Mb of disk space. At present the NMR data set consists of 734 tables of ^{1}H NMR chemical shifts (Fig. 3), essentially corresponding to oligosaccharides derived from glycoproteins, and 258 tables of ^{13}C NMR chemical shifts, generally associated with fragments of polysaccharides.

To search for the primary structure of a carbohydrate chain within the NMR database program, the construction of a search profile is required. This profile consists of a list of chemical shift values and an optional list of monosaccharide residues that can be used to specify the 'background' of the chemical shift values that produce a hit. The structures in the database are organized by carbohydrate chain type (e.g. *N*-linked, *O*-linked or polysaccharide; see Box 2) into different sections, which can be searched separately. In addition, the minimum percentage of matching chemical shift values per structure that will result in a hit and the tolerated variation of matching chemical shift values can be defined.

After a search, the hitlist of structures and corresponding NMR tables are simultaneously displayed, such that the matching monosaccharide residues in the structures and the chemical shift values in the tables are highlighted. This presentation clearly demonstrates which part of the carbohydrate structure displayed is recognized by the search profile, and which chemical shift values are involved in this recognition. In this way, searches for known carbohydrate structures can be performed easily. A novel aspect of the program, which is a major advantage, is that it can also assist in determination of the structure of unknown carbohydrate chains by suggesting structural elements that may be part of these unknown chains.

As the number of published tables of NMR chemical shift values is growing exponentially, a computerized approach to NMR data storage and structure retrieval is essential. The growing size of the NMR database inevitably leads to porting the database to other platforms (e.g. UNIX).



```
O-0402-000655
CCSD:655
Dua VK; Rao BNN; Wu SS; Dube VE; Bush CA
J. Biol. Chem. (1986) 261: 1599-1608
R6

GalpNAc-(1-3)┐
             │
        β-D-Galp-(1-3)-D-GalNAc-ol
             │
-L-Fucp-(1-2)┘

300
297
320
```

...e	Linkage	Proton	PPM	J	Hz	Note
...Ac-ol		H-1	3.805	1,2	7.0	
		H-2	4.304	2,3	2.2	
		H-3	4.100	3,4	10.6	
		H-4	3.605	4,5	1.5	
		H-5	4.125	5,6	6.7	
		H-6	3.675			
		NAc	2.048			
...p	3	H-1	4.710	1,2	7.4	
		H-2	3.904	2,3	10.3	
		H-3	4.109	3,4	3.5	
		H-4	4.225	4,5	<1.0	
		H-5	3.717			
...pNAc	3,3	H-1	5.189	1,2	3.8	
		H-2	4.247	2,3	11.4	
		H-3	3.930	3,4	3.4	
		H-4	4.019	4,5	<1.0	
		H-5	4.156	5,6	5.9	
		H-6	3.770			
		NAc	2.048			
...p	2,3	H-1	5.389			a
		H-2	3.83			
		H-3	3.83			
		H-4	3.81	4,5	<1.0	
		H-5	4.337	5,6	6.4	
		CH3	1.237			

...ine shape distorted due to virtual coupling.

Fig. 3

...xample of a single ^{1}H NMR record from the NMR database. Each record ...tains a carbohydrate structure, bibliographic information and a ^{1}H NMR ... the header of the table, 'PPM' marks the column with the chemical shift ... of the protons (or carbons), 'J' indicates which protons are coupled, and 'Hz' indicates the values of the coupling constants.

Future prospects for carbohydrate databases

In the future, we expect to see the creation of new carbohydrate databases, such as a database of three-dimensional structures of carbohydrate chains. All such databases together will provide the carbohydrate chemist with the necessary tools to keep up to date with the growing amount of available information.

Acknowledgements

This work has been supported by grants from the EC Biotechnology Action Program BAP-0364-NL, and by the EC Biotechnology Research for Innovation, Development and Growth in Europe (BRIDGE) BIOT-CT90-0184.

erences

Paulson, J.C. (1989) *Trends Biochem. Sci.* 14, 272–275
Karlsson, K.A. (1991) *Trends Pharmacol. Sci.* 12, 265–272
Elbein, A.D. (1991) *Trends Biotechnol.* 9, 346–352
Gersow, M.J. (1991) *Trends Biotechnol.* 9, 221–225
Jentoft, N. (1990) *Trends Biochem. Sci.* 15, 291–294
Van Boeckel, C.A.A. (1986) *Recl. Trav. Chim. Pays-Bas* 105, 35–53
Doubet, S., Bock, K., Smith, D., Darvill, A. and Albersheim, P. (1989) *Trends Biochem. Sci.* 14, 475–477
van Kuik, J.A. and Vliegenthart, J.F.G. (1992) *Carbohydr. Res.* 235, 53–68
Vliegenthart, J.F.G., Dorland, L. and Van Halbeek, H. (1983) *Adv. Carbohydr. Chem. Biochem.* 41, 209–373
Bock, K. and Pedersen, C. (1983) *Adv. Carbohydr. Chem. Biochem.* 41, 27–66
Bock, K., Pedersen, C. and Pedersen, H. (1984) *Adv. Carbohydr. Chem. Biochem.* 42, 193–225