greater scope, the EBI report recommends a budget said to be some 20% greater.

Both reports more or less agree that the future of a sequence-related bioinformatics centre should be built on the back of the present EMBL data library and that an out-station would be the best solution (in order to speed up negotiations etc.). The EMBL Council have apparently accepted this recommendation and talks between the CEC and EMBL, to see what areas of complete agreement etc. exist, are expected to take place in the not too distant future. Initial discussions have been delayed because the Commission appeared reluctant to debate matters of such long-term importance until their own immediate medium term future had been discussed internally and with the various international committees etc. that control their running.

One interesting point in the PA report concerns the recommendation that the centre might not be in Heidelberg. In fact, many current users of the EMBnet service feel that unless Germany drastically improves connectivity between Heidelberg and elsewhere, another location will be required. In any case, several countries, awakening to the importance of bioinformatics, are said to be interested in hosting the out-station- centre and so tenders are to be invited and judged; again on a strict and short timetable.

The EMBL Data Library will continue its present activities under the leadership of Graham Cameron who is also to lead the group during this interim period. Much still has to be done as the PA Report still actually has to be formally reacted to by the Commission. Unfortunately, the changes within DG XII, reported in our last issue, mean that there is no clear line of authority/day-to-day control. One of the resulting confusions is that several people see the ENSC as a CEC initiative, forgetting that it was just a reaction to an external body's recommendation.

Nevertheless, there seems a good chance that both the EMBL and the Commission will recognise the need for a more stable centre in Europe. Whether this remains solely linked to sequence-related bioinformatics, or expands to cover the science as a whole, remains to be seen; but we have a long way to go before the first stage is reached anyway.■

CarbBank and the Complex Carbohydrate Structure Database

R. Stuike-Prill and K. Bock*
A. Kleen and H. Paulsen**
J.A. van Kuik and J.F.G. Vliegenthart***
S. Doubet, D. Smith and P. Albersheim****
*Department of Chemistry, Carlsberg Laboratory, Gamle Carlsberg Vej 10, DK-2500 Valby-Copenhagen, Denmark
**Institute of Organic Chemistry, University of Hamburg, Martin-Luther-King-Platz 6, W-2000 Hamburg 13, Fed. Rep. Germany
***Bijvoet Center, Department of Bio-Organic Chemistry, Utrecht University, P.O. Box 80.075, NL-3508 TB Utrecht, The Netherlands
****Complex Carbohydrate Research Center, University of Georgia, 220 Riverbend Rd., Athens, GA 30602-4712, USA

The importance of carbohydrates as an energy storage medium in cells and as structural elements in the cell walls of plants and bacteria has been known for a long time. During the last 15 years, scientists have discovered the involvement of carbohydrates in an increasing number of different biological processes. Biologically important carbohydrates, which are often part of glycoproteins or glycolipids, function frequently as recognition signals. Due to the large number of different monosaccharides that can be linked together in many different ways, the number of different oligosaccharides found in nature is enormous. Oligosaccharides that have been isolated from glycoproteins or glycolipids reflect this large variability in the primary structure of carbohydrates. The complexity of carbohydrate structures may be demonstrated by comparing glycosyl residues with amino acid residues in their ability to form larger structures from individual units. Two different amino acid residues can be linked in only two ways to create dipeptides. In contrast, eight different disaccharides can be constructed from two different glycosyl residues. This number doubles if the different anomeric configuration

(α or β) of only the non-reducing residue is taken into account. The addition of a third residue in the peptide case increases the number of permutations to only six, but adding a third glycosyl residue in the carbohydrate case means that 132 different structures can be built. These examples demonstrate how complex oligosaccharides might be. Thus, oligosaccharides can carry considerably more information than, for example, peptides containing the same number of monomers.

In 1986, at the time the *CarbBank* project was initiated, it was impossible to systematically search existing databases for carbohydrate structures having more than two glycosyl residues. Structural information on carbohydrates normally stored in databases of chemical compounds, like Chemical Abstracts, could not easily be retrieved. It was not possible to search in any of the existing databases, for example, for oligosaccharide structures attached to proteins. This unsatisfactory situation led to a workshop sponsored by the U.S. Department of Energy to explore the possibility of establishing a database of complex carbohydrates. The new database emphasizes structural data, although it does provide linkage to the literature [1].

Organization: The *CarbBank* project is coordinated by an international Board of Overseers; the design and the development of the Complex Carbohydrate Structure Database (CCSD) and a management program (*CarbBank*) was undertaken by the Complex Carbohydrate Research Center (CCRC) at the University of Georgia in Athens, Georgia, USA. Curators in 12 countries were originally responsible for submitting data. However, today most of the data are obtained from Chemical Abstract Services (CAS) through a collaborative agreement. A database that combines the structures and bibliographic information contained in the CCSD with NMR spectroscopic data has been created as well [2,3].

Funding: In Europe the *CarbBank* project is funded by the EEC inside the BRIDGE project. The U.S. Department of Energy (DOE), the National Institutes of Health, the National Science Foundation, and the Department of Agriculture support the project in the USA.

Description of the Complex Carbohydrate Structure Database

The Complex Carbohydrate Structure Database (CCSD) is organized in a flat file format. Each CCSD record contains the complete primary structure of the oligosaccharide, bibliographic information, and additional text, such as keywords, biological activity, analytical methods used, etc. (*Fig. 1*).

As already mentioned, most of the data are obtained from Chemical Abstract Services and incorporated into the CCSD after further processing and verification. Each record is verified against the original literature, and additional information not included in the data from the CAS file are added. These are mostly keywords, analytical methods used to elucidate the carbohydrate structure, information about the biological source and activity, and information on possible aglycons. Many of the carbohydrate structures are isolated from glycoproteins that often carry more than one oligosaccharide chain. It is important, therefore, to know the amino acid residue to which the carbohydrate was attached. Cross-references to a corresponding entry in a protein database and the CAS are given when available.

The CCSD version that was released at the end of 1992 contains about 22000 records. Most of the published literature on carbohydrates up to 1990 has been included. For the years 1991 and 1992, about 2000 records are being processed. The reason for this backlog is that carbohydrate literature from the past had to be processed first. This task is almost completed, so the database should be more up to date in the future, i.e., only about six months to one year behind the current literature. Currently, an update of the database is released every six months.

Description of the *CarbBank* program

The database management program *CarbBank* enables the CCSD to be searched for whole structures, substructures and text information. Records can be edited, imported and exported in various formats, and different kinds of reports can be created. The program is written in Pascal and runs on IBM-compatible personal computers under MS-DOS. Currently, the *CarbBank* program is ported to C, so that it can be used on more powerful computers like UNIX workstations that use the X windows system graphics user interface. An interface to the *Entrez* retrieval software developed at the National Center for Biotechnology Information (NCBI; at the National Institutes of Health, Bethesda, MD, USA) is also under development.

The CarbBank program offers several modules in order to manage the CCSD. The **Search** module: A search of bibliographic data or

structural elements can be performed in a variety of ways. The user may enter the search criteria manually or select search items from lists of all the items that are contained in the database. By combining several search items from a list, a complicated search profile can easily be generated. If one searches for a larger structure, it might be feasible to use a similar structure included in the database as a probe that can be manually edited to obtain the structure in question. These methods reduce the possibilities for spelling and format mistakes and eliminate searches for structures or text files that do not exist in the database. The search from indexed text items and structural objects assures that searches are efficient and fast.

The **Import/Export** facilities in *CarbBank* enable the user to export the records of the CCSD as ASCII text files in an ASN.1 format, a standard that is also used by other databases and in the internal CCSD format (e.g. to create subdatabases). Records can be imported from one of these three formats as well. The **Edit** facility allows changes to be made in the database entries and new records to be entered.

The **Report** module: The contents of the database can be analyzed using the Report module, as was done to generate the statistical evaluation presented below.

Some statistics on the contents of the CCSD
The latest version of the CCSD has a total of 22333 records with 10893 unique structures. A few examples are given below of the number of entries in the database that fulfill a particular search profile.

The search for carbohydrate structures which are N©linked to a glycoprotein via an asparagine results in 1343 structures. About 545 of these structures are unique. The second major class of oligosaccharides are linked to glycoproteins via an O-glycosidic linkage to serine or threonine. About 872 structures fulfill this criterium, of which 406 structures are unique.

The branching trimannoside, α-D-Man-(1-3)-[α-D-Man-(1–6)-]β-D-Man, which is common to oligo-saccharides of the complex type, occurs in 4274 structures as an structural element; 1715 of these structures are unique.
Records included in the database have been published in 810 different journals.
Fifteen of the most frequently cited journals are listed in *Table 1.*
These journals cover about 62 % of the entries contained in CCSD.

The years in which publications include carbohydrate structures are 1941 through the present. The distribution of the records according to the publication year is shown in *Table 2.*

The structures present in the database contain from 1 to 43 residues. Occurrences of structures with a specified number of residues per structure are shown in *Table 3.*

The complexity of the structures in the CCSD is further indicated by the occurrence of highly branched carbohydrates, as shown in *Table 4.*.

Distribution
Since the fall of 1992, the CCSD and the *CarbBank* program has been distributed on a CD-ROM by the NCBI. The CD-ROM can be ordered for a handling fee. Users without access to a CD-ROM drive can also receive the database and program on floppy disks; the CCSD and *CarbBank* described above use about 50 Mbyte of disk space.

Information about ordering *CarbBank* can be obtained from the authors or directly from:
Ms. Dana Smith
UGA/CarbBank/CCSD
114 West Magnolia Avenue, Suite 305
Bellingham, WA 98225
USA
Telephone: 206-733-7183
Fax: 206-733-7283
E-Mail: Internet:
76424.1122@Compuserve.com
or:
CarbBank@UGA.bitnet. ■

References:

[1] S. Doubet, K. Bock, D. Smith, A. Darvill and P. Albersheim, *Trends Biochem. Sci.* **14** (1989) 475-477.

[2] J.A. van Kuik and J.F.G. Vliegenthart, *Trends Biotechnol.* **10** (1992) 182-185.

[3] .A. van Kuik, K. Hård and J.F.G. Vliegenthart, *Carbohydr. Res.* **235** (1992) 53-68.