ELSEVIER

Carbohydrate Research 290 (1996) C1–C5

CARBOHYDRATE RESEARCH

Announcement

Report on a Workshop to Establish Databases of Carbohydrate Spectra

A workshop to formulate guidelines to establish databases of carbohydrate spectra was held August 18–20, 1995, in conjunction with the XIII International Symposium on Glycoconjugates in Seattle. The workshop was sponsored by the U.S. Department of Energy, Division of Energy Biosciences (grant DE-FG-5-95ER20167) and organized by the Complex Carbohydrate Research Center of the University of Georgia (USA) and the Department of Bio-Organic Chemistry of the University of Utrecht (The Netherlands). A total of 23 scientists from around the world participated.

The workshop was organized to answer the need for carbohydrate spectral databases. Carbohydrate chemists are frequently forced to expend research effort and funds to duplicate the structural characterization of previously characterized complex carbohydrates because of the absence of spectral databases that could alert them beforehand to the fact that they are analyzing a known molecule. The workshop participants sought to initiate centralized collections of mass and nuclear magnetic resonance (NMR) spectra as a foundation for their subsequent development into searchable databases. Then scientists will need only to query the databases with the spectrum or with information defining the spectrum of an unidentified carbohydrate to find out if it has been previously characterized.

The great interest in complex carbohydrates has led to a rapid increase in the number of published complex carbohydrate structures, most of which are contained in the Complex Carbohydrate Structure Database (CCSD), a database organized by a 1986 U.S. Department of Energy, Division of Energy Biosciences-sponsored workshop. The CCSD (searchable by CarbBank) is designed to contain the structures of all published complex carbohydrates together with text files that provide the citation of the publication containing the structural characterization, the biological source of the carbohydrate, and, for example, the associated protein and point of attachment of the carbohydrate when the carbohydrate is part of a glycoprotein. Many of the structures in the CCSD have been reported in the literature multiple times, and each such report is a separate, cross-referenced record in the CCSD. The latest release of the CCSD contains more than 42000 records, each assigned a unique accession number. The workshop participants recognized that the task of developing spectral databases will be simplified by attaching the CCSD accession number to each spectrum so that the analyst can easily access the structure and associated text information of the complex carbohydrate that generated the spectrum.

In 1995 alone, 15000 to 20000 records will be published, which is about four times as many as were published in 1991. However, most of the structures of complex carbohydrates published today are identical to structures already contained in the CCSD that were isolated from other organisms, tissues, cells, or glycoconjugates. Since each substance entered into the CCSD is assigned its own accession number, a single oligosaccharide has as many accession numbers as the number of times the oligosaccharide is listed in the database. To facilitate the development of collections of spectra, the CarbBank/CCSD administrators agreed to attach an identical substance identifier number to all entries of a substance in the CCSD. The substance identifier number will provide the user with all CCSD records of that substance.

The workshop participants agreed that a separate collection of spectra should be formed for each class of complex carbohydrate. Thus, the spectra of *N*-linked carbohydrate side chains of glycoproteins will be stored in one collection and the spectra of *O*-linked (serine/threonine-linked) side chains in another. The spectra of glycolipids, glycophosphoinositols, glycosaminoglycans, polysaccharides, lipopolysaccharides, saponins, etc., will each be located in a unique collection. Partitioning in this manner, in addition to reducing the number of spectra in each collection, will enable the databases to be managed by curators who are experts on the spectra in their collections (see below).

The collections will be further divided into those that contain mass spectra and those that contain NMR spectra, again making the responsibilities of curators more manageable. However, any one complex carbohydrate collection may have a variety of types of mass spectra or NMR spectra within it. The mass spectra might include fast-atom bombardment, electrospray, matrix-assisted laser-desorption, electron-impact, and/or chemical-ionization spectrometry. Separate collections or databases will be organized for each type of low molecular weight compound, such as the partially methylated alditol acetates that are widely used to determine glycosyl linkage compositions. Standard protocols will be distributed (see below) for acquiring mass spectra of low molecular weight carbohydrates so that the spectra of samples being analyzed can be compared more readily to the spectra in the databases.

Mass spectra of large complex carbohydrates are valuable but less definitive than the spectra of smaller carbohydrates because slight variations in experimental conditions can cause large variations in the spectra of larger molecules. Furthermore, the structures of large molecular weight carbohydrates are often deduced from the mass spectra of oligosaccharide fragments of the parent molecule. Whenever possible, all of these mass spectra will be included in the collections. Since it is not yet possible to define protocols for acquiring mass spectra of larger complex carbohydrates, in order for their spectra to be of value, the details of the instrument settings and experimental conditions used to acquire each spectrum must accompany it. Guidelines for the information needed in these 'data sheets' will be made available on the CHO-DATA listserv (see below). The text information accompanying a spectrum will include the CCSD substance identifier number to provide access to the structure of the molecule, its calculated molecular weight, and appropriate literature citations.

NMR spectra will include both proton (^{1}H) and carbon-13 (^{13}C) when available. ^{1}H NMR spectra will be acquired at field strengths of 500, 600, 750, or 800 MHz. The

identity and chemical shift (offset) of the reference compound used to acquire the ^{13}C and ^{1}H NMR spectra must be included, as well as the temperature, solvent, pH, and instrument settings used (standardized as much as possible). The most useful conditions vis-à-vis the data collections for obtaining one-dimensional ^{1}H NMR spectra will be agreed upon by the appropriate collection curators and distributed. Spectra of new complex carbohydrates will be entered into the collections only following acceptance for publication of a refereed paper describing the structure.

Professor Vliegenthart and colleagues in Utrecht have developed Sugabase, a large, useful database based on the structural-reporter-group concept of well-defined chemical shifts of ^{1}H and ^{13}C NMR signals obtained from the spectra of various types of complex carbohydrates. The chemical shifts listed in Sugabase are sufficient in many instances to identify the complex carbohydrate. Sugabase can be downloaded from the Internet and then queried by presenting it with a list of the chemical shifts of an unidentified complex carbohydrate to ascertain whether Sugabase contains information to identify a complex carbohydrate with the same chemical shifts. Although most NMR spectra will be exchanged and transmitted as free induction decays (FIDs) in a to-be-agreed-upon digital format, the workshop participants agreed that all available NMR spectra of complex carbohydrates should be submitted to Sugabase as lists of the chemical shifts of identified signals.

Several other collections of carbohydrate spectra have been started at other locations, including ^{1}H and ^{13}C NMR spectra of bacterial polysaccharides, ^{1}H NMR and mass spectra of *N*- and *O*-linked carbohydrates, electron-impact mass spectra of partially methylated alditol acetates and of partially methylated anhydro alditol acetates, NMR and mass spectra of glycolipids, the spectra of glycophosphoinositols, and saponin spectra. These collections will be utilized in forming FTP (file transfer protocol) sites of specialized collections. The workshop participants agreed on the following procedures to 'grow' these data collections into useful databases.

1. The group recommended that an Internet listserv be organized to handle e-mail communications between those wishing to establish and utilize collections of carbohydrate spectra. Dr. William York (will@mondl.ccrc.uga.edu) has set up the CHO-DATA listserv that scientists can join by sending an e-mail message to:

listserv@uga.cc.uga.edu

with the body of the message containing the command:

sub CHO-DATA ⟨first-name⟩ ⟨last-name⟩

2. Carbohydrate spectra data collections will be organized by volunteer curators (see list below) at FTP sites on the Internet. The sites will be connected electronically to form an interactive distributed collection of carbohydrate spectra. The FTP sites are identified on the CHO-DATA listserv.

3. All carbohydrate spectra will be made freely available to anyone who wishes to download them from the FTP site. Each time a spectrum from a collection is used to identify a carbohydrate and the result is published, the author is expected to cite the collection that provided the spectrum and the original publication in which the spectrum was characterized.

Table 1
Organization for the collection of spectra

Type of complex carbohydrate	Type of spectra	Curator
Various	NMR	Vliegenthart, Van Kuik (Utrecht)
N-Linked chains	NMR	Van Halbeek (CCRC)
Xyloglucan	Mass, NMR	York (CCRC)
Methylated alditol acetates	Mass, NMR	Valafar (CCRC)
O-Linked chains	NMR	To be selected
Larger carbohydrates	Mass	Costello (Boston), Dell (London)
Glycophosphoinositols	Mass, NMR	Homans (St. Andrews)
Glycosphingolipids	Mass, NMR	Peter-Katalinic (Bonn)
Bacterial polysaccharides	Mass, NMR	Jansson (Stockholm) and Brisson (Ottawa)
Saponins	Mass, NMR	Tadahiro Takeda (Tokyo)
Glycosaminoglycans	Mass, NMR	to be selected

4. The spectral collections will be linked to the CCSD by attaching to each spectrum the appropriate CCSD substance identifier number.

5. The collections of spectra will constitute a foundation for the formation of databases with easy-to-use search engines that utilize appropriate semantics and syntax. It is hoped that one or more databases will be formed from each collection of spectra. The databases may be formed by academic or commercial groups. Those who develop databases from the collections of spectra may charge a fee for use of the databases if they wish.

6. An international Board of Overseers is being organized to stimulate the organization of collections of spectra by attracting and coordinating curators, recommending standards, and encouraging the conversion of the data collections into databases.

7. The Chair of the Board of Overseers will conduct the affairs of the Board with the assistance of a small, executive committee which the Chair will select from members of the Board of Overseers. Hans Vliegenthart will chair the Board of Overseers.

8. The workshop participants agreed to initiate FTP sites containing collections of spectra as described in Table 1. (Persons who have already volunteered to manage particular collections are indicated in the table.) After establishment of the FTP sites, carbohydrate scientists will be urged to submit their published spectra to the appropriate FTP sites using standardized protocols available at each site. The protocols will also be published on the CHO-DATA listserv.

The databases developed as a result of this workshop will accelerate progress in glycoscience by making knowledge of the structural biology of complex carbohydrates more accessible and by opening up the structural characterization of complex carbohydrates to analysts with little formal training in carbohydrate chemistry.

Submitted by:

Peter Albersheim
Complex Carbohydrate Research Center
University of Georgia

220 Riverbend Road
Athens, GA 30602-4712
USA
Tel: +1-706-542-4404
E-mail: palbersh@mondl.ccrc.uga.edu

Hans Vliegenthart
Department of Bio-Organic Chemistry
Bijvoet Center for Biomolecular Research
Utrecht University
P.O. Box 80.075
NL-3508 TB Utrecht
The Netherlands
Tel: +31-30-2532184
E-mail: vlieg@accucx.cc.ruu.nl