DOE/ER/13970--T3

The University of Georgia

Complex Carbohydrate Research Center

FACULTY:
Peter Albersheim, *Director*
706-542-4404
Alan Darvill, *Director*
706-542-4411
Karl-Erik L. Eriksson*
706-542-4453
Michael G. Hahn
706-542-4457
Bernd Meyer
706-542-4454
Debra Mohnen
706-542-4458
Kelley W. Moremen*
706-542-1705
Ron Orlando
706-542-4429
Michael Pierce*
706-542-1702
Herman van Halbeek
Associate Director-Instruction
706-542-4438

STAFF:
Russell W. Carlson
Technical Director (Plants and Microbes)
706-542-4439
Scott Doubet
CarbBank Director
206-733-7183
Roberta K. Merkle
Technical Director (Biomedical)
706-542-4441
Rosemary C. Nuri
Administrative Manager
706-542-4403
Lydia J. Snyder
Business Manager
706-542-4408

June 3, 1993

Dr. Robert Rabson
Division of Energy Biosciences
Office of Basic Energy Research, ER-17
Department of Energy
Washington, DC 20545

Dear Bob:

Thank you for organizing the May 14 meeting at NCBI to discuss the CarbBank/CCSD situation. We greatly appreciated the warm reception and concern for CarbBank and the CCSD expressed by you and the other representatives of the granting agencies.

The CarbBank project got off to an impressive start in the second phase of its development with the advent of intergovernmental agency support and the agreement with the Chemical Abstracts Service (CAS) to supply the project with the structures and citations for all carbohydrates containing more than two glycosyl residues. With the data from CAS, some help from the European curators, and the hard work of Scott Doubet and Dana Smith, the process of entering and verifying the structures and citations worked well. Over 18,000 records were added to the Complex Carbohydrate Structure Database (CCSD), bringing the total records in the CCSD to over 24,000, and virtually eliminating the backlog of records that existed prior to September 1991.

The success of CarbBank and the CCSD was evident. I include a list of the industrial, academic, and government locations that have received CarbBank directly from Scott and Dana. The number of individuals using CarbBank at each location is unknown. CarbBank and the CCSD are supposed to be distributed twice a year on the NCBI Repository CD-ROM. The second distribution, originally scheduled for October 1992, is imminent. The director of the NCBI Repository informed Scott at the beginning of this year that the NCBI had received over 500 requests for the Repository, and that about 70% of those requests were generated by an interest in CarbBank. We have no information on the number of individuals who have obtained CarbBank and the CCSD from NCBI via Internet.

The CAS data amounted to, on average, about 2,000 records per year for the years 1985 through 1990, and would now, if functioning, be providing about 3,000 records per year. Another 500 or so records per year were being found by Scott and Dana and the curators. We believe that an additional 500 or so records per year were overlooked by these surveillance procedures. In other words, about 85-90% of the published carbohydrate structures were being entered into the CCSD.

CCRC Building: 220 Riverbend Road, Athens, Georgia 30602-4712 USA
CCRC Telephone: 706-542-4401 Facsimile: 706-542-4412
*Located in Life Sciences Bldg.: c/o Biochemistry Dept., Life Sciences Bldg., Athens, Georgia 30602-7229 USA

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

CAS has undergone a change in management. The new management is unwilling to abide by the agreement with the CCRC. CAS has now decided to send us only those carbohydrate structures that have not been previously cited. This situation would be equivalent to the gene sequence database including only the first citation of a gene encoding a protein kinase and missing all the additional protein kinase genes found in the same and additional organisms. If the same asparagine-linked oligosaccharide is found at different places on the same protein, on different proteins in the same cell or organism, or even on proteins in different organisms, the users of CarbBank would not be informed, and users have no alternative way to search the literature! There would be no way, for example, to compare carbohydrates in different plants or between plants, microbes, and animals.

The transfer rate from CAS has been reduced from 3,000 to 600 records per year, or about 20% of all published carbohydrate structures and citations. Scott and Dana, with some help from the European curators, are verifying these citations and adding another 10-15%. In other words, only about 30% of the appropriate citations are now being added to the CCSD.

The impact of the loss of the CAS data comes at the same time that it has become apparent that we need to use Scott Doubet's time for software development rather than for entering structures into the database. We believe it was appropriate for Scott to spend much of his time building the database when we needed to reduce the large number of backlog records received from CAS. On the other hand, it is now increasingly important that Scott spend all his time, or as much of his time as possible, on software development. We have received a large number of requests for access to CarbBank and the CCSD on computer platforms other than PC-DOS. The MAC, UNIX, and NEC people all want it; so do others. Clearly, the way to achieve this multi-platform accessibility is to form a close collaboration with NCBI that provides nucleotide and amino acid sequences to all major platforms.

It is also important that the CCSD function as an integral part of NCBI's protein sequence and gene sequence databases. The demand for such a resource is already large, and it will grow rapdily. It became apparent at our meeting at NCBI on May 14 that the way to achieve this is to see to it that CarbBank/CCSD function as a part of the Internet version of the nucleotide and amino acid sequence databases. This would avoid the problem of intermittent distribution of CarbBank on the Repository CD, as well as the problems of operating such a large system on PCs. What is clearly needed is a way to search the database while it resides on Internet, so that the system does not have to be downloaded as is required now when using the CCSD available through the Repository/Internet.

The development of graphical interface search routines is another advantage that will benefit users when CarbBank/CCSD are adapted to the NCBI system. This interface will make it easier for users to find what they want as the size of the CCSD grows.

All parties at the NCBI meeting agreed that making CarbBank and the CCSD compatible with the NCBI database system must have the highest priority for Scott's time. In order to comply with this priority, Scott will no longer be able to abstract data to build the database, as his efforts will be devoted to reprogramming CarbBank to form a seamless system with NCBI's other biosequence database software. Success in this endeavor will mean that CarbBank and the CCSD will be accessible to a wide range of users on a variety of computer platforms, ensuring that CarbBank and the CCSD become an indispensable tool used by those interested in biosequence information. Researchers interested in the medical aspects of glycoproteins, for example, would be able to gain easy access to information about both the carbohydrate and protein portions, as well as any known functional aspects of the molecules, by using what would appear to the user to be a single program and data source on Internet. Although it will require considerable effort to make the CarbBank/CCSD software compatible with the NCBI system, we are convinced these efforts are critical to the future success of CarbBank and to its acceptance and usefulness as a major database in the research community.

The present funding of CarbBank/CCSD ($225,000) covers only the salaries of Scott Doubet and Dana Smith (the database manager), and their operating expenses. We were able to carry over some funds from the previous grant to hire chemist Nancy Shough half time. Nancy spends all of her time abstracting records from the literature. The funds supporting Nancy will run out in September, the end of the second year of the present five-year grant. Nancy's salary was supposed to be covered by the $25,000 that was, at the last minute, cut from the CarbBank budget.

We estimate that we will require, in addition to Scott and Dana, the equivalent of two full-time individuals to enter all the carbohydrate structures and citations in the CCSD. However, as we agreed at the May 14 meeting, we are willing to try going ahead with one full-time person in addition to Nancy, Scott, and Dana, with Nancy working half-time. We believe that with this help we will be able to find and enter, next year, ~70% of the current records. Perhaps, we can do even better.

A $250,000 budget for year three and a $100,000 supplemental request for year three are attached. Supplemental funds are requested for one additional full-time chemist/abstracter with necessary office space, phone, computer, and supplies. The supplemental funds requested also include an additional hard disk drive to store computer programs and data, and computer program development software (PC-DOS, Macintosh, and UNIX compilers and program development tools) to create CarbBank/CCSD software compatible with the NCBI mutli-platform system. We also

request supplemental funds for telephone and network access to NCBI during program development and for printing and mailing the hard copy CarbBank/CCSD manual to 500 subscribers.

Thank you ever so much for your continued support.

Sincerely yours,

Peter Albersheim
Executive Director of CarbBank

PA:ksh

cc: Alan Darvill
Scott Doubet
Joe L. Key

Enclosures