April 21, 1994

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

Dear Bob,

Thank you for encouraging and attending the meeting with Drs. Klaus Bock, Hans Vliegenthart, and Hans Paulsen in San Diego on March 16. We reached conclusions on how to productively continue the CarbBank Project as it undergoes several modifications.

As you know, the participants of a 1986 U.S. Department of Energy workshop determined that there was a critical need for a public domain database of complex carbohydrate sequences for glycobiologists, carbohydrate chemists, and others interested in complex carbohydrate research. The participants were aware that the Chemical Abstracts Services (CAS) database of complex carbohydrate substances was large, but expensive and cumbersome to use. As a result, an international partnership of U.S., European, and Japanese scientists was given the task of developing, building, and widely distributing a computer database called the Complex Carbohydrate Structure Database (CCSD) and CCSD access software, called CarbBank.

Note: The term "CCSD/CarbBank software", used throughout this document, refers to both the CCSD and CarbBank database access software.

This report summarizes our progress and requests additional resources for the CarbBank Project as the carbohydrate field grows larger and increases in importance to the scientific community.

CarbBank Project Status.

A nearly complete database covering articles published up to 1993 has been built by the U.S., European, and (to a small extent) Japanese CarbBank partners. The CCSD contains over 29,200 records (from 7,000 articles) and is growing at a rate of over 3,300 records (derived from over 1,000 articles) per year. The CCSD now has over 11,000 unique structures, far surpassing the original (1986) projection of 4,000 - 5,000 unique structures.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
The CCSD/CarbBank software, created for IBM PC-compatibles, has been well received. CCSD/CarbBank software and operation manuals have been released on CD-ROM discs by the National Center for Biotechnology Information (NCBI) and by the Protein Information Resource (PIR). NCBI has sent out its final Repository CD-ROM (January 1994 issue); however, PIR will continue to send out quarterly releases of its CD-ROM.

NCBI will continue to post CCSD/CarbBank software on its Internet file server.

We expect that continuous availability on the Internet and quarterly releases on CD-ROM discs will allow our subscriber base to grow beyond its present level of 600 subscribers. Of course, each CD-ROM may be used by several researchers. For example, Bayer AG (in Leverkusen, Germany), Glycomed (in Alameda, California), and others have placed CarbBank on their local area networks to accommodate multiple users.

Building the CCSD.

Although the idea of the CCSD was readily accepted by the scientific community, they requested that the CCSD be built quickly, that the literature be covered thoroughly, and that links between the CCSD and other biosequence data bases be established.

Our original model of data acquisition, in which volunteer curators sent reprints to the CarbBank office for data entry, was too slow and inefficient. Our European partners concluded that the CCSD could be built quickly if CAS would provide the data from its large commercially available database. Negotiations with CAS to exchange its data for CarbBank technology were completed in 1991. As a result, we agreed to supply CAS with the CarbBank source code, the CASON program (a program that could translate CAS data to a CCSD format), consulting services, and the CCSD. In return, CAS agreed to supply the CarbBank Project with accumulated data containing the backlog of oligosaccharide structures (trisaccharides and larger). CAS also agreed to supply CarbBank with new data for five years (1991 through 1996).

CAS made the initial transfer of a large amount of data in 1991. The CASON program was used to re-format CAS data for direct inclusion in the CCSD. Unfortunately, the CAS data set contained many errors, and the U.S. and European partners were forced to compare every record in the CAS data set with original citations (when available) and to make corrections and additions.

CAS continued to send data from newly abstracted publications to the U.S. team; unfortunately, data sets were found to be incomplete. CAS missed (or did not transfer) most of the structures in newly published citations, and they often missed entire citations.

CAS transferred its last complete data set for the publication year 1990. CAS alerted us to 600 of the 750 unique citations (now present in the CCSD) published in 1990. The situation deteriorated dramatically, and by 1993, CAS alerted us to only 150 unique citations.

The number of records that CAS transferred to us dropped steadily at a time when the actual number of articles being published was growing. To make up for the
incompleteness of CAS data, the U.S. team began to search through ISI Current Contents on diskette (1992 to present). The U.S. team also subscribed to and searched through selected carbohydrate journals to find citations containing complex carbohydrate structures. As a result of these actions, the CCSD contains 150 CAS-assisted articles and 930 non-CAS-assisted articles published in 1993.

Clearly, we can find articles that contain complex carbohydrate structures, without help from CAS. Our European (and eventually our Japanese) partners have agreed to work with us to cover as much of the literature as possible. The Europeans will use additional document “alert” services, such as Chemical Abstracts Selects. They have also agreed to scan selected journals for pertinent articles.

The process of building the database is more than just finding articles. Each article must be abstracted, and information must be transferred to CCSD records. This is the most technically challenging aspect in the database building process. The abstractor must be familiar with carbohydrate science, molecular biology, and chemical nomenclature. The abstractor must also be able to find structures in articles even when structures appear in a variety of formats.

Our customers have told us that providing structures, authors, and citation information is not enough. To make the database as useful as possible, each record must be annotated. The person abstracting each article must specify, when appropriate, the biological source of the substance, protein attachment site, binding information (e.g., to receptors, antibodies, and lectins), non-carbohydrate components, cross-references to other databases (e.g., points of attachment to proteins), etc. The annotation process is not an attempt to re-create the article; rather, it provides essential descriptive information to database users and links CCSD entries to other databases.

We have developed a system for abstracting articles and entering records at the Bellingham location. Document information (title, authors, and citation) is abstracted by one person. Structural information, keywords, and cross-references are abstracted by a second person. We use software to check that structures have been entered correctly and to obtain controlled lists of keywords and cross-references. Finally, we randomly select and check a fraction of the new records. If we had additional personnel, a higher percentage of entries would be re-checked before merging with the main CCSD.

We have also developed a system that will allow our European and Japanese partners to participate in the abstraction process. The partners will each be responsible for finding articles from a specified set of journals. Each article that contains a carbohydrate structure will be completely and clearly abstracted in a hard-copy format and sent to the data entry person at the Bellingham office. After each record is typed into a file, it is returned by electronic mail to the original abstractor for verification and if necessary, correction. Finally, records will be merged with the main database in Bellingham.

We expect this system to permit the entry of over 3,600 records located in 1,200 articles in 1994.
Software Enhancement.

We are working with NCBI and PIR to integrate the CarbBank Project with other biosequence projects. Advances in computer technology, dramatic growth in the size of databases, and a high level of cross-discipline interest are forcing database producers to make fundamental changes. These changes must be made as we contend with an accelerating flow of published sequence data.

The easy changes have already been made. Frequently updated repositories and access through advanced computer networks are now available.

The hard challenges remain. If CarbBank and the CCSD are to be functionally integrated with other biosequence databases, we must re-define the structure of the CCSD and completely re-write CarbBank.

We have already begun to change the CCSD into a database that will be accessible to a wide range of computer programs. We are working with NCBI to re-cast the CCSD in a new format called Abstract Syntax Notation 1 (ASN.1), a standard format for open databases. We foresee no major problems in this transition and expect to finish this process by the end of 1994.

Re-writing the CarbBank computer program is a major undertaking. The present version is a stand-alone, DOS-Text program. When compared to modern "Windows-like" programs produced by commercial vendors, CarbBank can only be described as an archaic relic from a previous age, tolerable only because it has no competition. We are using modern software tools and techniques available from both NCBI and commercial vendors to re-write CarbBank. When we finish, CarbBank will be network-aware, ASN.1-compliant, and available for a wide variety of computer environments.

The modernized CarbBank will run on the most popular computer platforms, will be used more widely than the present version, and will be easier to use, since it will incorporate elements from the latest industry standard, graphical user interfaces.

The new version will also be more powerful. CarbBank users will be able to display CCSD records, as they do now, and they will be able to easily access and view citation abstracts, structures, and entire texts of citations found in other databases. We are working to make connections from CarbBank into Medline (from the National Library of Medicine), GenInfo (NCBI), and ATLAS (PIR). The reverse process will also be possible. Software currently being developed elsewhere (NCBI, PIR, CAS, Oxford Glycosystems, and others) will have access to CarbBank methods and to CCSD data.

Academic and industrial users of CCSD/CarbBank software will benefit from these enhancements. For example, manufacturers of therapeutic glycotechnology products must maintain certification. To do so, they must control the quality and verify the identity of their products. When dealing with recombinant glycoproteins, manufacturers must be aware of the primary amino acid sequence (available through amino acid sequence databases) and the locations and identities of carbohydrate side chains (available through the annotation fields of CCSD records).

Building ASN.1 compliance and network awareness into CarbBank and constructing an ASN.1 version of the CCSD will guarantee that the CCSD will be used widely and have an essential place in the future of biotechnology.
Personnel Needs.

Overall, the project needs a staff of at least six full-time workers, including one programmer, one keyboarder, and four abstractors. We have two and one-half full-time equivalents at Bellingham, and we are asking for a total of four full-time equivalents in this renewal. This will require that the Europeans and the Japanese together contribute two full-time abstractors. We place a high value on their contribution; however, we are concerned that their contributions will remain in doubt until they stop using temporary, part-time postdoctoral fellows and establish data processing centers staffed by permanent specialists.

Even with European and Japanese contributions, the Bellingham office needs one and one-half additional positions to do a minimally acceptable job. Scott has been spending nearly all of his time over the past few years building the database. Even with Scott's help, over 1,200 recently published citations, representing a potential 3,600 records, have accumulated in Bellingham and await processing. Scott will now have to spend the majority of his time programming software, writing manuals, and coordinating the abstractors.

Without increased staff, the number of records entered into the database at the U.S. site will probably not exceed 1,700 records per year. Thus, the backlog of un-entered data will grow more rapidly. Citations will continue to be acquired and stored until additional people are made available. In an effort to enter the structures that are most valuable as links to other databases, we propose to give the highest priority for citations that deal with glycoprotein oligosaccharides (N-linked and O-linked glycoprotein side chains, and GPI-anchor structures). We will place the highest priority on any abstracted articles that the Europeans and Japanese send to us. We will also process citations from the journals to which we subscribe, namely, Carbohydrate Research, Glycoconjugate Journal, and Glycobidogy. Citations that deal with glycolipids, as well as plant, bacterial, and fungal oligosaccharides will be processed as time allows. Citations containing saponins, steroids, glycosides, aminocyclitols, glycosaminoglycans, and antibiotics will not be processed until the necessary personnel are available.


Funds are requested to support further development of CarbBank software and the CCSD. Our budget request of $365,000 represents an increase of $125,000 over last year to cover the requested one and one-half positions. The new budget covers overhead (at an "off-campus" rate of 26% of modified total direct costs), office and travel expenses, and salaries for four full-time positions.

Funds are requested for the following personnel:

(1) Executive Director of CarbBank, Peter Albersheim (5% time). Responsibilities include overall responsibility for CarbBank and the CCSD including interaction with granting agencies and the CarbBank Board of Overseers, and coordination of CarbBank with government agencies and corporations.

(2) CarbBank Director/Programmer, Scott Doubet (100% time). The major responsibilities of this position are computer program development (70%), building the
database and writing program documentation (15%), and administration and coordination with European and Japanese partners (15%).

(3) Database Manager, Dana Smith (100% time). The major responsibilities of this position are article acquisition, data entry, and user support.

(4) Data Entry Specialists (two full-time positions). Nancy Shough (50% time), persons to be named (50% and 100% time). The responsibilities of these positions are to abstract articles and enter data.

(5) A part-time Accountant (5% time) to administer the grant at the University of Georgia.

Funds are requested to cover the cost of office rent and equipment for the CarbBank staff. Funds are requested for computers and a printer to be used by the data entry specialists and to replace existing, failing equipment. Funds are requested for acquiring new (or buying upgrades of existing) software used to generate programs, write manuals, and maintain customer databases. Funds are also requested to cover the cost of acquiring articles for abstraction, including requests to authors, subscriptions to journals, and citation alert services. Funds are requested to cover the cost of communicating with our European and Japanese partners and with the data distribution vendors. Funds are requested to cover the costs of printing and mailing manuals and communicating with customers. Travel funds are requested to allow the CarbBank Director to attend international symposia coinciding with meetings of the CarbBank Board of Overseers and with the European and Japanese data entry specialists.

Summary.

We have created a useful database in an exciting and important area of science. The importance of the carbohydrate database and the publication rate in the carbohydrate sciences are growing rapidly. The problem that we face is having enough people to do the work. Adequate funding will allow us to continue to provide an excellent resource to the scientific community.

Thank you for your continued support of this project.

Sincerely yours,

[Signatures]

Peter Albersheim  
PI and Executive Director  
of CarbBank

Scott Doubet  
Co-PI and CarbBank  
Director

Encs.
cc: Klaus Bock  
Alan Darvill  
PA/SD:rn

**DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.