59 CODATA / NEWSLETTER

MARCH 1992

Task Group on Biological Macromolecules

The CODATA Task Group on Biological Macromolecules held its annual meeting at the CODATA Secretariat in Paris March 5-7, led by Chairman Professor A. Tsugita of the Science University of Tokyo. Twenty scientists from eight different countries, including regular, consulting, and corresponding members, all all represented different components of the projects devoted to archiving
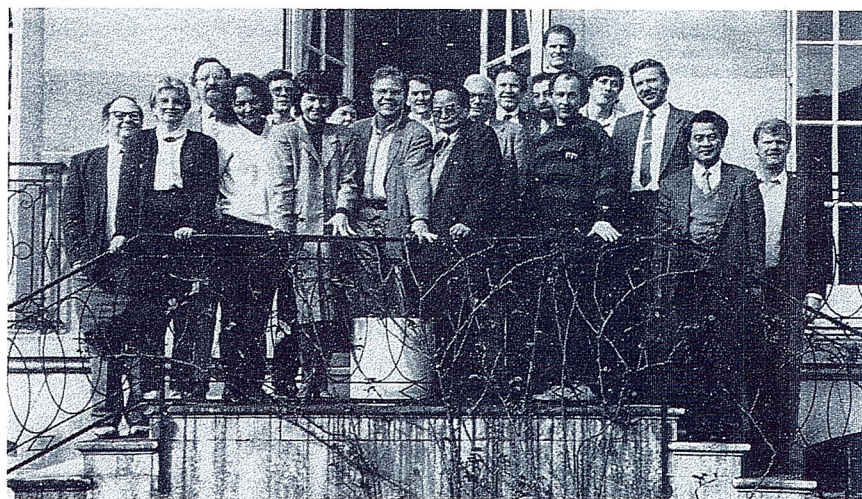
The Biological Macromolecules TG on the CODATA Secretariat Stairs. (Individuals are identified in caption on page 2.)

HIGHLIGHTS

The **Committee on Data for Science and Technology (CODATA)** was established in 1966 by the International Council of Scientific Unions.

Working on an interdisciplinary basis, CODATA seeks to improve the quality, reliability, processing, management, and accessibility of data of importance to science and technology.

data about biological macromolecules. They were joined by an observer: Dr. I. W. Tanaskovic of the United Nations University. This Task Group is an important, and in some cases unique, channel for information exchange and problem recognition. Action is taken whenever possible to resolve issues involving the different data banks. This meeting overlaps with a subsequent one on two-dimensional gels, used to characterize the protein components of cells.

In living organisms, genetic sequences in nucleic acids code for the amino acid sequences of proteins, which in turn determine their three-dimensional structures and functions. Information from each of these stages is measured, archived, distributed, and made the basis of information-retrieval systems supporting research efforts in molecular biology, medicine, and biotechnology. Reports from the database organizations represented at the meeting included descriptions of current activities at GenBank and the National Center for Biotechnology Information (NCBI) in the U.S.A., the DNA Data Bank of Japan (DDBJ) dealing with Nucleic Acid Sequences, and the tripartite International Protein Sequence Data banks (PIR-International, which comprises the groups at the National Biomedical Research Foundation (NBRF) in Washington, D.C., the Max-Planck-Institute für Biochemie in Munich (MIPS) and the Science University of Tokyo (JIPID). The establishment of this partnership was initiated by the Task Group in 1986 and its conspicuous success is one of which the Task Group is justly proud. The group also heard reports from the Protein Data Bank of three-dimensional structures of biological macromolecules, the BioMagRes Bank database of NMR data, and the Carbohydrate Data Center. Interaction between the Carbohydrate Data Center and the Chemical Abstracts Service organization was also fostered by the Task Group.

Despite the best efforts of the practitioners, the data archiving projects seem doomed to continual instability as a result of changes on the technical, political and financial fronts, of which only the first can be regarded as making steady progress. The explosive growth in the quantity and nature of data has required the data banks to revise their formats for archiving and distribution in many cases, and to investigate new media of data dissemination. For example, there was discussion of the relative merits and potential problems of the traditional "burst mode" of distribution whereby a complete set of data is distributed via magnetic media (now more often compact disk) and sent out at periodic intervals from central sources, and the "trickle mode" whereby data are made available from individual laboratories over computer networks. Is there some optimal combination? Two salient issues are the problem of quality control, the importance of which is very widely recognized, and

the particular commitment of CODATA—not so widely shared—to ensure that data will be available to all countries and to speak for scientists in those in which the latest technology is not commonly available.

Specific problems exposed at the meetings are generally turned over to specialist subcommittees for action. The meeting on 2-D gels is one such activity. Another, initiated at this meeting, addresses a nomenclature problem: that of names and abbreviations for unusual or modified amino acids. (There is a canonical set of twenty amino acids which accounts for most of the residues in natural proteins, and the nomenclature for these and many others has already been set by the International Union of Biochemistry and the International Union of Pure and Applied Chemistry, but numerous special cases remain and must be included in data banks of protein sequences, structures and NMR data, as well as in the Chemical Abstracts file itself.) A subcommittee will study the problem and, it is hoped, produce results useful to the appropriate committees of the above-mentioned Scientific Unions.

The group discussed the very serious problems concerning the legality, wisdom, and ethics (which is not to say that this order implies a priority of importance) of attempting to patent or copyright biological information. These problems have very broad significance and it is essential for the committee to consider them.

From l.- r. in photo on p.1: A.M. Lesk (University of Cambridge), W.C. Barker (National Biomedical Research Foundation), J.L. Markley (BioMagResBank NMR Database), M. Attimonelli (CNR Bari), F. Pfeiffer (Max-Planck-Institut für Biochemie, MIPS), P.S. Wilson (Chemical Abstracts Service), B.R. Seavey (BioMagResBank NMR Database), J.-M. Claverie (National Center for Biotechnology Information, NIH), R.J. Simpson (Ludwig Institute for Cancer Research), A. Tsugita (Research Institute for Biosciences), K. Loening (Topterm), T.F. Koetzle (Protein Data Bank, Brookhaven National Laborabory), R.Stuik-Prill (Carbohydrate Data Center), H.W. Mewes (Max-Planck-Institut für Biochemie, MIPS), P. Gilna (GenBank), D.G. George (National Biomedical Research Foundation), D. Schomberg (Enzyme Data Bank, Gesselschatt für Biotechnologische Forschung), J.L. Wu (Academia Sinica, Taiwan), M. Cinkowsky (GenBank).

CODATA BULLETIN 23(4) (1991)
"INFORMATION INTEGRATION FOR BIOLOGICAL MACROMOLECULES"

A synopsis of the current status of the major databases pertaining to biological macromolecules.

This concise monograph includes reports from the following data collections: The DNA **GENBANK** ● The **EMBL** Data Library ● The DNA Data Bank of Japan ● **PIR** - The Protein Identification Resource ● **MIPS** - European Node for Protein Sequence Data ● **JIPID** - Protein Information Database ● **CARBANK** - The carbohydrate database project ● **PDB** - The Brookhaven Protein Data Bank ● **BioMagResBank** - A Protein NMR database ● **ATCC** - The American Type Culture Collection database of recombinant materials ● **HDB** - The CODATA/IVIS Hybridoma Data Bank ● **NCBI** - The National Center for Biotechnology Information Database program ● **CAS/STN** - The Chemical Abstracts Service/Scientific and Technical Information Network.

Other topics addressed include Genome computing (Lawrence Berkeley Laboratory Human Genome Center) and emerging databases (*e.g.*, 2-D gel protein databases).

"...a valuable tool and information source for laboratories working in the Biosciences."

--Richard J. Simpson
Head, Protein Structure Laboratory
Ludwig Institute for Cancer Research (Melbourne)

Executive Committee Helps to Define CODATA Progress

The Executive Committee reviewed Task Groups and Commissions, monitored the expenditure of funds, and the plans for International Conferences in the near future, as well as relationships with Data Centers, ICSU bodies, and a host of scientific unions, international, and national bodies for three days during the ides of March.

CODATA's President (David Abir, Tel-Aviv, Israel) ponders a weighty decision.

The hardest tasks were possibly the scientific evaluation of its pluridisciplinary task and working groups, devising more effective publication plans, and coping with the rapidly evolving systems of data management and dissemination.

Enhancement of the efficiency of communications together with dissemination of information and developments within CODATA are two prime objectives. Provision of requested advice to the General Assembly on scientific and administrative matters was not neglected.

CODATA's Past President (David R. Lide, Jr.—Gaithersburg, Md., U.S.A.) and Treasurer (James Crease, Surrey, U.K.) consult on an administrative matter.

CODATA Executive Committee pauses for a portrait in the Conference Room.
From left to right (standing): Richard Simpson, G. H. Wood, David R. Lide, Jr., David Abir, Jacques-Emile Dubois, F. A. Kuznetsov, James Crease, Lev Gurvich. (Seated): Phyllis Glaeser (Ex. Sec.), A. Tsugita, Richard Sinding-Larsen, G. C. Carter. (Excused): Jack Westbrook, R. Eckermann.

Repertoire des Banques de Données

The Association Française des Documentalistes et Bibliothécaires Specialistes have just published the *Repertoire des Banques de Données Professionnelles 1992* which contains details on 1600 scientific and technical bibliographic and numerical data banks worldwide. The book may be purchased from ADBS, 25 rue Claude Tilleur, 75012 Paris, France. FAX: +33 1 43723041.

The CODATA Beijing International Conference

Gleanings from the Welcome letters to Beijing (*by Xu Zhihong*) and
to CODATA '92 (*by J.-E. Dubois*)

"October in Beijing --- China's greatest cultural city --- notable for golden autumn scenery --- for famous sights and ancient sites (Palace Museum, Temple of Heaven, the Great Wall....) --- for 800 years the historical capital of China.... --- "Hundreds of academic research institutes --- big industrial ministries --- educational and health institutions --- more than 100 universities and colleges..." --- center for delicious and tasteful cuisine....

"A chance --- in a pluridisciplinary meeting to approach the new data challenges as related to the technico-scientific visions of our world.

"World changes involving various paradigms and models, such as those of Global Change, Geosciences and Space sciences—based on a multitude of simple, fundamental, or complex scientific and technical data—and their dissemination as processed data.

"As we move from our post industrial age to what is considered the future Information Age, CODATA has asked our Scientific Program Committee to shape this Meeting to face fundamental transnational problems and to examine the relevant data required to feed any global system analysis. For the first time, a significant association with the World Data Centers will provide an overall view of global change problems. This tempted us to survey aspects of Asian national resources.

"Alongside these central themes, CODATA will also show its continuous progress toward interdisciplinary focus, using the powerful tools of computer science (valuable both for logistics and for logical handling of complex data knowledge synergies) --- in such areas as biology, agriculture, materials, and geosciences."

CODATA Publications

CODATA Bulletins (1991)

A Glossary of Terms Relating to Data, Data Capture, Data Manipulation, and Databases, *CODATA Bulletin*, No. 23 (1-2), January-June 1991.(a)

Database Developments in Asian-Oceanic Countries, *CODATA Bulletin*, No. 23 (3), July-September 1991. (b)

Information Integration for Biological Macromolecules, *CODATA Bulletin*, No. 23 (4), October-December 1991. (c)

CODATA Bulletins (1992)

Materials Database Management—Global Trends, Standards, and Issues, *CODATA Bulletin*, No. 24 (1), January-February 1992. (d)

Abstracts from the 12th International CODATA Conference, *CODATA Bulletin*, No. 24 (2), 1992. (e)

International Workshop—Crop Modelling.(f)

Directory of Asian and Oceanic Data Sources. (g)

(a) Available from Hemisphere Publishing Corp., New York, 1991. vii + 196 pp. ISSN 0366-757X.

(b) Available from Hemisphere Publishing Corp., New York, 1991. xxix + 138 pp. ISSN 0366-757X.

(c) Available from Hemisphere Publishing Corp., New York, 1991. iv + 107 pp. ISSN 0366-757X.

(d) Available from Begell House, New York, 1992. v + 97 pp.

(e) Forthcoming from Begell House, NY.

(f) Forthcoming from Begell House, NY.

(g) Forthcoming from Begell House, NY.

Thermodynamic Database

The Thermocenter of the Russian Academy of Sciences has issued an updated version of its IVTANTHERMO 90 Database containing internally consistent thermodynamic property data for 85 elements and about 2300 compounds (oxides, sulfides, halides, hydroxides, carbides, salts, etc.) over a wide range of temperatures in solid, gaseous, and liquid states. presented with new user-friendly software. Orders for the database (price: $1500) may be sent to Professor L. Gurvich, c/o CODATA Secretariat, 51 Bd. de Montmorency, 75016 Paris, France.

Using Environmental Data for Crop Modeling Research

The first activity of CODATA's new Commission on Global Change Data, a CODATA International Workshop on the "Integration, Dissemination, and Use of Environmental Data for Research (Applications to Crop Modeling)" was held in Chambery, France, 2-5 March 1992. Local arrangements were managed by Prof. Martin-Bouyer of the University of Savoie, located in Chambery. The technical program was organized by Mr. Paul Uhlir, of the U.S. National Research Council, and by Dr. G. Cynthia Carter, of the U.S. Department of Energy and a member of CODATA's Executive Committee. Funding for the Workshop was provided by Unesco, ICSU, and NOAA. About 30 scientists from 13 countries attended the Workshop, including representatives from WMO, FAO, UNEP, and USDA.

The first day and a half was devoted to presentations that reviewed the current status of crop modeling research, with particular emphasis on the requirements for data and data integration of the various types of crop models. Also reviewed were the needs of developing countries, particularly those in the Sahel region of Africa, and the dissemination of the results of crop modeling to those countries. Data requirements were shown to vary significantly for the various applications of crop models, which range from international research through national policy making, as well as district crop planning, to day-by-day farm management. The increased use of satellite remote sensing data received much attention.

The remainder of the Workshop was devoted to the assembly of a publication which will be the main product of the meetings. The attendees were split into four task groups, which developed written material on topics that included applications of crop modeling, data requirement, impacts of uncertainty in data and models, the applications of advanced technologies such as GIS, and the infrastructure improvements necessary to promote the dissemination of model results. A draft version of this report, together with papers from the invited speakers, will be completed by the end of April. Final publication is expected in mid 1992 as a *CODATA Bulletin*.

The participants in the Workshop felt that the discussions were extremely useful, and that the resulting publication will be an important contribution in this field. Chambery was an ideal location for a meeting of this kind, and the local arrangements were excellent.

--Michael A. Chinnery, Chairman
CODATA Commission on Global Change Data

DB-The Plant Fossil Record

Version 1 of this database (PFR 1.0) consisting of 10,478 Records representing the type species of plant fossil genera became available in April 1991 from the International Organization of Palaeobotany (IOP). The package includes a copy of the Textmaster retrieval program.

Textmaster is a freeform content searchable database program and can search for most words in any Field. Simple questions of the 10,478 Records in PFR 1.0 can be answered in a few seconds. More complex searches involving two or more words, with optional limitations such as one part of a word, may take a little longer.

Each Record is organized in the format and standard adopted in the Frankfurt Declaration at a special meeting of IOP. A maximum of 43 Fields contain taxonomic, nomenclatural, bibliographic, descriptive, stratigraphical, geographical and curatorial detail. Most of the Records are incomplete.

The Records represent *nearly* all plant fossil genera. They include details of fossils of different organs (leaves, wood, stems, pollen, etc.) of the vascular plants as well as those of remains of bryophytes, dinoflagellate cysts and other algae, fungi, nannoplankton and acritarchs. The details have been studiously edited from existing catalogs—Index Nominum Genericorum, and those of Andrews, Blazer, Watt, and Meyen.

The PFR database has the potential of answering more complex and deep questions, especially when details of species are included in later versions and when occurrence Records are added as well. For now, PFR 1.0 has very limited scope for this kind of use. Work is in progress to produce a PFR occurrence database but it will obviously take many thousands of hours of work to produce a version which matches contemporary knowledge. Such a tool would help reconstruct evolutionary processes, plot migratory changes, check other hitherto hidden patterns within the record, identify wrong identifications, etc., reconstruct palaeoenvironments and accommodate biorecords. All human knowledge of past vegetation would be open to inspection and intelligent consideration.

The package requires an IBM personal computer (or compatible) running on DOS version 3.0 or higher with at least 27 Mb of dedicated hard disk; however, selected portions can be utilized in less space.

Further information on its procurement on a stepped basis for personal, educational, and commercial cost between $100 and $1000 (US) may be had from the IOP Secretary (Prof. M. C. Boulter, Polytechnic of East London, Romford Road, E15 4L7, U.K.).
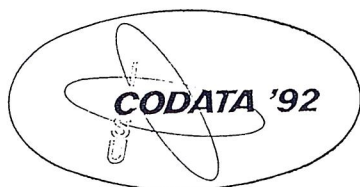
Reminder: Register for
"New Data Challenges...."
13th International CODATA Conference!