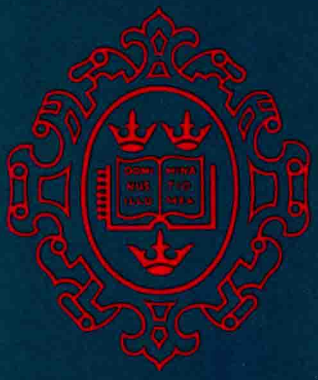
Biomolecular Data

A Resource in Transition

Edited by

RITA R. COLWELL



OXFORD SCIENCE PUBLICATIONS

Biomolecular Data

A Resource in Transition

Editor

RITA R. COLWELL

Maryland Biotechnology Institute
University of Maryland

Associate Editors

DAVID G. SWARTZ

and

MICHAEL T. MACDONELL

OXFORD NEW YORK TOKYO

OXFORD UNIVERSITY PRESS

1989

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press, New York

British Library Cataloguing in Publication Data
Biomolecular data
1. Biochemistry. Machine-readable files
I. Colwell, Rita R. (Rita Rossi), 1934–
574.19'2'0285574
ISBN 0–19–854247–X

Library of Congress Cataloging in Publication Data
Biomolecular data: a resource in transition/editor, Rita R. Colwell; associate editors, David G. Swartz and Michael T. MacDonell. Summaries of the results of the First CODATA Workshop on Nucleic Acid and Protein Sequencing Data, held at the National Bureau of Standards, Gaithersburg, Md., on May 3–6, 1987, sponsored by CODATA and other organizations.
Bibliography Includes index.
1. Nucleotide sequence—Congresses. 2. Information storage and retrieval systems—Nucleotide sequence—Congresses. 3. Amino acid sequence—Congresses. I. Colwell, Rita R., 1934– II. Swartz, David G. III. MacDonell, Michael Terrell. IV. CODATA. V. CODATA Workshop on Nucleic Acid and Protein Sequencing Data (1st: 1987: National Bureau of Standards)
QP625.N89B56 1989 (574.87'328—dc19) (88–38087)
ISBN 0–19–854247–X

Printed in Great Britain
at the University Printing House, Oxford
by David Stanford
Printer to the University

£30-00

BIOMOLECULAR DATA

PREFACE

In 1966 the International Council of Scientific Unions established CODATA, the Committee on Data for Science and Technology, with the goal of improving the quality, reliability, management, and accessibility of important data in all fields of science and technology. Working with its National and Union Members, CODATA has attempted to provide an umbrella for cooperative efforts of an international and interdisciplinary nature. These efforts include the dissemination of recommended data sets in key areas of science, development of standard formats, preparation of directories to sources of data, pilot projects to develop databases, and a variety of educational endeavors.

In recent years CODATA has given particular emphasis to issues arising from the rapidly growing store of quantitative data in the biosciences. A pilot data bank on hybridomas and monoclonal antibodies was established in 1982. A task group to coordinate the collection of protein sequence data has been active since 1984. CODATA is also a sponsor of the Microbial Strain Data Network, which has been established in conjunction with the World Federation of Culture Collections and other organizations.

As a further step in this direction CODATA was pleased to join a number of other organizations in sponsoring the First CODATA Workshop on Nucleic Acid and Protein Sequencing Data, held at the National Bureau of Standards, Gaithersburg, Maryland, on May 3-6, 1987. This workshop brought together leading scientists involved in the determination, storage, and use of protein and nucleic acid sequences at a strategic time when plans for large-scale projects on the human genome were under active discussion. *Biomolecular Data: A Resource in Transition* summarizes the results of this workshop and sets out significant issues and directions that must be considered in future database development. As planning progresses for such development, the ideas discussed in this book should prove to be of great value to the scientific and technical community.

David R. Lide
President, CODATA
August 1988

Editors may be contacted at these addresses:

Dr. Rita R. Colwell
Maryland Biotechnology Institute
University of Maryland
College Park, MD 20742, USA

David Swartz
Data Center Manager
Maryland Sea Grant College
University of Maryland
College Park, MD 20742, USA

Dr. Michael MacDonell
AMBIS Systems
3939 Ruffin Road
San Diego, CA 92123, USA

ACKNOWLEDGEMENTS

We are grateful to the many authors and workshop participants who made *Biomolecular Data: A Resource in Transition* possible. We are also grateful to Mark Jacoby and Merrill Leffler of the Maryland Sea Grant College for assistance with editing. Special thanks to Sandy Harpe, also of the Maryland Sea Grant College, for book design and layout and to Lisa Griffin and Barbara Burnett for word processing. We would like to recognize Dr. David Lide for his support from the beginning of this project through to its completion and also to Kathy Stang of his staff.

We wish to acknowledge the following organizations for their support: the Maryland Sea Grant College, the National Bureau of Standards, CODATA, the Maryland Biotechnology Institute, the National Library of Medicine, the Food and Drug Administration, and Monsanto, Inc.

—*The Editors*

CONTENTS

INTRODUCTION

In a few short years biotechnology, once viewed as the science of the future, has rapidly become the technology of the present. What began largely as a descriptive science has led to the development of specific genetic engineering techniques that will affect the way we live and the way we think. Scientists are now able to capitalize on the efforts of researchers who have worked for several decades on projects that are revealing the structure and function of the genetic material of organisms. Helping to stimulate this rapid progress are the numerous molecular biology databases organized to house and assemble data so that they could be put to the best use for furthering scientific knowledge.

We are in a phase of rapid growth in the number and variety of information resources and in the quantity of data available to support fundamental and applied research. Numerous information sources are available including databases on nucleic acid sequences, protein sequences, carbohydrate sequences, molecular structure, human genetic information, computer networks, bulletin boards and file servers, bibliographic information, restriction enzymes, vectors, microbial strains and cell lines, hybridomas, genetic maps and probes, and even databases describing other databases. (See the appendix for a brief description of many of these information resources.)

Of special importance to the progress of biological research and biotechnology are the nucleic acid and protein sequence databases. These databases have proved to be a valuable resource for the planning and evaluation of the results of sequencing experiments. They have also provided the basis for statistical analysis and the comparison of large numbers of sequences.

The availability of sequence databases has helped further biological knowledge in a number of areas. In oncogene research, for example, the complete sequences of genes found in malignant human cells have been determined, and researchers have been able to compare these sequences with similar ones of genes present in certain RNA tumor viruses that are implicated in the cause of cancer in mice and chickens. More recently Dr. Russell Doolittle has discovered that one

of these oncogenes, v-sis, found in a simian sarcoma virus, is very similar to a human gene giving rise to the growth factor PDGF, or platelet-derived growth factor. In the body, PDGF stimulates epithelial cells to grow, thereby playing an important role in wound healing. PDGF also plays a role in the proliferation of cells that clog blood vessels, creating the conditions that may lead to a heart attack. Doolittle's computerized comparison established a surprising link between heart disease and cancer. This application has demonstrated that information on the function of a gene could be derived from comparison of its sequence to others with known functions in a database.

The availability of sequence data has also furthered our understanding of evolutionary relationships among many forms of life. Using the aligned sequences of highly conserved molecules such as rRNA or certain enzymes, biologists have been able to infer the phylogenetic relatedness of different organisms. These approaches are helping to develop more stable systems of classification, particularly with regard to prokaryotic organisms.

There are other areas of biotechnology and the life sciences that are growing as sequence information is accumulated. One area involves protein structure determinations. The elucidation of the structure and/or function of a protein often begins with the direct determination of its amino acid sequence, or by inferring the sequence from the corresponding nucleotide sequence (cDNA) of the gene that encodes the protein. This capability has assisted the growth of the kind of protein engineering where the structure of proteins is determined through X-ray crystallographic studies. These studies generate and require large quantities of data—amino acid sequence, crystallization conditions, and coordinates of the 3-D representation of the protein, among them—that are analyzed by elaborate computer programs. Many of these data are being stored in large databases such as the Protein Identification Resource or the Protein Data Bank of the Brookhaven National Laboratory.

The growth in sequence data has arisen as a result of technological breakthroughs occurring in the 1970s. Advances in sequencing made it clear that sequencing had become an important tool of the life sciences. After these developments came a rapid increase in the number of sequences. Nucleic acid sequences are now being reported for a range of known functions such as protein coding, RNA-coding, regula-

tory regions of both DNA and mRNA, and also for structural regions of RNA.

Rapid growth of the volume of data has led to numerous problems that demand the scientific community's attention. For example, there are long lags in the submission and entry of data; there are problems ensuring sufficient peer review and quality control; and while more documentation on data is often needed, there persists an even greater need for better access to data and for more innovative analysis.

In recognition of such problems arising from the rate at which nucleic acid and protein sequence data are being accumulated by the international scientific community, as well as growing problems encountered in the management, quality control, distribution and publication of these data, an international workshop was held May 3-6, 1987. Emphasis was placed on methods of standardization, networking, global interfacing of databases, publication of sequences, as well as on exploration of the potential use of databases for storage and retrieval of higher order information such as periodicity, pattern and signal, secondary, and tertiary structures, and genome maps.

An international committee composed of representatives of GenBank, the University of Maryland Biotechnology Institute, Berlin RNA Databank, Nagoya University, Protein Identification Resource (PIR), Pasteur Institute, CODATA, National Cancer Institute, National Institutes of Health, and the National Bureau of Standards was formed to define the general nature of the problems associated with international sequence databases. This group met in February, 1986 and again in September, 1986 to outline a workshop in which these problems could be addressed by experts drawn from the international scientific community. From these meetings emerged the consensus that some of the most critical problems facing sequence databases are (1) standardization, (2) global interfacing, (3) cooperation among journals and databases, (4) criteria for publication of sequences and whether database entry constitutes valid publication, (5) education, and (6) users' requirements and applications. To address these points, a workshop format consisting of interactive panels of specialists along with a number of invited speakers was organized to address specific problem areas. In addition, demonstrations and posters were solicited to encourage an exchange of ideas and innovations throughout the field of sequence data management and analysis. This workshop was sup-

ported in part by the National Library of Medicine, the Food and Drug Administration, the Maryland Biotechnology Institute, the National Bureau of Standards, CODATA and Monsanto, Inc.

This book has evolved out of the initial effort of the workshop and from the diligent efforts of its many contributors. The purpose is to help bring attention to the recent advances in the molecular biology databases, and also to the needs and requirements of the individuals who maintain and manage them, and the users who depend on them. Special emphasis has been placed on the sequence databases; however, due to the interrelatedness of sequence data with other forms of molecular biological data we have also included discussion of other important data as well.

The book is divided into four sections:

Section 1 provides a history and some important background information on the emergence of biomolecular data and the problems which have arisen as a result of rapid growth of this data.

Section 2 discusses the management of biomolecular data, including a discussion of current and anticipated problems, and proposals for possible solutions. Managers of the many different information resources available in molecular biology and related areas contributed to this section.

Section 3 focuses on the users of molecular biological data, providing information on current applications and insight into the requirements of users. The authors also provide suggestions to the providers of data and to funding agencies concerning the availability and need for certain services, such as training and support, access to and support for certain software, and documentation and integration of information resources.

Section 4 provides a window into the future of biomolecular information resources. Papers prepared by noted specialists discuss important trends and changes occurring in the management, distribution and generation of biomolecular data and related information. An important topic is the integration of different information resources

and the availability of these resources via on-line access through computer networks or through portable media such as compact disks. A recurring topic is the changing role of peer-reviewed journals and their relationship with the sequence databases.