John A. A. Sillince · Maria Sillince

Molecular Databases for Protein Sequence and Structure Studies

Springer-Verlag

John A. A. Sillince · Maria Sillince

Molecular Databases for Protein Sequences and Structure Studies

An Introduction

With 27 Figures

Springer-Verlag
Berlin Heidelberg NewYork
London Paris Tokyo
Hong Kong Barcelona Budapest

Dr. John A. A. Sillince
Lecturer in Management Information Systems
British Sheffield University
Management School
Crookesmoor Building
Conduit Road
Sheffield S10
1FL, UK

Dr. Maria Sillince
Assistant Subject Librarian
Wolverhampton Polytechnic
Wolverhampton
UK

ISBN 3-540-54332-5 Springer-Verlag Berlin Heidelberg NewYork
ISBN 0-387-54332-5 Springer-Verlag NewYork Berlin Heidelberg

Library of Congress Cataloging-in-Publication Data
Sillince, John
Molecular Databases for protein sequences and structure studies :
an introduction / J. A. A. Sillince, M. Sillince
Includes bibliographical references and index.
ISBN 0-387-54332-5
1. Amino acid sequence--data processing.
2. Proteins--Analysis--Data processing.
3. Expert systems (Computer science)
I. Sillince, M. (Maria)
II. Title.
QP551.S56 1991
574.19'245'0285--dc20 91-25898

Typesetting: Camera ready by authors
Printing: Color-Druck Dorfi GmbH, Berlin; Binding: Lüderitz & Bauer, Berlin
51/3020-5 4 3 2 1 0 Printed on acid-free paper

Acknowledgements

We are very grateful to all those scientists who replied to the questionnaire discussed in the last chapter of this book. Although lack of space prevents listing them all, the replies of the following were particularly helpful: Professor T.R. Blundell, Dr. A. Sali, Dr. J.P.Overington, Dr. M.S. Johnson, Dr. J.M. Thornton and Dr. S.P. Gardner at the Department of Crystallography at Birkbeck College, London University; Dr. D.A. Clark, Dr. G.J. Barton and Dr. C.J. Rawlings at the Biomedical Computing Unit, Imperial Cancer research Fund, London; Dr. E.M. Mitchell , Dr.P.J. Artymiuk, Dr.D.W. Rice, and Dr. P. Willet at the Department of Information Studies, University of Sheffield; Dr.S.Karlin at the University of California for making available copies of papers prior to publication; Dr. P.E. Jansson, Dr. L. Kenne, and Dr. G. Widmalm of the Department of Organic Chemistry, University of Stockholm.

Similarly for the latest information on biomolecular databanks we would like to thank Dr. R.F. Doolittle, Dr. W.C. Barker and Dr. D.F. Feng at the Centre for Molecular Genetics, University of California, San Diego; Dana Smith at the Complex Carbohydrate Research Center, University of Georgia; Dr. J.R.Lawton, Dr. F.A. Martinez and Dr. C. Burks at Los Alamos; also information scientists at Fisons Pharmaceuticals, Loughborough, and at Courtaulds, Coventry; also we are grateful to Professor J. Meadows and Dr. J.F.B. Rowland at the Department of Library and Information Studies at Loughborough University for their assistance

in planning and criticising parts of the manuscript; also we would like to thank librarians at Loughborough and Warwick University Libraries and at Coventry and Wolverhampton Polytechnic Libraries; and at the British Library Document Supply Centre, Weatherby.

Also we would like to acknowledge the kind permission of the following for allowing us to reproduce material: Figure 1.1 is reprinted with permission by Watson J.D., "Molecular biology of the gene", Copyright (c) 1976, Benjamin/ Cummings Publishing Company; Figure 1.2a is reprinted with permission from McGammon J.A., and Harvey S.C.,Copyright (c) 1987, "Dynamics of proteins and nucleic acids", Cambridge University Press; Figure 1.2b is reprinted with permission from McGammon J.A., and Harvey S.C., Copyright (c) 1987, "Dynamics of proteins and nucleic acids", Cambridge University Press; Figure 1.3 is reprinted with permission from McGammon J.A., and Harvey S.C.,Copyright (c) 1987, "Dynamics of proteins and nucleic acids", Cambridge University Press; Figure 1.4 is reprinted with permission from F.H.C.Crick, "The genetic code", in "Recombinant DNA", Copyright (c) 1978, W.H.Freeman, Figure 1.5. is reprinted with permission from Bishop M.J., Ginsburg M., Rawlings C.J., and Wakeford R., "Molecular sequence databases", in Bishop M.J., and Rawlings C.J., (Eds) "Nucleic acid and protein analysis: a practical approach", Copyright (c) 1987, IRL Press, by permission of Oxford University Press; Figure 1.6 is reprinted with permission from Brown D.D., "The isolation of genes", in "Recombinant DNA", Copyright (c) 1978, W.H.Freeman, Figure 1.7 is reprinted with permission from Bishop M.J., Ginsburg M., Rawlings C.J., and Wakeford R., "Molecular sequence databases", in Bishop M.J., and Rawlings C.J., (Eds) "Nucleic acid and protein analysis: a practical approach", Copyright (c) 1987, IRL Press,by permission of Oxford University Press; Figure 2.1 is reprinted with permission from Nucleic Acids Research, Vol. 11, Hawley D.K., and

McClure W.R., Copyright (c) 1983, IRL Press, by permission of Oxford University Press; Figure 4.4 is reprinted with permission from Ash J.E., Chubb P.A., Ward S.E., Welford S.M., and Willett P., "Communications storage and retrieval of chemical information", Copyright (c) 1985, Ellis Horwood; Figure 4.5 is reprinted with permission from Ash J.E., Chubb P.A., Ward S.E., Welford S.M., and Willett P., "Communications storage and retrieval of chemical information", Copyright (c) 1985, Ellis Horwood; Figure 4.6 is reprinted with permission from Lynch M.F., Harrison J.M., Town W.G., and Ash J.E., "Computer handling of chemical structure information", Copyright (c) 1971, Macdonald; Figure 4.7 is reprinted with permission from Computers and Chemistry, Vol.13, Haines R.C., "Computer graphics for processing chemical substance information", Copyright (c) 1989, Pergamon Press PLC; Figure 5.1 is reprinted with permission from Database, Vol. 10, Barnard J.M., "Online graphical searching of Markush structures in patents", Copyright (c) 1987, Online Inc; Figure 5.2 is reprinted with permission from Lynch M.F., Harrison J.M., Town W.G., and Ash J.E., "Computer handling of chemical structure information", Copyright (c) 1971, Macdonald; Figure 7.1 is reprinted with permission from Keil B., "Cooperation between databases and the scientific community", in Doolittle R.F., (Ed) "Molecular evolution: computer analysis of protein and nucleic acid sequences", Copyright (c) 1990, Methods in Enzymology, Vol. 183, Academic Press; Figure 7.3 is reprinted with permission from Bishop M.J., Ginsburg M., Rawlings C.J., and Wakeford R., "Molecular sequence databases", in Bishop M.J. and Rawlings C.J. (Eds) "Nucleic acid and protein analysis: a practical approach", Copyright (c) 1987, IRL Press, by permission of Oxford University Press; Figure 8.1 is reprinted with permission from Database, Vol. 121, Allen F.C., and Ferrell W.R., "Numeric databases in science and technology: an overview", Copyright (c) 1989, Online Inc; Figure 8.2

is reprinted by permission from Nature Vol. 335, pp. 745-748, Copyright (c) 1988, MacMillan Magazines Ltd; Figure 9.1 is reprinted with permission from Doolittle R.F., "Of ORFs and URFs: a primer on how to analyse derived amino acid sequences", Copyright (c) 1986, University Science Books; Figure 9.2 is reprinted by permission from Nature Vol. 296, pp. 171-173, Copyright (c) 1982, MacMillan Magazines Ltd; Figure 9.3 is reprinted with permission from Doolittle R.F., "Of ORFs and URFs: a primer on how to analyse derived amino acid sequences", Copyright (c) 1986, University Science Books; Figure 10.2 is reprinted with permission from Thornton J.M. and Taylor W.R., "Structure prediction", in Findlay J.B.C. and Geisow M.J., (Eds), "Protein sequencing: a practical approach", Copyright (c) 1989, IRL Press, by permission of Oxford University Press.

Preface

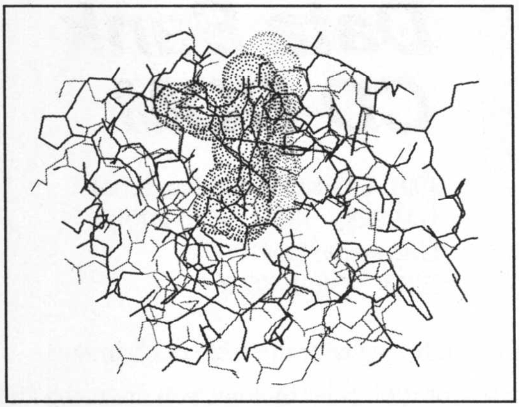
Molecular information is now too vast to be acquired without the use of computer-based systems, which can select according to supplied criteria. Developments in programming have created the ability to extend molecular science in ways that would have been impossible without their help. New databases are being established which enable previously unanswerable questions to be considered. One of these questions is whether or not one can predict the three dimensional structure of a protein from information about its sequence of amino acids.

In order to help the reader to understand this new field, several topics are explained in this volume. The structure and function of proteins and nucleic acids are described, in order to emphasise the way in which three dimensional structure reflects a protein's role in the organism. Also it is important to consider what is involved in molecular data, and how it is represented and registered in software and on the screen. Another aspect to consider is how computer-based research tools are used in molecular science, in particular for manipulating sequence and structure information. Sequence and structure are at the centre of research problems in molecular science, in the identification of a new protein (14000 are known so far) or its three dimensional structure (only 400 are known so far), in patent writing and patent searching, and in modelling proteins.

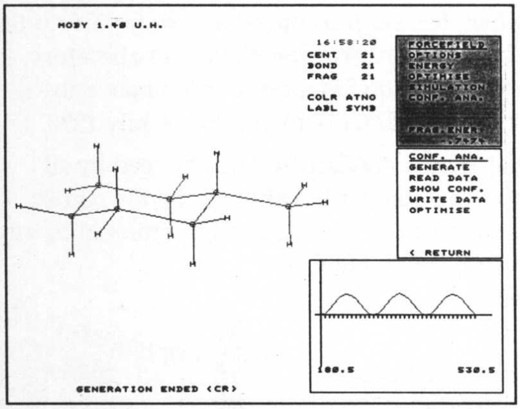
There is also a description of the state of the art in what databanks exist, both for sequences and for structures, and what types of system are available for using them, for modelling, for searching, and for integrating the operations of the online database and the local system such as a PC in a laboratory. New developments in knowledge-based systems and database technology are described. The case study on protein structure prediction, which includes developments in the specification of an expert system for such a problem, is intended to exemplify the integrated nature of modelling and search (both computer-based) and laboratory experiment in molecular science.

A New Molecular Modeling Program for the PC

U. Höweler ___ MOBY ___ Version 1.4



Structure analysis of a hemoglobin subunit of Cytochrome C Van der Waals representation of the hemoglobin structure embedded in the peptide environment. The iron atoms is coordinated by the four nitrogen atoms of the ring system, by a nitrogen of a histidine residue and the sulfur atom of a methionine side chain.



```
MOBY 1.40 U.H.

16:58:20
CENT    21
BOND    21
FRAG    21

COLR ATNO
LABL SYMB

GENERATION ENDED <CR>
```

Confirmational analysis of methyl cyclohexane which shows MOBY menu at right, additional information about the molecule at left and rotational profile at bottom right.

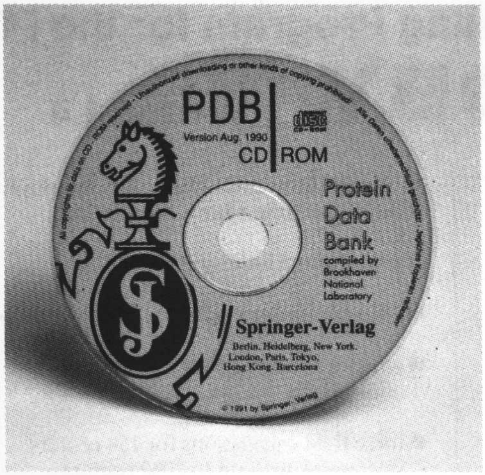
Hardware requirements:
IBM PC or compatible computer, 640 kB RAM, 80x87 arithmetic coprocessor, MS-DOS version 2.x or higher, hard disk, 1.5 MB free disk space, graphics card EGA or VGA, HERCULES supported, mouse optional.

Moby is a Molecular Modeling Program for IBM PC and compatible computers.

It provides the following functions:

- 3D graphic display for up to 2000 centers
- structure and property analysis and comparison
- force field calculations for 150 centers interacting with up to 2000 centers
- geometry optimization and conformation analysis and molecular dynamics simulation
- semiempirical quantum chemical calculation (MNDO, AM1)
- MOBY reads and writes standard structure file formats (e. g. Protein Structure Database format)
- MOBY reads and writes 3D geometries in any format
- MOBY writes HPGL plot files and generates hardcopy output

Springer-Verlag
Berlin
Heidelberg
New York
London
Paris
Tokyo
Hong Kong
Barcelona
Budapest


PDB
Version Aug. 1990
CD ROM
Protein Data Bank
compiled by Brookhaven National Laboratory
Springer-Verlag
Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona
© 1991 by Springer-Verlag

Brookhaven National Laboratory,
Upton, Long Island, NY (Ed.)

Protein Data Bank CD-ROM

1991. CD-ROM and handbook
DM 998,- ISBN 3-540-14101-4
University: DM 498,-
ISBN 3-540-14102-2

No more queuing in the network, no more quibbling with the I&D department!

You are working in the field of biochemistry or molecular biology, you are using an industry-standard PC, you have powerful software at your disposal for the visualization of macromolecular structures, and you have been waiting for a comprehensive database of proteins, enzymes and polynucleic acids, accessible from your PC at your bench and at any time? Here it is: The **Protein Data Bank CD-ROM** comprises all the files of the original Protein Data Bank which has been compiled and so far distributed on tapes (DATAPRTP and DATAPRFI) by the Brookhaven National Laboratory, Upton, NY (USA). This particular issue contains the 554 atomic coordinate entries, the source codes and bibliographic records of the Data Bank release of July 1990.

The CD-ROM has been produced in High Sierra standard and can be read by all brands of CD-ROM drives complying with this standard. Although the file names follow the MS-DOS convention, the file structure can be read and interpreted by any operating system (MS-DOS, UNIX, MacIntosh-OS, etc.), independently of hosts, data lines and with no costs for online searchers.

Do you go in for molecular modelling? if so, we have a software package which will prove particularly useful to you: MOBY. MOBY can display structures with up to 2000 centres (or atoms), and its tools for probing into and modifying these structures help you to perform the decisive steps of your analyses right at your desk and may even take you as far as the question on hand demands.

Springer-Verlag
Berlin
Heidelberg
New York
London
Paris
Tokyo
Hong Kong
Barcelona
Budapest

Contents

CHAPTER SEVEN: THE MAIN SEQUENCE DATABANKS IN MOLECULAR SCIENCE

List of Figures