

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 New York
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

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways, and storage in data banks. Duplication of this publication or parts thereof is only permitted under the provision of the German Copyright Law of September 9, 1965, in its current version and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin, Heidelberg 1991

Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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 Courtauld, 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

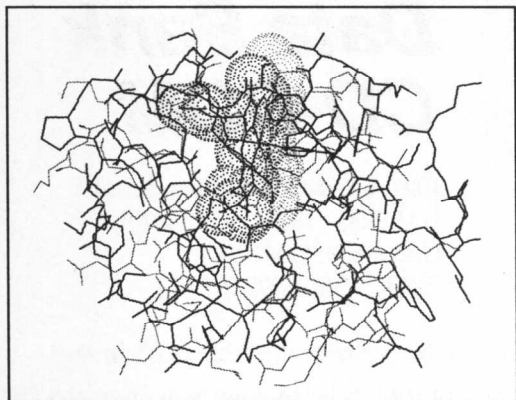
Molecular information is now too vast to be acquired without the use of computer-based systems, which can select according to supplied criteria. Developments in programming have created the ability to extend molecular science in ways that would have been impossible without their help. New databases are 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.

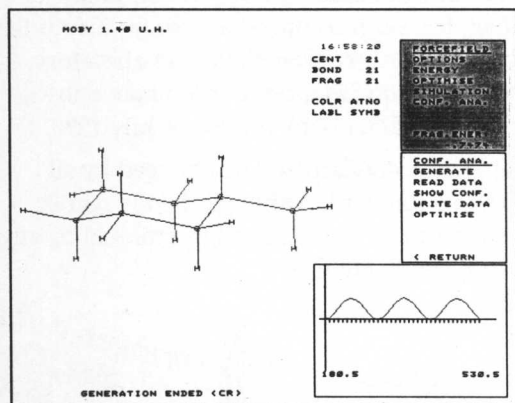
There is also a description of the state of the art in what data-banks exist, both for sequences and for structures, and what types of system are available for using them, for modelling, for searching, and for integrating the operations of the online database and the local system such as a PC in a laboratory. 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 atom 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.



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:

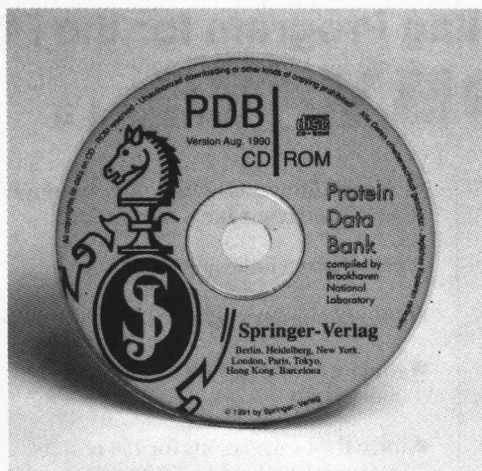
IBM PC or compatible computer, 640 kB RAM, 80x87 arithmetic coprocessor, MS-DOS version 2.x or higher, hard disk, 1.5 MB free disk space, graphics card EGA or VGA, HERCULES supported, mouse optional.

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



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

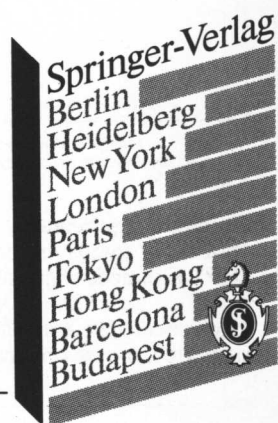
Protein Data Bank CD-ROM

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.



Contents

CHAPTER ONE: INTRODUCTION	1
1.1 Aims of book	1
1.2 The structure and role of proteins and nucleic acids	2
1.2.1 protein structure	2
1.2.2 primary structure	4
1.2.3 secondary structure	4
1.2.4 super—secondary structure	6
1.2.5 tertiary structure	7
1.2.6 quaternary structure	7
1.2.7 DNA structure	7
1.3 The nature of molecular data and its representation	10
1.3.1 sequence data and its representation	11
1.3.2 structure data and its representation	13
1.4 The importance of protein structure and function studies	14
1.5 References	17
CHAPTER TWO: COMPUTER—BASED RESEARCH TOOLS FOR MOLECULAR SCIENCE	20
2.1 The use of computers and online facilities in sequencing	20
2.1.1 computer—based sequencing projects	20
2.1.2 computer—based sequence analysis	21
2.2 The importance of sequence databanks in sequence analysis	26
2.2.1 proposed second—generation databanks	27
2.3 Integration of databank searching with sequence determination	27
2.4 References	28
CHAPTER THREE: ONLINE DATABASES IN BIOCHEMISTRY AND MOLECULAR SCIENCE	31
3.1 The importance of online databases	31
3.2 Why use online services?	33
3.3 What problems motivate using online services?	34
3.4 Types of online databases and CD—ROMs	37
3.5 The financing of databases	41
3.6 Training end—users	43
3.7 Current awareness and in—house systems	44
3.8 References	46

CHAPTER FOUR: METHODS OF COMPUTER REPRESENTATION AND REGISTRATION		
4.1 Ambiguous versus unambiguous representation		51
4.1.1 ambiguous representation		51
4.1.2 unambiguous representation		52
4.2 Graphical data representation		53
4.2.1 internal representation of graphical information		59
4.2.2 screen representation of graphics		59
4.3 Interconversion of structure representation		65
4.4 Registration		67
4.4.1 registration with canonical description		68
4.4.2 isomer sort registration		70
4.4.3 translation between systematic nomenclatures		70
4.4.4 automatic indexing		71
4.5 References		71
		73
 CHAPTER FIVE: DATABASE SEARCHING IN BIOCHEMISTRY AND MOLECULAR SCIENCE		
5.1 Bibliographic searching		77
5.2 Patent searching		77
5.3 Substructure searching		81
5.3.1 screens		84
5.3.2 structure elucidation		86
5.3.3 pattern matching searches		87
5.3.4 chemical structure searching and modelling software		88
5.4 References		90
		92
 CHAPTER SIX: USING EXPERT SYSTEMS FOR DATABASE SEARCHING IN MOLECULAR SCIENCE		
6.1 Introduction		96
6.1.1 the ANSI/SPARC database standard		96
6.1.2 expert systems		97
6.2 Elements of database systems (DBS)		98
6.2.1 the relational approach		100
6.2.2 object-oriented databases		100
6.2.3 semantic databases		104
6.2.4 object-oriented and semantic databases compared		106
6.3 Elements of knowledge base systems (KBS)		107
6.4 Elements of a knowledge-based management system (KBMS)		108
6.4.1 dynamically—updated knowledge bases		109
6.4.2 natural language interfaces		110
6.4.3 user modelling		110
6.4.4 self—modifying consistency checks		113
6.4.5 knowledge representation		114
6.4.6 knowledge manipulation and retrieval		114
6.4.7 reasoning		115
6.4.8 using reasoning to update the knowledge base		116
6.4.9 explanation		117
		118

6.4.10 security	118
6.4.11 integrity	118
6.4.12 protection	119
6.4.13 pictorial databases	119
6.4.14 moving images	120
6.5 References	121

CHAPTER SEVEN: THE MAIN SEQUENCE DATABANKS IN MOLECULAR SCIENCE

7.1 Definitions	125
7.2 Short history of sequence databanks	126
7.3 What databases are available?	128
7.4 The main sequence databases	134
7.4.1 the GenBank databases	134
7.4.2 the EMBL Data Library	135
7.4.3 GENINFO	136
7.4.4 the NBRF—PIR protein sequence databases	137
7.4.5 the NEWAT database	138
7.4.6 the PRF—SEQDB databank	139
7.4.7 the DNA Data Bank of Japan (DDBJ)	140
7.4.8 GENESEQ	140
7.4.9 Institut Pasteur databanks	141
7.5 Data structure and management	141
7.5.1 management of databases	142
7.5.2 cooperation between databanks	144
7.5.3 data acquisition	146
7.5 Data retrieval and manipulation	146
7.6.1 search strategy	146
7.6.2 database management systems	147
7.7 An example of protein sequence analysis software	150
7.8 References	152

CHAPTER EIGHT: THE MAIN STRUCTURE DATABANKS IN MOLECULAR SCIENCE

8.1 The significance and history of structure databanks	157
8.2 The Protein Data Bank (PDB) at Brookhaven	158
8.3 The Cambridge Crystallographic Databank	160
8.4 The Complex Carbohydrate Structure Database (CCSD)	161
8.5 Integrated access to structure and sequence data	165
8.5.1 the Canadian Scientific Numeric Database Service (CAN/SND)	167
8.5.2 the ISIS Integrated Sequence and Structure Database	167
8.6 References	170

CHAPTER NINE: SEQUENCE SEARCHING

9.1 Introduction	173
9.2 First example of a sequence search	179
9.3 Second example of a sequence search	181
9.4 Protein structure prediction	182
9.5 References	186

CHAPTER TEN: CASE STUDY: SPECIFICATION OF AN EXPERT SYSTEM FOR PROTEIN STRUCTURE PREDICTION	
10.1 Introduction	189
10.2 Problem description	189
10.3 The main functions of a structure prediction expert system	191
10.3.1 expert system components	193
10.3.2 avoidance of a flat expert system	193
10.3.3 well—definedness	195
10.4 Elicitation of knowledge from protein scientists	196
10.4.1 the need for integration	199
10.4.2 graphical interface	199
10.4.3 making structure prediction programs intelligent	201
10.5 References	203
	211
CHAPTER 11: APPENDIX	
11.1 Source of information	216
11.2 Addresses	216
11.3 References	216
INDEX	233
	234

List of Figures

Figure 1.1	The 20 common amino acids found in proteins	3
Figure 1.2a	The backbone or sidechain of a polypeptide	5
Figure 1.2b	A tyrosine sidechain of a polypeptide	5
Figure 1.3	The alpha—helix (a) and the beta—sheet (b)	6
Figure 1.4	The DNA molecule	8
Figure 1.5	Codes for nucleotides, including uncertainties	9
Figure 1.6	Synthesis of protein molecules	10
Figure 1.7	The one—letter and three—letter codes for amino acids	11
Figure 1.8	The genetic code	13
Figure 2.1	Matrix evaluation of a sequence	23
Figure 2.2(a)	Matrix for —10 promoter region for consensus sequence	24
Figure 2.2(b)	Matrix for —10 promoter region with different penalties for different mismatches to consensus	24
Figure 2.2(c)	Matrix with elements proportional to the frequency of each base at each position in a collection of promoters	24
Figure 2.3	Mutation substitution matrix	25
Figure 4.1	Two distinct compounds with identical GREMAS fragmentation codes	53
Figure 4.2	Generalisation of a specific molecule	54
Figure 4.3	Connection table for a simple acyclic structure	55
Figure 4.4	CAS Registry (III) connection table	56
Figure 4.5	WLN ring notations	57
Figure 4.6	Three different linear notations	59
Figure 4.7	Graphical Data Structure	61
Figure 5.1	A Markush Structure Description	81
Figure 5.2	Example of substructure query	84
Figure 7.1	Comparison of data volumes in nucleotide sequences	126
Figure 7.2	Online databases for biological macromolecules and their components	133
Figure 7.3	An example of CODATA recommended format in the NBRF—PIR database	143
Figure 8.1	Summary of information contained in CCDC database	161
Figure 8.2	The ISIS integrated protein sequence and structure gateway	168
Figure 9.1	Alignment of alpha and beta chains in human haemoglobin	174

Figure 9.2	Search of short sequences carried out by Brown et al	179
Figure 9.3	Different rates of change of different proteins	181
Figure 10.1	The current approach to structure prediction	199
Figure 10.2	Composition graph of secondary structure	206