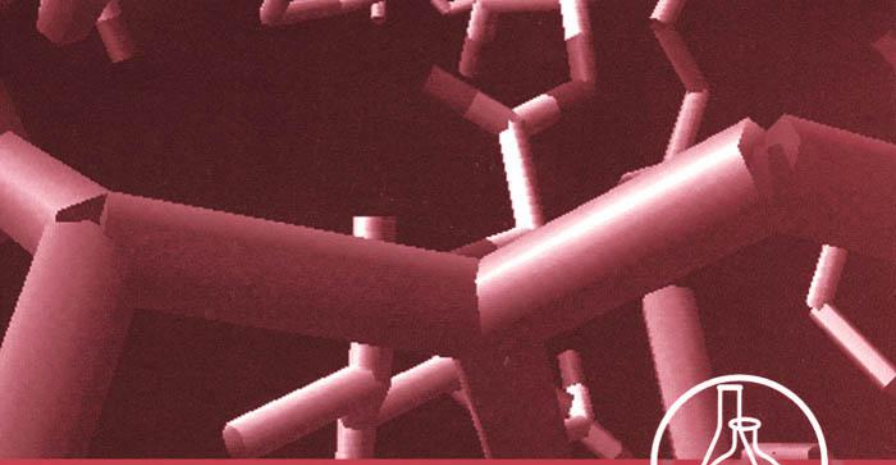

Bioinformatics

Sequence, structure and databanks

Edited by
Des Higgins
Willie Taylor

PRACTICAL APPROACH

Bioinformatics: Sequence, structure, and databanks

'Atomics is a very intricate theorem and can be worked out with algebra but you would want to take it by degrees because you might spend the whole night proving a bit of it with rulers and cosines and similar other instruments and then at the wind-up not believe what you had proved at all.

'Now take a sheep', the Sergeant said. 'What is a sheep only millions of little bits of sheepness whirling around and doing intricate convolutions inside the sheep? What else is it but that?'

(from *The Third Policeman*, Flann O'Brien)

The Practical Approach Series

Related **Practical Approach** Series Titles

Protein-Ligand Interactions: structure and spectroscopy*
Protein-Ligand Interactions: hydrodynamic and calorimetry*
DNA Protein Interactions
RNA Protein Interactions
Protein Structure 2/e
DNA and Protein Sequence Analysis
Protein Structure Prediction
Antibody Engineering
Protein Engineering

* indicates a forthcoming title

Please see the **Practical Approach** series website at
http://www.oup.co.uk/pas
for full contents lists of all Practical Approach titles.

Bioinformatics: Sequence, structure, and databanks

A Practical Approach

Edited by

D. Higgins

Department of Biochemistry,
University College, Cork,
Ireland

and

W. Taylor

Division of Mathematical Biology,
National Institute for Medical Research,
The Ridgeway, Mill Hill,
London NW7 1AA, UK

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Cape Town
Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul Karachi
Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw
with associated companies in Berlin Ibadan
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2000
Reprinted 2001

A catalogue record for this book is available from the British Library
Library of Congress Cataloging in Publication Data
(Data available)
ISBN 0 19 963791 1 (Hbk.)
ISBN 0 19 963790 3 (Pbk.)
10 9 8 7 6 5 4 3 2 1
Typeset in Swift by Footnote Graphics, Warminster, Wilts
Printed in Great Britain on acid-free paper by
The Bath Press, Avon

Preface

Bioinformatics an emerging field

In the early eighties, the word 'bioinformatics' was not widely used and what we now know as bioinformatics, was carried out as something of a cottage industry. Groups of researchers who otherwise worked on protein structures or molecular evolution or who were heavily involved in DNA sequencing were forced, through necessity, to devote some effort to computational aspects of their subject. In some cases this effort was applied in a haphazard manner, but in others people realized the immense potential of using computers to model and analyse their data. This small band of biologists along with a handful of interested computer scientists, mathematicians, crystallographers, and physical scientists (in no particular order of priority or importance), formed the fledgling bioinformatics community. It has been a unique feature of the field that the most useful and exciting work has been carried out as collaborations between researchers from these different disciplines.

By 1985, there was the first journal devoted (largely or partly) to the subject: Computer Applications in the Biosciences. Bioinformatics articles tended to dominate it and the name was changed to reflect this, a few years ago when it was re-christened, simply, as Bioinformatics. By then, the EMBL sequence data library in Heidelberg had been running for four years, followed closely by the US based, GenBank. The first releases of the DNA sequence databases were sent out as printed booklets as well as on computer tapes. It was routine to simply dump the tape contents to a printer anyway as computer disk space in those days was expensive. This practice became pointless and impossible by 1985, due to the speed with which DNA sequence data were accumulating.

During the 1990s, the entire field of bioinformatics was transformed, almost beyond recognition by a series of developments. Firstly, the internet became the standard computer network world wide. Now, all new analyses, services, data sets, etc. could be made available to researchers across the world by a simple annoucement to a bulletin board/newsgroup and the setting up of a few pages on the World Wide Web (WWW). Secondly, advances in sequencing technology have made it almost routine to think in terms of sequencing the entire genome of organisms of interest. The generation of genome data is a completely

computer-dependent task; the interpretation is impossible without computers and to access the data you need to use a computer. Bioinformatics has come of age.

Sequence analysis and searching

Since the first efforts of Gilbert and Sanger, the DNA sequence databases have been doubling in size (numbers of nucleotides or sequences) every 18 months or so. This trend continues unabated. This forced the development of systems of software and mathematical techniques for managing and searching these collections. Earlier, the main labs generating the sequence data in the first place had been forced to develop software to help assemble and manage their own data. The famous Staden package came from work by Roger Staden in the LMB in Cambridge (UK) to assemble and analyse data from the early DNA sequencing work in the laboratory of Fred Sanger.

The sheer volume of data made it hard to find sequences of interest in each release of the sequence databases. The data were distributed as collections of flat files, each of which contained some textual information (the annotation) such as organism name and keywords, as well as the DNA sequence. The main way of searching for sequences of interest was to use a string-matching program or to browse a printout of some annotation by hand. This forced the development of relational database management systems in the main database centres but the databases continued to be delivered as flat files. One important early system, that is still in use, for browsing and searching the databases, was ACNUC, from Manolo Gouy and colleagues in Lyon, France. This was developed in the mid-eighties and allowed fully relational searching and browsing of the data base annotation. SRS is a more recent development and is described fully in Chapter 10 of this volume.

A second problem with data base size was the time and computational effort required to search the sequences themselves for similarity with a search sequence. The mathematical background to this problem had been worked on over the 1970s by a small group of mathematicians and the gold standard method was the well-known Smith and Waterman algorithm, developed by Michael Waterman (a mathematician) and Temple Smith (a physicist). The snag was that computer time was scarce and expensive and it could take hours on a large mainframe to carry out a typical search. In 1985, the situation changed dramatically with the advent of the FASTA program. FASTA was developed by David Lipman and Bill Pearson (both biologists in the US). It was based on an earlier method by John Wilbur and Lipman which was in turn based on an earlier paper by two Frenchmen (Dumas and Ninio) who showed how to use standard techniques from computer science (linked lists and hashing) to quickly compare chunks of sequences. FASTA caused a revolution. It was cheap (basically free), fast (typical searches took just a few minutes), and ran on the newly available PCs (personal computers). Now, biologists everywhere could do their own searches and do them as often as they liked. It became standard practice, in

laboratories all over the word, to discover the function of newly sequenced genes by carrying out FASTA searches of databases of characterized proteins. Fortunately, by this time the databases were just big enough to give some chance of finding a similar sequence in a search with a randomly chose gene. Sadly, the chances were small initially, but by the early nineties they had risen to 1 in 3 and now are well over 50%.

By 1990, even FASTA was too slow for some types of search to be carried out routinely, but this was alleviated by the development of faster and faster workstations. A parallel development was the use of specialist hardware such as super-computers or massively parallel computers. These allowed Smith and Waterman searches to be carried out in seconds and one very successful service was provided by John Collins and Andrew Coulson in Edinburgh, UK. The snag with these developments was the sheet cost of these specialist computers and the great skill required to write the computer code so networks were important. If you could not afford a big fast box of specialized chips, you might know someone who would allow you to use theirs and you could log on to it using a computer network.

In 1990, a new program called BLAST appeared. It was written by a collection of biologists, mathematicians and computer scientists, mainly at the new NCBI, in Washington DC, USA. It filled a similar niche to the FASTA program but was an order of magnitude faster for many types of search. It also featured the use of a probability calculation in order to help rank the importance of the sequences that were hit in the search (see Chapter 8 for some details). Probability calculations are now very important in many areas of bioinformatics (such as hidden Markov models; see chapter 4).

Protein structure analysis and prediction

Protein structure plays a central role in our understanding and use of sequence data. A knowledge of the protein structure behind the sequences often makes clear what mutational constraints are imposed on each position in the sequence and can therefore aid in the multiple alignment of sequences (Chapters 1, 3, and 6) and the interpretaion of sequence patterns (Chapter 7). While computational methods have been developed for comparing sequences with sequences (which, as we have already seen, are critical in databank searching), methods have also been developed for comparing sequences with structures (something called 'threading') and structures with structures (Covered in Chapters 1 and 2, respectively). All these methods support each other and roughly following the progression: (1) DB-search → (2) multiple alignment → (3) threading → (4) modelling. However, this is often far from a linear progression: the alignment can reveal new constraints that can be imposed on the databank search, while at the same time also helping the threading application. Similarly, the threading can cast new (structural) light on the alignment and all are carried out under (and also affect) the prediction of secondary structure.

Before the advent of multiple genome data, this favoured route often came to a halt before it started: when no similar sequence could be found even to make an alignment. However, with the genomes of phylogentically widespread organisms either completed or promised soon (bacteria, yeast, plasmodium, worm, fly, fish, man) there is now a good chance of finding proteins from each that can compile a useful multiple-sequencing alignment. At the threading stage (2) in the above progression, the current problem and worry is that there may not be a protein structure on which the alignment can be fitted. Failiure at this stage generally compromises any success in the final modelling stage (unless sufficient structural constraints are available from other experimental sources). This problem will be eased by structural genomics programmes (often associated with a genome program) for the large-scale determination of protein structures. As with the genome, these data will greatly increase the chance of finding at least one structure onto which the protein can be modelled.

The future of a mature field

With several complete genomes and a reasonably complete set of protein structures, the problems facing Bioinformatics shifts from its past challenge of finding weak similarities among sparse data, to one of finding closer similarities in a wealth of data. However, concentrating on protein sequence data (as distinct from the raw genomic DNA) eases the data processing problem considerably and the increased computation demands can be met by the equally rapid increase in the power of computers. In this new situation, perhaps all that will be needed is a good multiple sequence alignment program (such as CLUSTAL or MULTAL) with which to reveal all necessary functional and structural information on any particular gene.

The most fundamental impact of the 'New Data' is the realization that the biological world is finite and, at least in the world of sequences, that we have the end in sight. We have already, in the many bacterial genomes and in yeast, seen the minimal complement of proteins required to maintain independent life—and at only several thousand proteins, it does not seem unworkably large. This will expand by an order-or-magnitude in the higher organisms but it is already clear that much of this expansion can be accounted for by the proliferation of sequences within tissue or functionally specific families (such as the G-protein coupled receptors). Removing this 'redundancy' might still result in a set of proteins that, if not by eye, can be easily analysed by computer.

The end-of-the-line in protein structures may take a little longer to arrive, but, by implication from the sequences, it too is finite—and indeed, may be much more finite than the sequence world. This can be inferred from current data by the number of protein families that have the same overall structure (or fold), but otherwise exhibit no signs of functional or sequence similarity. Besides comparing and classifying the different structures, an interesting aspect is to develop models of protein structure evolution, perhaps allowing very distant relationships between these different folds to be inferred. It might be hoped that this

will shed light on the most ancient origins of protein structure and on the distant relationships between biological systems.

The ultimate aim of Bioinformatics must surely be the complete understanding of an organism—given its genome. This will require the characterization and modelling of extremely complex systems: not only within the cell but also including the fantastic network of cell–cell interactions that go to make-up an organism (and how the whole system boot-straps itself). However, as Sergeant Pluck has told us: what is an organism but only millions of little bits of itself whirling around and doing intricate convolutions. If a genome can tell us all these bits (and sure it will be no time till we have the genome for a sheep) then all we have to do is figure out how it all whirls around. For this, without a doubt, the Sergeant would have recommended the careful application of algebra—and, had he known about them, I'm sure he would have used a computer.

D.H. and W.T., 2000

Contents

Protocol list

Abbreviations

AACC	amino acid class covering
API	application programming interface
CASP	critical assessment in structure prediction
3-D	three-dimensional
EBI	European Bioinformatics Institute
EM	expectation-maximization
e-value	expectation value
EVD	extreme value distribution
FDF	fast data finder
FN	false negative
HMM	hidden Markov model
ILP	inductive logic programming
LAMA	local alignment of multiple alignments
MAST	multiple alignment searching tool
MDL	minimum description length
MP	membrane protein
NCBI	National Centre for Biotechnology Information
NCGR	National Centre for Genomic Research
NNSSP	nearest neighbour secondary structure prediction
PD	pattern driven
PDB	protein data bank
PHD	profile secondary structure predictions from Heidelberg
PPV	positive predictive value
PSSM	position-specific scoring matrices
QOS	quality of service
RMSD	root mean square deviation
RTT	round trip time
SAP	structure alignment program
SD	sequence driven
SP	sum of pairs
SRS	sequence retrieval system

SSE	secondary structure element
TM	transmembrane
TP	true positive
WWW	World Wide Web