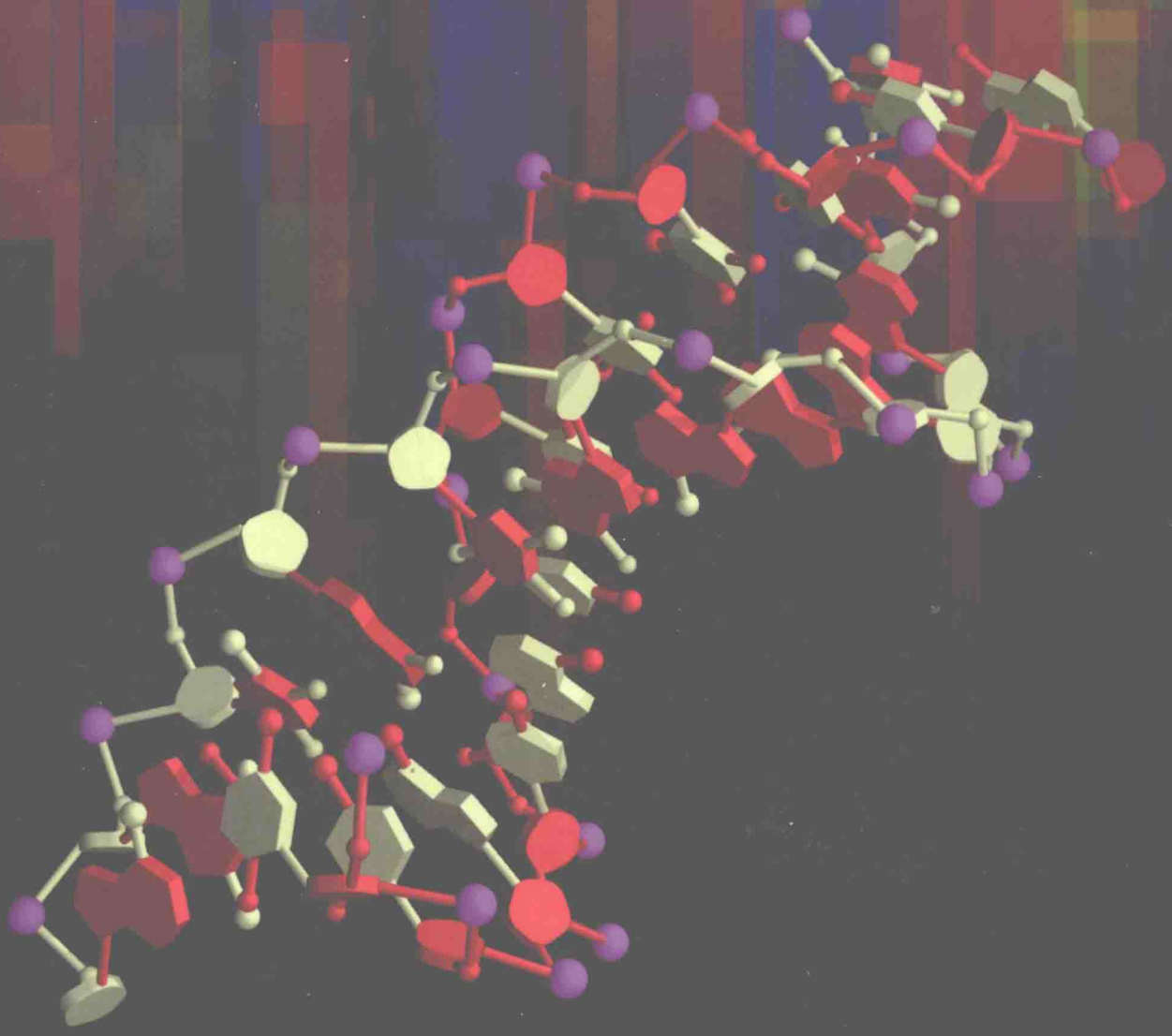
OXFORD



Introduction to Bioinformatics

Arthur M. Lesk

Introduction to

Bioinformatics

Arthur M. Lesk
University of Cambridge

In nature's infinite book of secrecy
A little I can read.

– *Anthony and Cleopatra*

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 2002, reprinted 2002

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Data available
ISBN (Pbk)0 19 925196 7
Typeset by Newgen Imaging Systems(P) Ltd, Chennai, India
Printed in Great Britain
on acid-free paper by The Bath Press, Bath

Introduction to bioinformatics

Dedicated to Eda, with whom I have merged my genes.

Preface

On 26 June, 2000, the sciences of biology and medicine changed forever. Prime Minister of the United Kingdom Tony Blair and President of the United States Bill Clinton held a joint press conference, linked via satellite, to announce the completion of the draft of the Human Genome. *The New York Times* ran a banner headline: 'Genetic Code of Human Life is Cracked by Scientists'. The sequence of three billion bases was the culmination of over a decade of work, during which the goal was always clearly in sight and the only questions were how fast the technology could progress and how generously the funding would flow. The Box shows some of the landmarks along the way.

Next to the politicians stood the scientists. John Sulston, Director of The Sanger Centre in the UK, had been a key player since the beginning of high-throughput sequencing methods. He had grown with the project from its earliest 'one man and a dog' stages to the current international consortium. In the US, appearing with President Clinton were Francis Collins, director of the US National Human Genome Research Institute, representing the US publicly-funded efforts; and J. Craig Venter, President and Chief Scientific Officer of Celera Genomics Corporation, representing the commercial sector. It is difficult to introduce these two without thinking, 'In this corner ... and in this corner...' Although never actually coming to blows, there was certainly intense competition, in the later stages a race.

The race was more than an effort to finish first and receive scientific credit for priority. Indeed, it was a race after which the contestants would be tested not for whether they had taken drugs, but whether they and others could discover them. Clinical applications were a prime motive for support of the human genome project. Once the courts had held that gene sequences were patentable – with enormous potential payoffs for drugs based on them – the commercial sector rushed to submit patents on sets of sequences that they determined, and the academic groups rushed to place each bit of sequence that *they* determined into the public domain to *prevent* Celera – or anyone else – from applying for patents.

The academic groups lined up against Celera were a collaborating group of laboratories primarily but not exclusively in the UK and USA. These included The Sanger Centre in England, Washington University in St. Louis, Missouri, the Whitehead Institute at the Massachusetts Institute of Technology in Cambridge, Massachusetts, Baylor College of Medicine in Houston, Texas, the Joint Genome

Institute at Lawrence Livermore National Laboratory in Livermore, California, and the RIKEN Genomic Sciences Center, now in Yokahama, Japan.

Both sides could dip into deep pockets. Celera had its original venture capitalists; its current parent company, PE Corporation; and, after going public, anyone who cared to take a flutter. The Sanger Centre was supported by the

Landmarks in the Human Genome Project

1953	Watson–Crick structure of DNA published.
1975	F. Sanger, and independently A. Maxam and W. Gilbert, develop methods for sequencing DNA.
1977	Bacteriophage ϕX–174 sequenced: first 'complete genome.'
1980	US Supreme Court holds that genetically-modified bacteria are patentable. This decision was the original basis for patenting of genes.
1981	Human mitochondrial DNA sequenced: 16 569 base pairs.
1984	Epstein–Barr virus genome sequenced: 172 281 base pairs
1990	International Human Genome Project launched – target horizon 15 years.
1991	J. C. Venter and colleagues identify active genes via Expressed Sequence Tags – sequences of initial portions of DNA complementary to messenger RNA.
1992	Complete low resolution linkage map of the human genome.
1992	Beginning of the *Caenorhabditis elegans* sequencing project.
1992	Wellcome Trust and United Kingdom Medical Research Council establish The Sanger Centre for large-scale genomic sequencing, directed by J. Sulston.
1992	J. C. Venter forms The Institute for Genome Research (TIGR), associated with plans to exploit sequencing commercially through gene identification and drug discovery.
1995	First complete sequence of a bacterial genome, *Haemophilus influenzae,* by TIGR.
1996	High-resolution map of human genome – markers spaced by $\sim$ 600 000 base pairs.
1996	Completion of yeast genome, first eukaryotic genome sequence.
May 1998	Celera claims to be able to finish human genome by 2001. Wellcome responds by increasing funding to Sanger Centre.
1998	*Caenorhabditis elegans* sequence published.
1 September, 1999	*Drosophila melanogaster* genome sequence announced, by Celera; released Spring 2000.
1999	Human Genome Project states goal: working draft of human genome by 2001 (90% of genes sequenced to >95% accuracy).
1 December, 1999	Sequence of first complete human chromosome published.
26 June, 2000	Joint announcement of complete draft sequence of human genome.
2003	Fiftieth anniversary of discovery of the structure of DNA. Target date for completion of high-quality human genome sequence by public consortium.

UK Medical Research Council and The Wellcome Trust. The US academic labs were supported by the US National Institutes of Health, and Department of Energy.

On 26 June, 2000 the contestants agreed to declare the race a tie, or at least a carefully out-of-focus photo finish.

The human genome is only one of the many complete genome sequences known. Taken together, genome sequences from organisms distributed widely among the branches of the tree of life give us a sense, only hinted at before, of the very great unity *in detail* of all life on Earth. They have changed our perceptions, much as the first pictures of the Earth from space engendered a unified view of our planet.

The sequencing of the human genome sequence ranks with the Manhattan project that produced atomic weapons during the Second World War, and the space program that sent people to the Moon, as one of the great bursts of technological achievement of the last century. These projects share a grounding in fundamental science, and large-scale and expensive engineering development and support. For biology, neither the attitudes nor the budgets will ever be the same. Soon a 'one man and a dog project' will refer only to an afternoon's undergraduate practical experiment in sequencing and comparison of two mammalian genomes.

The human genome is fundamentally about information, and computers were essential both for the determination of the sequence and for the applications to biology and medicine that are already flowing from it. Computing contributed not only the raw capacity for processing and storage of data, but also the mathematically-sophisticated methods required to achieve the results. The marriage of biology and computer science has created a new field called bioinformatics.

Today bioinformatics is an applied science. We use computer programs to make inferences from the data archives of modern molecular biology, to make connections among them, and to derive useful and interesting predictions.

This book is aimed at students and practising scientists who need to know how to access the data archives of genomes and proteins, the tools that have been developed to work with these archives, and the kinds of questions that these data and tools can answer. In fact, there are a lot of sources of this information. Sites treating topics in bioinformatics are sprawled out all over the Web. The challenge is to select an essential core of this material and to describe it clearly and coherently, at an introductory level.

It is assumed that the reader already has some knowledge of modern molecular biology, and some facility at using a computer. The purpose of this book is to build on and develop this background. It is suitable as a textbook for advanced undergraduates or beginning postgraduate students. Many worked-out examples are integrated into the text, and references to useful web sites and recommended reading are provided.

Problems test and consolidate understanding, provide opportunities to practise skills, and explore additional topics. Three types of problems appear at the ends of chapters. Exercises are short and straightforward applications of material in the text. Problems also involve no information not contained in the text, but require lengthier answers or in some cases calculations. The third category, 'Web–lems,' require access to the World Wide Web. Weblems are designed to give readers practice with the tools required for further study and research in the field.

What has made it possible to try to write such a book now is the extent to which the World Wide Web has made easily accessible both the archives themselves and the programs that deal with them. In the past, it was necessary to install programs and data on one's own system, and run calculations locally. Of course this meant that everything was dependent on the facilities available. Now it is possible to channel all the work through an interface to the Web. The web site linked with this book will ease the transition (see inside front cover.) To ensure that readers will be able freely to pursue discussions in the book onto the Web, descriptions of and references to commercial software have been avoided, although many commercial packages are of very high quality.

A serious problem with the web is its volatility. Sites come and go, leaving trails of dead links in their wake. There are so many sites that it is necessary to try to find a few gateways that are stable – not only continuing to exist but also kept up-to-date in both their contents and links. I have suggested some such sites, but many others are just as good. The problem is not to create a long list of useful sites – this has been done many times, and is relatively easy – but to create a short one – this is much harder!

Some computing is introduced in this book based on the widely available language PERL. Examples of simple PERL programs appear in the context of biological problems. Many simple PERL tasks are assigned as exercises or problems at the ends of the chapters.

Where might the reader turn next? This book is designed as a companion volume – in current parlance, a 'prequel' – to *Introduction to Protein Architecture: The Structural Biology of Proteins* (Oxford University Press, 2001), and that title is of course recommended. Other books on sequence analysis range from those oriented towards biology to others in the field of computer science. The goal is that each reader will come to recognize his or her own interests, and be equipped to follow them up.

I am grateful to many colleagues for discussions and advice during the preparation of this book, and to the universities of Uppsala, Umeå, Rome 'Tor Vergata' and Cambridge for the opportunity to try out this material.

I thank S. Aparicio, T. Baglin, D. Baker, A. Bench, M. Brand, G. Bricogne, R.W. Carrell, C. Chothia, D. Crowther, T. Dafforn, R. Foley, A. Friday, M.B. Gerstein, T. Gibson, T. J. P. Hubbard, J. Irving, J. Karn, K. Karplus, B. Kieffer, E.V. Koonin, M. Krichevsky, P. Lawrence, D. Liberles, A. Lister, E.L. Lesk, M.E. Lesk, V.E. Lesk, V.I. Lesk, L. Lo Conte, D.A. Lomas, J. Magré, C. Mitchell,

J. Moult, E. Nacheva, H. Parfrey, A. Pastore, D. Penny, F.W. Roberts, G.D. Rose, B. Rost, J. Sulston, M. Segal, E.L. Sonnhammer, R. Srinivasan, R. Staden, G. H. Thomas, A. Tramontano, A.A. Travers, A. Venkitaraman, G. Vriend, J.C. Whisstock, S.H. White, C. Wu, and M. Zuker for advice and critical reading.

I thank the staff of Oxford University press for their skills and patience in producing the book.

Cambridge A.M.L.
January 2002

My goal is that readers of this book will emerge with

- An appreciation of the nature of the very large amount of detailed information about ourselves and other species that has become available.
- A sense of the range of applications of bioinformatics to molecular biology, clinical medicine, pharmacology, biotechnology, agriculture, forensic science, anthropology and other disciplines.
- A useful knowledge of the techniques by which, through the World Wide Web, we gain access to the data and the methods for their analysis.
- An appreciation of the role of computers and computer science in the investigations and applications of the data.
- Confidence in the reader's basic skills in information retrieval, and calculations with the data, and in the ability to extend these skills by self-directed 'field work' on the Web.
- A sense of optimism that the data and methods of bioinformatics will create profound advances in our understanding of life, and improvements in the health of humans and other living things.

Plan of the book

- Chapter 1 sets the stage and introduces all of the major players: DNA and protein sequences and structures, genomes and proteomes, databases and information retrieval, the World Wide Web and computer programming. Before developing individual topics in detail it is important to see the framework of their interactions.
- Chapter 2 presents the nature of individual genomes, including the Human Genome, and the relationships among them, from the biological point of view.
- Chapter 3 imparts basic skills in using the Web in bioinformatics. It describes archival databanks, and leads the reader through sample sessions, involving information retrieval from some of the major databases in molecular biology.
- Chapter 4 treats the analysis of relationships among sequences – alignments and phylogenetic trees. These methods underlie some of the major computational challenges of bioinformatics: detecting distant relatives, understanding relationships among genomes of different organisms, and tracing the course of evolution at the species and molecular levels.
- Chapter 5 moves into three dimensions, treating protein structure and folding. Sequence and structure must be seen as full partners, with bioinformatics developing methods for moving back and forth between them as fluently as possible. Understanding protein structures in detail is essential for determining their mechanisms of action, and for clinical and pharmacological applications.

Contents

1 Introduction

Biology has traditionally been an observational rather than a deductive science. Although recent developments have not altered this basic orientation, the nature of the data has radically changed. It is arguable that until recently all biological observations were fundamentally anecdotal – admittedly with varying degrees of precision, some very high indeed. However, in the last generation the data have become not only much more quantitative and precise, but, in the case of nucleotide and amino acid sequences, they have become *discrete*. It is possible to determine the genome sequence of an individual organism or clone not only

completely, but in principle *exactly*. Experimental error can never be avoided entirely, but for modern genomic sequencing it is extremely low.

Not that this has converted biology into a deductive science. Life does obey principles of physics and chemistry, but for now life is too complex, and too dependent on historical contingency, for us to deduce its detailed properties from basic principles.

A second obvious property of the data of bioinformatics is their *very very large amount*. Currently the nucleotide sequence databanks contain 16×10^9 bases (abbreviated 16 Gbp). If we use the approximate size of the human genome – 3.2×10^9 letters – as a unit, this amounts to five HUman Genome Equivalents (or 2 *huges*, an apt name). For a comprehensible standard of comparison, 1 *huge* is comparable to the number of characters appearing in six complete years of issues of *The New York Times*. The database of macromolecular structures contains 16 000 entries, the full three-dimensional coordinates of proteins, of average length ~400 residues. Not only are the individual databanks large, but their sizes are increasing at a very high rate. Figure 1.1 shows the growth over the past decade of GenBank (archiving nucleic acid sequences) and the Protein Data Bank (PDB) (archiving macromolecular structures). It would be precarious to extrapolate.

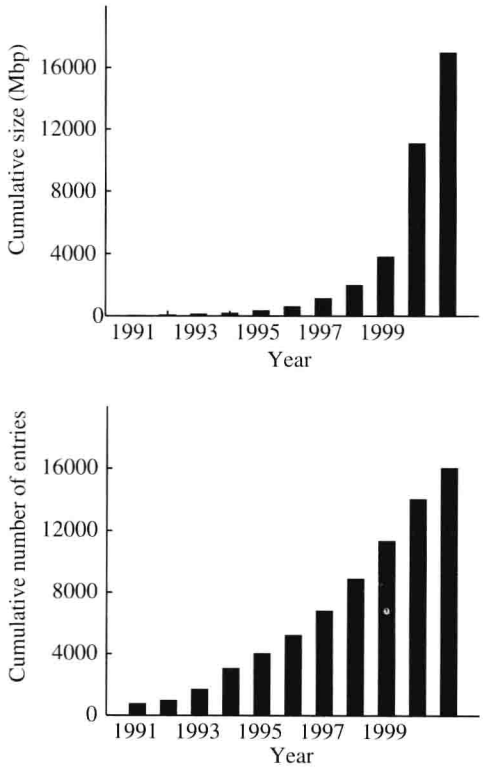


Fig. 1.1 (a) Growth of GenBank, the US National Center for Biotechnology Information genetic sequence archival databank. (b) Growth of Protein Data Bank, archive of three-dimensional biological macromolecular structures.