PREFACE TO THE FIRST EDITION

On June 26, 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 3 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 Table shows some of the landmarks along the way.

Next to the politicians stood the scientists. John Sulston, Director of the Wellcome Trust Sanger Institute in the UK, had been a key player since the beginning of high-throughput sequencing methods. He had grown with the project from the 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 Wellcome Trust Sanger Institute 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 Wellcome Trust Sanger Institute was supported by the 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 June 26, 2000 the contestants agreed to declare the race a tie, or at least a carefully out-of-focus photo finish.

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. Craig 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 UK Medical Research Council establish the Sanger Centre for large-scale genomic sequencing, directed by J. Sulston.
1992	J. Craig 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 ≈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.
September 1, 1999	*Drosophila melanogaster* genome sequence announced, by Celera Genomics; released Spring 2000.
1999	Human Genome Project states goal: working draft of human genome by 2001 (90% of genes sequenced to >95% accuracy).
December 1, 1999	Sequence of first complete human chromosome published.
June 26, 2000	Joint announcement of complete draft sequence of human genome.
2003	Fiftieth anniversary of discovery of the structure of DNA. Announcement of completion of human genome sequence.

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 subjects. Three types of problems appear at the ends of chapters. Exercises are short and straightforward applications of material in the text. Problems also require no information not contained in the text, but require lengthier answers or in some cases calculations. The third category, 'Weblems,' require access to the Worldwide 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 Worldwide 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. 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, 2000), 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.

PREFACE TO THE SECOND EDITION

Bioinformatics has grown since the first edition of this book appeared.

The most striking change has been a refocus on integration; that is, of trying to see life processes as unified systems. As I wrote at the end of *Introduction to Protein Science: Architecture, Function and Genomics*, 'During the last century, molecular biologists have been taking living things apart. Our task now is to understand how to put them back together.' We have had large amounts of data. Now we are trying to see how they interrelate. At the heart of life processes, are complicated patterns of interaction among the components, in space and in time. To understand these patterns the field has moved towards combining information into networks, and trying to understand their structures and dynamics.

Supporting this venture are the growing streams of data. The human genome, available in draft form when the first edition appeared, is now complete. It is joined by the complete genomes of 18 archaea, 155 bacteria, over 30 eukarya, and many other organelle and viral sequences. These genomes illuminate each other. One story that they tell is about unsuspected underlying unities of all living things, despite the obvious and profound differences in morphology and lifestyle.

Genomic sequences are supplemented by other data streams, notably the proteome. Knowing patterns of gene expression, and networks of regulatory interactions, shows how cells and organisms implement the information in the DNA. The potential for the life of an organism is contained in its genome, but it would be impossible to deduce a biography from it. Genomes are not formulas or scripts. It is in the proteins, and their interactions with themselves and with DNA, that we must seek the set of activities, contingent on and responsive to, the environment. Proteomics is giving us the information we need to see how the system works.

Research and applications require that the data be available in useful form. It is not enough to make the data public. The information must be subjected to quality control, annotation, and a logical structure must be imposed on it to make information retrieval possible. For this we are indebted to the institutions that archive, curate, organize and distribute the data. A recent trend has seen mergers of these groups into collaborative projects spanning the continents. In accord with the need to integrate the study of different types of data, we are moving in the direction of a single biological data repository. Individual scientists will be able to define 'virtual databanks' tailoring access to the information to suit particular needs and interests.

A gratifying consequence of academic bioinformatics is its contributions to applications in medicine, agriculture and technology. A better understanding of life processes empowers us to deal with them when they go wrong.

PREFACE TO THE THIRD EDITION

Major changes in molecular biology since the second edition most prominently involve the great growth in new complete genome sequences that have become available. These are results of enhancements in methods of sequence determination. The extension to metagenomics—the survey of distributions of sequences in a region of the earth or ocean—is new.

Major changes in information distribution involve the accelerating transition from paper to electronic libraries. A new chapter treating this subject, appears in this edition. The implications for scientific research are only a part of the great social revolution that has flowed from the development of the Web; comparable to, if not exceeding, the one impelled by the printing press 500 years ago.

There are many different possible points of view from which to present molecular biology. Bioinformatics is one of them. I have also written about genomics, and about proteins, in companion volumes also published by Oxford University Press: *Introduction to Protein Science: Architecture, Function and Genomics* and *Introduction to Genomics*. As a result, this book is focussed more tightly on the applied science of bioinformatics. Readers are urged to put the books together for a more rounded appreciation of the pageant and mechanisms of life.

PREFACE TO THE FOURTH EDITION

The natural habitat of bioinformatics is the web. Previous versions of this book recognized this, to some extent, with an Online Resource Centre supplementing the text. With this edition, the online material assumes a full partnership.

To learn bioinformatics means to understand basic concepts and principles, and to develop a set of skills. The paper text contains an exposition of the concepts and principles; the Online Resource Centre is the equivalent of a 'laboratory' or 'practical' component of the course. An icon in the text  indicates the appearance in the Online Resource Centre of material related to current discussion.

The data of bioinformatics are accessible on the web. Programs to analyse them are available on the web. Indeed, many authors of programs provide web servers for remote access to the calculations. Links from databases to servers streamline the passage from data retrieval to data analysis. Such facilities supersede the old procedure of 'download the data onto your computer, install the program on your computer, and run it locally'.

All research in contemporary molecular biology depends on data, and programs to retrieve and analyse them. There is consensus that all biomedical scientists must achieve a minimum of programming skills, but there is vigorous debate over what this minimum level should be. The point of view expressed in this book is that molecular biologists based primarily in a 'wet' lab must dip no more than their toes into the stream; those based primarily at a computer must wade in up to their waist perhaps; but only those specializing in computer science and software development must undergo total immersion.

Indeed, one of the arguments for the suggestion that sophisticated programming skills are not generally required is the great panoply of freely available programs, written by acknowledged professionals. What is essential is developing skill in *using* these programs, and in *intelligent interpretion* of the results that they produce.

This is the goal of the problems and projects in the Online Resource Centre. Many of them are 'weblems' based on data and facilities on the web. Some are programming exercises, based on the PERL language. PERL is a relatively simple but extremely effective programming language. It is one of the languages popular in the bioinformatics community. Similar languages include PYTHON and RUBY; each of these has its adherents. For PERL (and for the other languages), an extensive repertoire of utilizable program components is available, both general (see, e.g. T. Christiansen and N. Torkington, *Perl Cookbook*, 2nd edn, O'Reilly Media, Sebastopol, CA, 2003) and specialized (www.bioperl.org).

Some of the PERL exercises in the Online Resource Centre involve modifying programs. Such challenges can be more focused than writing programs from scratch. Some of the exercises, problems, and weblems, although not requiring any programming, can be solved more easily by writing short PERL programs. Readers are encouraged to try this approach whenever appropriate.

In addition to PERL, the minimal computing skills essential for a biomedical scientist would include facility with using social media for communication (it is assumed that readers are familiar with Facebook and YouTube, but there are others that are in use for communication among scientists), and the ability to create a website. Studying from this book and the Online Resource Centre affords an opportunity to practise these skills. You might, for instance, 'turn in' the answers to homework assignments by gathering them into a web page. Questions about statements that you and the other students found unclear in your instructor's lectures—or, conceivably, even in this book—could be shared and discussed in a blog. Indeed, there is now a trend to integrating websites and social media. However, there are security issues. Your instructor might be unhappy if everyone copied the answers to the exercises from the first student to post them. A class taught from this book would afford a fine opportunity to explore the possibilities and challenges.