Introduction to Computational Genomics

Where did HIV and SARS come from? Have we inherited genes from Neanderthals? How do plants use their internal clock? How do odor receptors function? The genomic revolution in biology, started in the late 1990s, enables us to answer such questions. But the revolution would have been impossible without the support of powerful computational and statistical methods that enable us to exploit the massive amounts of genomic data. Accordingly, many universities are introducing the courses to train the next generation of bioinformaticians: biologists who are fluent in the language of mathematics and computer science, and data analysts who are familiar with the problems and techniques of biology. The entry cost into this field is very high, requiring knowledge of several disciplines and of a large and fast-growing literature. Students need a road map to navigate entry to this field. This readable and entertaining book, based on successful courses taught at the University of California and elsewhere, provides that. It guides the reader step by step through some of the key achievements of bioinformatics. The hands-on approach makes learning easier and experimentation natural, as well as equipping the reader with the necessary tools and skills. Statistical sequence analysis, sequence alignment, hidden Markov models, gene and motif finding, gene expression data analysis and reconstruction of evolutionary relations between genomes are all introduced in a rigorous yet accessible way. A companion website provides the reader with all the data used in the examples, links to publicly available software, and `Matlab`™ demonstration for reproducing and adapting all the steps performed in the book.

Nello Cristianini is Professor of Artificial Intelligence at the University of Bristol.

Matthew Hahn is Assistant Professor, Department of Biology and School of Informatics, Indiana University.

Introduction to Computational Genomics

A Case Studies Approach

Nello Cristianini and Matthew W. Hahn

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521856034

First published 2007

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN-13 978-0-521-85603-4 hardback
ISBN-10 0-521-85603-5 hardback

ISBN-13 978-0-521-67191-0 paperback
ISBN-10 0-521-67191-4 paperback

Contents

Preface

Nothing in biology makes sense except in the light of evolution.

Theodosius Dobzhansky

Modern biology is undergoing an historical transformation, becoming – among other things – increasingly data driven. A combination of statistical, computational, and biological methods has become the norm in modern genomic research. Of course this is at odds with the standard organization of university curricula, which typically focus on only one of these three subjects. It is hard enough to provide a good synthesis of computer science and statistics, let alone to include molecular biology! Yet, the importance of the algorithms typical of this field can only be appreciated within their biological context, their results can only be interpreted within a statistical framework, and a basic knowledge of all three areas is a necessary condition for any research project.

We believe that users of software should know something about the algorithms behind the results that are presented, and software designers should know something about the problems that will be attacked with their tools. We also believe that scientific ideas need to be understood within their context, and are often best communicated to students by means of examples and case studies.

This book addresses just that need: providing a rigorous yet accessible introduction to this interdisciplinary field, one that can be read by both biologically and computationally minded students, and that is based on case studies. It has evolved from a course taught at UC Davis, where both authors were doing research in computational biology, one coming from computer science (N.C.) and the other from biology (M.W.H.).

The authors had to understand the other's field in order to do research in the hybrid science of computational genomics. The goal of this book is to develop a simple, entertaining, and informative course for advanced undergraduate and graduate students. Based on carefully chosen case studies, the chapters of this book cover ten key topics that we feel are essential to a scientist conducting research in bioinformatics and computational genomics. We will be satisfied if at the end of this first course the reader is able to understand and replicate the main results from the classic papers in this field.

This book benefited from the input of many colleagues and students. In particular, many of the case studies have been based on projects developed by students, post-docs, and research visitors, as well as teaching assistants. In particular we want to thank Elisa Ricci, Tara Thiemann, Margherita Bresco, Chi Nguyen, Khoa Nguyen, and Javannatah Gosh. The authors have benefited from discussions with various outstanding colleagues, and it is difficult to list them all. However in particular we want to thank Linda Bisson, Stacey Harmer, Wolfgang Polonik, Dan Gusfield, Sergey Nuzhdin, Lucio Cetto, Leonie Moyle, Jason Mezey, and Andrew Kern. The first draft has been proof-read by various colleagues: Chris Watkins, John Hancock, Asa Ben Hur, Tom Turner, Quan Le, Tara Thiemann, Vladimir Filkov, and Rich Glor. N. C. acknowledges support from NIH grant R33HG003070-01.

The book's website contains the data, the algorithms and the papers used in the case studies, and much more. It can be found at the URL

`www.computational-genomics.net`

Bristol, Bloomington

Prologue

In praise of cells

- How cells work
- What is a genome
- The computational future of biology
- A roadmap to this book

The physicist Richard Feynman is credited with jump-starting the field of nanotechnology. In a talk at Caltech in December 1959, Feynman issued a famous challenge: he would pay $1000 to anyone who could write the entire *Encyclopedia Britannica* on the head of a pin. Feynman calculated that the size of the area was approximately 1/16 of an inch across (about 1.6×10^{-3} meters), and that in order to fit all 42 million letters of the *Encyclopedia* one would have to make each letter 1.0×10^{-8} meters across. It took (only) 26 years before the prize was finally claimed by a graduate student at Stanford University.

Now, consider the problem of having to write out the entire set of instructions needed to build and operate a human, and consider having to do so in each of the trillions of cells in the body. The entire human genome is 3.5 billion "letters" long, and each cell is only 2 microns (2×10^{-7} meters) across. (Actually, two complete copies of the genome are present in each cell, so we have to fit a bit more than 7 billion letters.) However all the organisms on earth overcome these packaging problems to live and prosper in a wide range of environments.

In the same 1959 lecture Feynman also imagined being able to look inside a cell in order to read all of the instructions and history contained within a genome. A few decades later the genomic era began – a time when technological advances in biology and computational advances in computer science came together to fulfill Feynman's dream. Bioinformatics and computational genomics make it possible to look inside a cell and read how an organism functions and how it got to be that way. This book endeavors to be a first course in this new field.

Why bioinformatics?

How are all forms of life related? What was the first cell like? How do species adapt to their environments? Which part of our genome is evolving the fastest? Are we descendents of Neanderthals? What genes are responsible for major human diseases? Why do we need new flu vaccines every year?

Modern biology is a goldmine of fascinating questions, and never before have we been so close to the answers. The main reason for this is the new, data-driven approach to biological investigation spawned by the availability of large-scale genomic data. The availability of these data has triggered a revolution in biology that can only be compared to the revolution in physics at the beginning of the twentieth century. The effects of this revolution have been felt in other fields of science, as well. Application of genomic

technology to medicine, drug design, forensics, anthropology, and epidemiology holds the promise to improve our life and enlarge our understanding of the world.

Almost as important as this scientific revolution is the cultural revolution that has accompanied it. Many of the questions asked in modern biology can only be answered by computational analysis of large quantities of genomic data. Researchers in computer science and statistics have been recruited to this effort to provide both a conceptual framework and technological support. Biologists, computer scientists, and statisticians now work together to analyze data and model living systems to a level of detail that was unthinkable just a few years ago. The impact of this collaboration on biology has been invaluable and has lead to the new discipline of bioinformatics.

Soon, a new kind of scientist (with knowledge in computer science, statistics, mathematics, biology, and genetics) will arise. Most major universities have already started various types of degrees in bioinformatics and are drawing students with a wide range of backgrounds. The purpose of this book is to provide a first course in the questions and answers to problems in bioinformatics and computational genomics (because many people have preconceived notions of the term "bioinformatics," we use these two phrases interchangeably). We hope to provide the biology necessary to computational researchers – though we obviously cannot cover everything – and the algorithms and statistics necessary to biologists. All of this is in the hope of molding a new type of researcher able to ask and answer all (or almost all) of the questions in modern biology.

A bit of biology

One of the most fundamental questions, the one that underlies many others, is: How do cells work? For both unicellular and multicellular organisms, we want to know how cells react to their environment, how genes affect these reactions, and how organisms adapt to new environments. The general picture is known, but many of the details are missing. Modern biology aims to answer this question in detail, at a molecular level. Here we review some of the most basic ideas in biology to provide the minimum knowledge needed to conduct research in bioinformatics. We stress at the outset that biology is a field of exceptions: all of the generalizations and rules we introduce here will be wrong for some organisms, but covering all of the exceptions would take another book. Throughout the text, we have tried to note when there are important exceptions that bear on the examples given.

Every organism's dream (so to speak) is to become two organisms. An organism reproducing faster, or exploiting its environment more efficiently, rapidly out-competes its rivals for resources. This was the basic point made by Darwin and is vital to understanding the way cells and organisms work. The conceptual framework of evolution is the most fundamental aspect of biological thinking and allows us to organize and interpret all of the data we will be analyzing in this book. No analysis of the genetic differences between individuals in a species, or between different species, makes sense outside of an evolutionary framework. Over the 3.5 billion years life has been on this planet, organisms have become extremely streamlined and efficient, shaped to a large extent by the evolutionary process of natural selection. If we want to understand cells, we have to understand both the power and the limitations of natural selection.

Conversely, in order to understand natural selection, we also must understand much about basic biology.

There are two basic types of cells, those with and without nuclei (called *eukaryotic* and *prokaryotic* cells, respectively). These types of cells largely share the same fundamental molecular machinery, but prokaryotes are simpler, unicellular organisms (such as bacteria), while eukaryotes are often more complex and include both unicellular and multicellular organisms (such as fungi, animals, and plants).

Unicellular organisms are the simplest free-living things; their ability to interact with the environment, derive the energy and materials needed to continually fabricate themselves, and then eventually to reproduce is controlled by a complex network of chemical reactions. Even the individual cells that make up a multicellular organism must each perform thousands of such reactions. These chemical reactions (called metabolism as a whole) are the very essence of cellular life: a cell needs to process various nutrients found in its environment, to produce the components it needs to operate, and then to breakdown components no longer needed. These tasks are carried out via biochemical processes that can be finely controlled by the cell. Each reaction needs to be catalyzed (triggered) by specific proteins – special molecules produced by the cell itself. Many proteins are enzymes, a kind of molecule involved in nearly every activity of the cell. Other proteins are used as structural elements to build cellular parts, as activation or repression agents to control reactions, as sensors to environmental condition, or take part in one of the many other tasks necessary for cellular function. There are thousands of different proteins in each cell, often specialized to control one specific reaction, or to be part of a specific cellular structure. Producing these proteins not only requires the cell to obtain energy and materials, but also requires detailed communication between different parts of a cell or between cells. Much of the cellular machinery is devoted simply to ensuring the production of proteins at the right moment, in the right quantity, in the right place.

A protein is a large molecule formed by a chain of amino acids, which folds into a characteristic shape. The same 20 basic amino acids are used by all known organisms. The exact composition of the chain (which amino acids are in which order) determines its shape, and its shape determines its function – i.e. which reactions it will facilitate, which molecules it will bind to, etc. The need to produce thousands of proteins means that a cell must have a way to remember the recipe for each of them, as well as have a way to produce them at the right time.

A cell's most reliable way to pass on the recipe for making proteins is contained in its genetic material and is passed on to daughter cells at each division. The machinery for reading this information is one of the core components of all living things and is highly similar in all types of cells; the machinery itself is formed by a complex of enzymes, specified by the very instructions it must read! This self-referential, auto-poietic, aspect of life can be mind-boggling.

The genetic material used by cells is formed by molecules of DNA (deoxyribonucleic acid), which have a sequential structure that enables them to act as information storage devices. The way in which they store the recipe for proteins and the information needed to control their production will be discussed in Chapters 1 and 2.

The quest to understand the way in which DNA is used by organisms to pass on genetic instructions has spanned the last two centuries of biology. The initial steps were taken in 1859 by a Moravian monk named Gregor Mendel. Mendel discovered that genetic information is contained in discrete units (what we now call *genes*), passed from generation to generation. The second major step came in 1944, when a group in New York led by Oswald Avery showed that nucleic acids were the molecules used to encode this information. Finally, with the proposal of the structure of DNA by James Watson and Francis Crick in 1953, the mechanism for the replication and retrieval of information stored in a DNA sequence was found. What came in the years following these discoveries has been an incredible series of events, with biologists unraveling the exact way in which proteins are specified by DNA, and revealing how cells use genetic information to synthesize proteins.

The future of biology

Although the big picture came to emerge gradually in the last decades of the twentieth century, it also became increasingly clear that the size and complexity of organisms meant that a detailed understanding of their inner-workings could not be achieved by small-scale experiments. By the end of the century it became possible to automate the acquisition of this knowledge and thus to collect gigabytes of data in a short period of time. The invention of sequencing machines that can read the entire DNA sequence of a bacterium in only a day, and of a larger eukaryote in a month, as well as machines that can identify and quantify all of the genes active in a cell or a tissue, has ensured a steady flood of biological information for the foreseeable future. The analysis of these data promises to be the biggest challenge to biology in the twenty-first century.

The details of the roles played by different proteins in cellular reactions, how these reactions are organized into pathways (whereby the product of one reaction becomes the substrate of another reaction), and how pathways are organized into a complex network that must continually reproduce itself are now the questions that biologists can address. In addition, crucial questions concerning the way in which genes are responsible for differences between species, between individuals of the same species, and the role genes play in evolution can be answered by a large-scale analysis of the entire collection of genetic material contained within each individual.

But there are also many simpler questions that we still cannot answer satisfactorily: How many different proteins do organisms have? How is their production coordinated? How did they arise? What – if not proteins – does the majority of DNA in a cell code for?

It is the aim of this book to provide the tools necessary to answer the above questions by computational analysis of genomic data. The ten chapters of this book cover ten topics that we feel are necessary to a scientist conducting research in bioinformatics and computational genomics. Below we outline these ten topics.

A roadmap to this book

This book is divided into ten chapters, each presenting a major idea or task in computational genomics. On top of this structure, however, the book is divided into three main threads. Chapters 1–4 provide the tools necessary to annotate

a genomic sequence – to describe the main features and structures found in a genome, and ideally also their function. In Chapters 5–8 we learn how to go from treating the genome as a static edifice to the dynamic, evolving object that it truly is. Finally, because each of the first eight chapters has looked solely at DNA sequences, we turn to the analysis of gene expression in Chapters 9 and 10, showing how this can be used to identify the function of genes and other structures found in the sequence. We show how the analysis of data produced by DNA microarrays differs from sequence analysis, but we also show how it can be synthesized with sequence data to reveal even more about the inner workings of cells. Below is a short synopsis of each of the ten chapters.

Chapter 1 describes the major features of a genome, by using as a leading example the first genomic sequence of a free-living organism ever obtained: that of the bacterium *Haemophilus influenzae*. We show how to retrieve and handle genomic data, perform some simple statistical analysis, and draw some conclusions. The chapter also introduces probabilistic models of biological sequences and important notation and terminology. After this chapter, the reader will be able to download and manipulate DNA sequences from public databases, and understand their statistical properties.

Chapter 2 explains what genes are and how to find them in a DNA sequence by locating particular regions called open reading frames (ORFs), again in the case of simple bacterial sequences. It also deals with other statistical signals to be found in genome sequences, and discusses a crucial point: how to assess the significance of a pattern and how to report it in terms of p-values. This chapter will enable the reader to find candidate genes and assess the significance of their findings.

Chapter 3 deals with the important algorithmic issue of assessing sequence similarity, the standard way to detect descent from a common ancestor. To this purpose the chapter introduces the technology of sequence alignment as an indispensable tool of bioinformatics, describing in detail the basic pairwise global and local alignment algorithms (based on dynamic programming) as well as briefly discussing multiple alignment and fast pairwise alignment algorithms (such as BLAST). This chapter will enable the reader both to decide if two given DNA sequences are likely to be homologous and to understand how to use common alignment tools.

Chapter 4 uses the example of odorant-receptor proteins to introduce another of the algorithmic workhorses of the field of bioinformatics: hidden Markov models (HMMs). This class of probabilistic models for sequences (and signals therein) underlies many modern gene finding algorithms, but is also used in sequence segmentation, multiple alignment, etc. The chapter demonstrates how to detect change points in the statistical make-up of biological sequences – a task that can help to identify features such as horizontally transferred segments of DNA – and how to summarize all the features of a protein family into a single probabilistic description. The reader should then be able to determine the likelihood that a protein belongs to a certain family, and therefore whether already annotated proteins can be used to assign function.

Chapter 5 introduces the issue of genetic variation among individuals of the same species, by comparing genetic sequences of Neanderthal and *Homo sapiens*. The fascinating question of our relation with these ancient inhabitants of Europe can be entirely answered by analyzing publicly available DNA

sequences, and in the process we can learn about single nucleotide polymorphisms (SNPs) and statistical models of sequence evolution. In order to account for the probability of multiple substitutions in DNA sequences, and hence to obtain better assessments of the genetic distance between individuals, the Jukes–Cantor and Kimura 2-parameter models are derived. An analysis of DNA from various apes also hints at fundamental questions about human origins. The reader will be able to assess genetic distance between sequences, understand the mathematics behind the models, and apply this to real data.

Chapter 6 directly addresses the question of sequence evolution under natural selection. A sequence evolves with different rates if it is under selective pressure either to change or to stay constant, and this selective pressure can be quantified by using statistical models and appropriate algorithmic methods. The example of HIV evolution is used in this chapter to illustrate how certain locations of this fast-evolving virus change at a high rate – to keep ahead of the immune system of the host – while others are fairly conserved. Evolution of drug resistance follows similar patterns, and can be similarly detected. The reader will become familiar with the computation and the interpretation of the K_a/K_s ratio on real sequence data.

Chapter 7 takes these ideas one step further, showing how it is possible to reconstruct the evolutionary history of a set of homologous sequences by constructing phylogenetic trees. This is not just important for evolutionary studies, but can have many practical applications, as is demonstrated by the case study of the SARS epidemic. In late 2002 a virus jumped from an animal to a human in China, triggering a violent epidemic that spread to many countries before being identified and isolated. But its time, place, and host of origin, as well as the trajectory followed by the infection, can be reconstructed by an analysis of the viral genetic sequences. Simple algorithms and advanced concepts of phylogenetic analysis are presented, including the basic neighbor-joining algorithm, and more advanced and sophisticated approaches. These methods are also used to answer questions about the origin of HIV, and to address questions about early human evolution. The reader will learn to construct phylogenetic trees from sequence data.

Chapter 8 discusses one of the most recent applications of computational genomics, namely whole-genome analysis of multiple species. This involves large-scale genomic comparisons between different species, and if the species are chosen carefully it can provide a wealth of information, from helping to identify functional regions to reconstructing the evolutionary mechanisms that led to speciation. We take the complete genomes of different species of *Chlamydia*, an internal parasite of eukaryotic cells, and we see how they differ from major large-scale rearrangements of the same genes. We also identify syntenic regions, gene families, and distinguish between orthologous and paralogous genes. The reader will become familiar with the basic concepts and tools of whole-genome analysis.

In Chapter 9 we address another major source of genomic information: gene expression data collected by using DNA microarrays. Exploiting patterns found in this type of data requires using pattern recognition technology, a mix of statistics and computer science. We demonstrate the power of this approach to functionally annotate genomes by studying the case of yeast. A series of landmark papers in the late 1990s introduced the analysis of gene expression

data by looking just at yeast genomes, and these studies are repeated in this chapter. The main tools are presented, including data processing, clustering, classification, visualization, and applied to the detection of cell-cycle regulated genes. The reader will be able to perform basic tasks of data mining with gene expression data, and to understand the assumptions underlying the most common algorithmic approaches.

Finally, in Chapter 10 we discuss the integration between expression and sequence information, by studying the circadian clock in plants. Genes regulated by the internal clock (as opposed, for example, to genes responding to external stimulations) can be identified by gene expression analysis, and clustered according to the time phase of their cycle. The upstream sequences of genes of equal phase can reveal common patterns, candidate binding sites for regulatory proteins. This analysis can be performed by the reader, illustrating how sequence and expression information can be synthesized, to annotate not only protein coding but also regulatory regions.

Many more important topics and approaches exist in computational genomics, but in order to make this introduction as gentle as possible, we have selected the above ten themes as representatives of the style of analysis typically found in this exciting scientific domain. More advanced approaches should be more easily accessible to the readers once they have become familiar with the contents presented in this book. Sections marked with a $*$ can be skipped at a first read.

Reading list

A general understanding of molecular biology, genetics, and evolution are all essential for researchers in computational genomics. This can be obtained in many introductory textbooks of biology, as well as in more specialized introductions to the field of genomics. The reader may refer to Brown (1999) and to Gibson and Muse (2004) for a general introduction to genomics and evolution, or follow the links in the book's website to online introductory material about molecular and cell biology. The lecture by Richard Feynman on nanotechnology can be found in the article Feynman (1960). The book's website:

`www.computational-genomics.net`

contains links to introductory articles and other online material.