

PREFACE

We currently see a vast amount of information being generated as a result of experimental work in biomedicine. Particularly impressive is the development in DNA sequencing. As a result, we are now facing a new era of genomics where a lot of different species, as well as many different human individuals, are being analysed. There are many important biological questions being addressed in such genome-sequencing projects, including questions of medical relevance. A critical technical part of all these projects is computational analysis. With the large amount of sequence information generated, computational analysis is often a bottleneck in the pipeline of a genomics project. Therefore, there is great demand for individuals with the appropriate computational competence. Ideally, such individuals should not only be proficient in the relevant mathematical and computer scientific tools, but should also be able to fully understand the different biological problems that are posed. This book was partly motivated by the urgent need for bioinformatics competence due to recent developments in genomics.

A student or scientist may enter into bioinformatics from different disciplines. This book is written mainly for the biologist that wants to be introduced to computational and programming tools. There are certainly books out there already for that type of audience. However, I was attracted by the idea of assembling a book that would cover a large number of relevant biological topics and, at the same time, illustrate how these topics may be studied using relatively simple programming tools. Therefore, an important principle of the book is that it will attempt to convince the reader that relatively simple programming is sufficient for many bioinformatics tasks and that you need not be a programming expert to be effective. Another important principle of the book is that I wanted the bioinformatics examples to be very practical and explicit. Thus, the reader should be able to follow all the details in a procedure all the way from a biological problem to the results obtained through a technical approach. As one demonstration of this principle, all files and scripts mentioned in this book are available for download at www.cambridge.org/samuelsson. This means the reader is able to try it all out on his/her own computer. I also wanted this book to illustrate the interdisciplinary nature of bioinformatics. Therefore, I have chosen to include a substantial amount of biological motivation as well as programming technology. As a result, the book has a number of rather sudden transitions from descriptions of biological topics to very technical computing matters.

This book is intended as a guide to Perl and Unix-based computing tools for students with a background in molecular biology, biochemistry, cell biology or genomics who have no previous background in this type of computing. In addition, PhD students and scientists at all levels in these fields who want to be introduced to such programming tools should hopefully benefit from the book. The computational parts of the text should be easy to understand for a student lacking a background in computer science; the programming examples presented are at a fairly basic level. The book is complemented by exercises for further study, with mixed levels of difficulty. For the benefit of the student without a mathematical background, the book is by and large non-mathematical, avoiding topics such as probability theory. In summary, the reader of this book should be any student or scientist with some insight into biology, but who also wants to learn about bioinformatics at a more technical level. I also think of the book as being of potential interest to the student or scientist with a background in computer science or programming, but who seeks biological motivation and wants to know more about biological problems that are typically addressed in genomics and bioinformatics.

I present a number of biological and medical topics related to DNA and protein sequences and show how they may be exploited using bioinformatics tools. The book inspires from a biological point of view by selecting relevant and interesting examples; some of the examples will be understood also by non-biologists. Many of the biology topics presented in the book are related to human genomics or human disease, emphasizing the importance of bioinformatics in human medicine. The examples chosen are mainly in the field of sequence bioinformatics. This is a classic area of bioinformatics that has been described previously in many textbooks, but is enjoying renewed interest following current developments in genomics. 'Personal genomics', as touched upon in the final chapters, will be an important area in biomedicine and clinical medicine.

The material in this book is divided into a number of major biological or bioinformatical topics; gene technology, human disease, evolution, gene function, information resources, gene identification methods and personal genomes. Within each of these topics there are different examples of problems that require bioinformatics tools.

The material is organized from the perspective of sequence bioinformatics. First, simple sequence operations such as translation and pattern matching are presented in Chapters 1–5. Chapter 6 deals with RNA secondary structure, and there is a discussion of pairwise alignments and sequence similarity searches in Chapters 7 and 11. Multiple alignments and molecular phylogeny are covered in Chapters 8–10. Different methods of functional assignment are discussed in Chapters 11–13, while molecular sequence databases are discussed in Chapter 14. Finally, gene prediction methods are covered in Chapters 15–17.

From a computational point of view the book focuses on the Unix operating system and the Perl programming language as these are the predominant bases of computational tools in the area of bioinformatics. The Perl content is also organized in a specific fashion, with new concepts introduced in each chapter. For this reason, it is a good idea to read the chapters of the book in the order they are presented. Should a reader contemplate studying the chapters in a different order, Appendix III, providing a short reference guide to Perl, might be helpful. The Perl examples are at a fairly simple level throughout the book, although the Perl code tends to get somewhat more complex towards the end. As mentioned above, a major principle in the design of the book is to convince the reader that relatively simple programming is sufficient to handle many common biological problems. It should also be pointed out that this book is not a complete Unix or Perl reference. In addition, there are more advanced areas of Perl programming that are not covered, such as references and object-oriented programming. A reader seeking information on such topics should consult additional books, such as those listed in Appendix III.