
Preface

Broadly speaking, Bioinformatics can be defined as a collection of mathematical, statistical and computational methods for analyzing biological sequences, that is, DNA, RNA and amino acid (protein) sequences. Numerous projects for sequencing the DNA of particular organisms constantly supply new amounts of data on an astronomical scale, and it would not be realistic to expect that biologists will ever be able to make sense of it without resorting to help from more quantitative disciplines.

Many studies in molecular biology require performing specific computational procedures on given sequence data, for example, simply organizing the data conveniently, and therefore analysis of biological sequences is often viewed as part of computational science. As a result, bioinformatics is frequently confused with *computational sequence analysis*, which is somewhat narrower. However, understanding biological sequences now increasingly requires profound ideas beyond computational science, specifically, mathematical and statistical ideas. For example, the protein folding problem incorporates serious differential geometry and topology; understanding the evolution of sequences is dependent upon developing better probabilistic evolutionary models; analysis of microarray data requires new statistical approaches. Generally, when one goes beyond algorithms and starts either looking for *first principles* behind certain biological processes (which is an extremely difficult task) or paying more attention to *modeling* (which is a more standard approach), one crosses the boundary between computational science and mathematics/statistics. These days many mathematicians are becoming interested in bioinformatics and beginning to contribute to research in biology. This is also my personal story: I am a pure mathematician specializing in several complex variables, but now I work in bioinformatics as well.

Despite this mathematical trend, bioinformatics is still largely taught by computational groups and computer science departments, and partly by engineering and biology (in particular, genetics) departments. Naturally, in courses developed by these departments emphasis is placed on algorithms and their implementation in software. Although it is useful to know how a particular

piece of software works, this software-oriented education does not always reveal the mathematical principles on which the algorithms are based. Such incompleteness may lead to certain problems for the graduates. Suppose, for example, that a commonly used model is implemented in software and the students are taught how to use it. Of course, the model makes some simplifying *assumptions* about the biological processes it attempts to describe, and these assumptions are buried in the mathematical core of the model. If the students are taught only how to use the software and are not taught the mathematical foundations of the model, they will know nothing about the assumptions and therefore *limitations* of the model; this in turn means that they will not be able to interpret correctly the results of applying the software to biological data.

This situation with education in bioinformatics is now beginning to change as mathematics departments around the world are starting to teach this subject. I have been teaching two bioinformatics courses at the Department of Mathematics of the Australian National University (ANU) in Canberra, for two years now. When I started teaching them I quickly realized that all the textbooks that I found on the subject were skewed towards computational issues, reflecting, of course, the dominant teaching culture at the time. Those textbooks were not very satisfying from a mathematician's point of view and were unacceptable for my purposes. What I needed was a clear and mathematically rigorous exposition of procedures, algorithms and models commonly used in bioinformatics. As a result, I began writing my own lecture notes, and eventually they formed the basis for this book.

The book has two parts corresponding to the two courses. The first course is for second-year students and requires two medium-level first year mathematics courses as prerequisites. It concerns four important topics in bioinformatics (sequence alignment, profile hidden Markov models, protein folding and phylogenetic reconstruction) and covers them in considerable detail. Many mathematical issues related to these topics are discussed, but their probabilistic and statistical aspects are not covered in much depth there, as the students are not required to have a background in these areas. The second course (intended for third-year students) includes elements of probability and statistics; this allows one both to explore additional topics in sequence alignment, and to go back to some of the issues left unexplained in the first course, treating them from the general probabilistic and statistical point of view.

The second course is much more demanding mathematically because of its probabilistic and statistical component. At the same time, the chapters on probability and statistics (Chaps. 6 and 8) contain very few proofs. The path taken in these chapters is to give the reader all the main constructions (for instance, the construction of probability measure) and to illustrate them by many examples. Such a style is more gentle on students who only have taken a couple of mathematical courses and do not possess the mathematical maturity of a student majoring in mathematics. In fact, this is the general approach taken in the book: I give very few proofs, but a lot of discussions and examples.

Nevertheless, the book is quite mathematical in its logical approach, rigor and paying attention to subtle details.

Thus, for someone who wants to get a mathematical overview of some of the important topics in bioinformatics but does not want to go too deeply into the associated probabilistic and statistical issues, Part I of the book is quite sufficient. But it should be stressed that without reading Part II, one's understanding of various procedures from Part I will be incomplete.

Although Parts I and II together cover a substantial amount of material, none of the topics discussed in the book is treated comprehensively. For example, the chapter on protein folding (Chap. 4) and the one on phylogenetic reconstruction (Chap. 5) could each easily be expanded into a separate book. The amount of material included in the book is what realistically can be taught as two one-semester courses. Certainly, if the probability and statistics components of the book are taught separately in a different course, one can fit in more genuine bioinformatics topics, for example, the analysis of microarray data, currently not represented in the book at all.

The book concentrates on the mathematical basics of bioinformatics rather than on recent progress in the area. Even the material included in the book is found quite demanding by many students, and this is why I decided to select for it only a few topics in bioinformatics. Thus, this book is by no means a comprehensive guide to bioinformatics.

This is primarily a textbook for students with some mathematical background. At the same time, it is suitable for any mathematician, or, indeed, anyone who appreciates quantitative thinking and mathematical rigor, and who wants to learn about bioinformatics. It took me a substantial effort to explain various bioinformatics procedures in a way suitable for a general mathematical audience, and hence this book can be thought of, at least to some extent, as a translation and adaptation of some topics in bioinformatics for mathematicians. On top of this, the book contains a mathematical introduction to statistics that I have tried to keep as rigorous as possible.

I would like to thank my colleagues Prof. Sue Wilson and Prof. Simon Eastaugh of the Mathematical Sciences Institute (MSI) and the Centre for Bioinformation Science (CBiS) at the ANU who first suggested that I should turn my lecture notes into a book and encouraged me during the course of writing. I would like to thank Prof. Peter Hall of the MSI for patiently answering my many questions on the theory of statistics. Finally, I am grateful to Prof. John Hutchinson of the Department of Mathematics for encouragement and general discussions.

Canberra,
March 2004

Alexander Isaev