error-detecting capabilities, and minimum distance properties of higher order dc-constrained codes were likewise discarded.

The text would surely benefit from an expanded treatment of recording codes designed for PRML, the channel technology that caused a sea change in magnetic recording during the late 1990s and is now in ascendance in optical recording. The advent of PRML spawned new directions in constrained coding research, including distance-enhancing codes, high-rate parity coding, and architectures for combined detection and decoding, all of which have found wide use in commercial storage devices.

Kees Immink deserves praise for this latest update of his classic 1991 book. Comprehensive in scope, thoughtfully organized, and written with both the motivated student and the practicing engineer in mind, the volume should find a place on the bookshelf of anyone with an interest in coding techniques for data storage systems. Visit the Shannon Foundation Publishers website for information on ordering a copy. And once you’re there, follow the link to Kees Immink’s personal website where you can enjoy the photo of him receiving a well-deserved Emmy for his pioneering contributions to recording code technology.

REFERENCES


Statistical Methods in Bioinformatics (Second Edition)—
Warren J. Ewens and Gregory R. Grant (New York: Springer-Verlag, 2005). Reviewed by M. Vidyasagar

It is an oft-repeated truism that the first half of the 21st century belongs to biology, just as the last half of the twentieth century belonged to microelectronics in its various forms (transistors, VLSI, microprocessors, systems on a chip, the worldwide web, etc.). As with all truisms, there is an element of truth, but also a lot of hype. The popular image of biology as predominantly an experiment-based science has now given way to the realization that biology is also now as much information-based as it is experiment-based. Rapid advances in experimental methods, such as high throughput DNA sequencing, microarrays (gene chips and now protein chips) and other related techniques, now permit biologists to generate data at a truly unprecedented pace. Coping with this flood of data and extracting useful “information” out of raw data is one of the major challenges that confronts today’s biologists. Bioinformatics is the name given to a loosely defined body of mathematical and computational methods that seek to address this challenge.

“Bioinformatics” is thus the intellectual successor to earlier branches of science that had names such as theoretical biology, mathematical biology, and so on. But there is an important difference. Bioinformatics

is perhaps the first theoretical subject that addresses not only the underlying mathematics of a biological problem, but also the computational complexity of any solution algorithms developed.

Scientists know that science proceeds mostly through evolution, and rarely if ever by revolution. However, the attention of the popular media and of society at large is captured by discrete events in time, which are then said to usher in a “revolution.” In this particular instance, the mapping of the human genome in February 2001 was considered to be a seminal event that would magically unlock all of the mysteries of life, if only we had sufficiently powerful computers at our disposal. The mapping of the human genome was rapidly followed by the publication of the genomes of several other organisms; so much so that Genbank, the repository of all-known genome sequences, holds (as of September 5, 2006) 1,796,667 sequences, totalling 80,369,977,826 base pairs in all. Similarly, the Protein Database (PDB) holds 38,620 structures of proteins as of September 5, 2006. It is interesting to note that both the Genbank and PDB seem to be doubling in size every six to nine months. The well-known (and often misquoted) Moore’s law of VLSI states that the computational power of computers will double, and their cost will halve, every eighteen months. Thus, the conclusion is unavoidable: The glut in “raw data” being generated by various biology laboratories around the world cannot be addressed merely by Moore’s law alone. New theoretical methods continue to be needed.

The promise of a genomic revolution being ushered in the mapping of the human genome also turned out to be premature. The readers of the Information Theory Transactions cannot be expected to be familiar with this point, but practically all known drugs are aimed at just about 500 “drug targets,” that is, proteins produced by the human body. New drugs are supposed to function by inhibiting or enhancing the action of the target protein (depending on whether the action of the protein is beneficial or not). During the initial days after the human genome was mapped, computational methods turned up as many as 5000 potential drug targets—roughly an increase by a factor of ten! However, validating these targets required a great deal of expensive and time-consuming laboratory work, thus belying to a large extent the supposed benefits of post-genomic biology.

It would however be wrong to swing to the opposite end of the spectrum and declare that computational biology is not important. Problems of determining which parts of a genome correspond to genes (and which parts are “junk DNA”), predicting the three-dimensional structure of a protein, determining the function of a gene or a protein by assessing its similarity to other known genes and proteins, are all computationally intensive problems whose solution continues to be relevant to biology, and its application to drug discovery which is the biggest commercial application of such work.

Thus, to repeat, bioinformatics is as relevant a topic today as it was during the late 1990s and early 2000s. Interestingly, much of the theoretical work and algorithmic development in this field took place as early as the 1980s, when computational power was strictly limited, and networking of computers practically unknown.

The book under review represents a serious attempt to gather, in a single book, many if not most of the popular statistical methods used in bioinformatics. The book is the second edition of an earlier book with the same title that was published in 2001. As there is no Preface to the second edition, it is difficult to determine the extent of the changes made from the first edition. At a superficial glance, it appears to me that about 90% of the material is the same. Thus the review can be applied practically in tact to the first edition as well.

In the preface (to the first edition) dated February 2001, the authors say “We take bioinformatics to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to the study of very large biological, and particularly genetic, data sets.”

Manuscript received September 12, 2006.

The reviewer is with Tata Consulting Services, Hyderabad 500081, India (e-mail: sagar@atc.tcs.co.in).

Communicated by S. Verdú, Associate Editor for Book Reviews.

Digital Object Identifier 10.1109/TIT.2006.885536
This is an apt summary of bioinformatics as a subject. Elsewhere in the preface the authors say “Our aim is to give an introductory account of some of the probability theory, statistics, and stochastic process theory appropriate to computational biology and bioinformatics. This is not a “how to” book, of which there are several in the literature, but it aims to fill a gap in the literature on the statistical and probabilistic aspects of bioinformatics.” I have chosen this particular passage as a benchmark against which the actual contents and exposition can be measured.

The book is of epic proportions (597 pages), and consists of 15 chapters plus four appendices, as follows:

1) Probability Theory (i): One Random Variable
2) Probability Theory (ii): Many Random Variables
3) Statistics (i): An Introduction to Statistical Inference
4) Stochastic Processes (i): Poisson Processes and Markov Chains
5) The Analysis of One DNA Sequence
6) The Analysis of Multiple DNA or Protein Sequences
7) Stochastic Processes (ii): Random Walks
8) Statistics (ii): Classical Estimation Theory
9) Statistics (iii): Classical Hypothesis Testing Theory
10) BLAST
11) Stochastic Processes (iii): Markov Chains
12) Hidden Markov Models
13) Gene Expression, Microarrays, and Multiple Testing
14) Evolutionary Models
15) Phylogenetic Tree Estimation
   A Basic Notions in Biology
   B Mathematical Formulæ and Results
   C Computational Aspects of the Binomial and Generalized Geometric Distribution Functions
   D BLAST: Sum of Normalized Scores

A detailed comparison with the table of contents of the first edition shows that, for all practical purposes, there are only three changes in the second edition: Chapter 3 is considerably enlarged in the second edition; Chapters 8 and 9 used to be a single chapter in the first edition; and Chapter 13 in the second edition is completely new.

From the table of contents it is obvious that the book is vast in its sweep, beginning with the fundamentals of a single random variable, and encompassing advanced topics such as continuous-time Markov chains (albeit on a finite state–space), hidden Markov models, and the like.

The authors say in the preface that “The material in this text assumes little or no background in biology.” In accordance with this statement, Appendix A entitled “Basic Notions in Biology” consists of precisely three pages. Taken very literally, that statement is indeed true. While all the examples are taken from biology (in particular genomics or proteomics), a reader who is willing to accept on faith that these problems are interesting and important by themselves will in no way be hampered by a lack of familiarity with biology. On the other hand, my past experience in reading mathematical texts suggests that, unless one is passionate about the problems that are attempted to be studied, it is not possible to focus one’s attention for very long on the book at hand. In short, mathematics is definitely not recreational reading. My point therefore is that the book would make for extremely dry reading except for a person who is interested in computational biology, and can thus be expected to have some familiarity with at least the basic concepts of biology.

From reading the book it is obvious that the authors are quite knowledgeable about both the underlying mathematics as well as the application of the theory to various problems in bioinformatics. Though the title speaks about “statistical” methods, there is some discussion about nonstatistical methods as well, for example, the alignment of two or more sequences, which can be solved via a purely deterministic algorithm. However, for the most part the book focuses on statistical methods.

I would summarize the book by saying that practically every important statistical technique that is used in bioinformatics is touched upon in the book. The list of references is exhaustive and well researched. Thus, the book is a valuable reference source for anyone who wishes to find out what a particular technique is about, and to find out the relevant references in the literature. Having said that, I also believe that the book cannot really be used for self-study, especially by a mathematically trained person. There is a lot of material that tells the reader what a particular subject is about, but in almost all cases the treatment stops short of being comprehensive. Because the authors have attempted to begin from first principles in probability, the level of exposition in the book is not uniform, but is somewhat variable. Naturally, the authors wish to keep the mathematical level as elementary as possible. As a consequence, at times they are forced to leave out vital details. Rather than deal in generalities, I will quote a specific passage to illustrate the point; see p. 264:

The derivative (with respect to $\theta$) of the right-hand side of (7.16) is 0. Since there are infinitely many possible values of $N$, the left-hand side is an infinite sum of functions of $\theta$. A basic theorem of calculus gives a criterion indicating when this derivative of a sum of functions can be calculated as the sum of the derivatives of the individual functions. This criterion holds in this case, so that . . .

What the authors wish to say is the following: “If we differentiate the infinite series term by term with respect to $\theta$, the resulting series is still uniformly convergent, so it is acceptable to do so.” The trouble with the above kind of exposition is that it satisfies neither the mathematically minded nor the mathematical novice. The novice neither knows nor cares that he/she is not supposed to differentiate inside the summation sign, without some safeguards. The mathematically minded person would like to know precisely why it is permissible to differentiate inside the summation, instead of being told that some vague, unnamed theorem in calculus allows him/her to do so.

A glance at the Index reinforces the point I am trying to make. I find it difficult to imagine why a book would have both Kolmogorov–Chapman equation and L’Hôpital’s rule in the Index. A person who is encountering L’Hôpital’s rule for the first time in this book is not going to be able to understand the Kolmogorov–Chapman equation. Conversely, a person who is able to grasp the Kolmogorov–Chapman equation should know L’Hôpital’s rule in his/her sleep.

Chapter 10 on BLAST (Basic Linear Alignment Search Technique) is among the longest chapters in the book. It can be argued that the advent of BLAST in the early 1990s by Simon Altschul, Samuel Karlin, and their co-workers, was indeed a seminal event in sequence alignment. To see why, let us suppose that $u, v$ are two sequences over a finite alphabet, of length $n$ and $m$ respectively. For example, $u, v$ can be nucleotide sequences assuming values in the four-symbol alphabet $\{A, C, G, T\}$, or the 20-symbol alphabet of amino acids. A fundamental problem in biology (specifically genomics and/or proteomics) is to find the best possible “match” between $u$ and $v$. If we insist on a perfect match between the two sequences, or subsequences thereof, the problem is easy to solve. But in biology one can never expect to find “perfect” matches. Instead one opts for optimal “gapped” alignment. This particular problem was solved as far back as 1981 (before the age of massive computation) via the so-called “Smith–Waterman algorithm.” This algorithm has complexity $O(nm)$, or quadratic complexity if both strings are of comparable length.

Unfortunately, the complexity is exponential in the number of strings. Thus if one attempts to find the optimal gapped alignment of $k$ strings $u_1, \ldots, u_k$, then the complexity is $O(2^{kT})$, where $T$ is the length of the longest string (assuming all strings are of comparable length). The work of Altschul, Karlin, Dembo, Zeitouni, and others cuts through this Gordian knot of exponential complexity by presenting a
probabilistic method of doing the optimal alignment. Though the letter B in BLAST stands for “Basic,” in fact there is nothing “basic” about the algorithm, as it makes use of extremely sophisticated asymptotic estimates of the tail probabilities of various random events. There are several versions of BLAST, depending on which stochastic model one uses for the sequences to be aligned (e.g., i.i.d., sample path of a Markov chain, etc.). Rather unfortunately, since the inception of the theory there has not been a widely accessible treatment of the theory. (Naturally, the original research papers are written in the terse style that is appropriate for peer-reviewed journals.) Thus the authors had a real opportunity to write a definitive treatment of BLAST theory for the mathematically minded. I am persuaded, on the basis of the contents of the book, that the authors are sufficiently knowledgeable to have written such a chapter. Instead, the basic problems are stated, and ways of addressing the problems using BLAST are presented—but the theory itself is mostly unstated. Thus, a comprehensive textbook-level treatment of BLAST still remains to be written. I consider it highly desirable for someone to write such a book (or chapter), because the basic statistical assumptions that underly the BLAST algorithm do not always hold in biological problems. Thus anyone attempting to modify BLAST theory to cater to different statistical models would first need to understand the existing theory thoroughly.

In spite of these quibbles, I did feel that I had learnt a great deal by reading the book. There is virtually no other book I know of, in which all the topics are listed in one place, and almost all of the literature is collected in the bibliography. The treatment of each mathematical or statistical topic, though it might be just adequate to whet the appetite, does indeed do at least that job quite well. Thus, a reader who is interested in any one topic can use this book as a starting point for reading, and then build up his/her knowledge by following up with the sources cited in the bibliography. For the purposes of teaching, the book is perhaps well suited for an expert teacher, as he/she would be able to do the supplementary reading and integration of the material alluded to in the preceding sentence. A nonexpert teacher might find it more difficult to give a thorough treatment of any of the topics listed here.