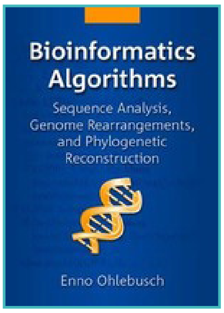
Bioinformatics Algorithms - Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction



Enno Ohlebusch

Oldenbusch Verlag, 2013

ISBN: 978-3000413162

604 pp.

Josè R. Valverde

CSIC, Centro Nacional de Biotecnología, Madrid, Spain

Received 11 June 2014; **Published** 16 July 2014

Valverde JR (2014) *EMBnet.journal* **20**, e781. http://dx.doi.org/10.14806/ej.20.0.781

Competing Interest: none

There are many books on bioinformatics in circulation, many of them dealing with issues concerning how analytical methods are implemented, and the algorithms that underpin them. I was therefore curious to know whether a new book on Bioinformatics Algorithms was actually needed, and whether this book really could provide something that others didn't. Consequently, in this review, I will try to give an idea of what the book offers, what you may expect from it, and who is most likely to benefit from it.

The roots of bioinformatics were, and still are, firmly embedded in the management and analysis of biological sequence information. This book therefore focuses on the core technologies that underlie modern sequence analysis in the context of genomics and phylogeny.

The book starts with a general introduction (chapter 1), followed by traditional string-comparison methods (chapter 2), and quickly moves to the core of modern genomic methods: suffix arrays (chapters 3 and 4). Suffix arrays have gained increasing popularity, as they provide an efficient way to perform linearly scaling queries of huge textual data-sets; common practical applications of these algorithms are thus presented in chapter 5. Chapters 6 and 7 then address how to make these algorithms and data structures more efficient, introducing methods that work with compressed data, such as the Burrows-Wheeler Transform. These methods address exact string matching efficiently (in linear time), and have direct applications in problems such as genome assembly and short-read mapping. The traditional approaches to sequence comparison are then introduced in chapter 8, where the Needleman-Wunsch algorithm is described, followed by methods used to build multiple sequence alignments and whole genome alignments, touching on topics such as genome rearrangements. Chapter 9 deals with sorting by reversals to introduce methods that can be used to address these issues. Finally, chapter 10 ties everything together in a clear exposition of a practical application: phylogenetic analysis.

The overall layout is organised as a book on Computer Science (CS). This means that algorithms are minutely described, generally with an accompanying step-by-step walk-through using a small example data-set, followed by detailed algorithm validation and complexity analysis of its time and space requirements in 'big O' notation. The descriptions are clear, concise and illustrated with many opportune Figures, easy to follow and understand. The description of algorithm goals, and validation and complexity analysis, use a formal language that will appeal to pure computer scientists.

The orientation towards CS is also shown in the topic layout: purely algorithmic issues are often presented before their practical application in bioinformatics, often with forward references to later chapters. Many surrogate techniques related to these algorithms (e.g., SVMs, which use the kernel methods presented), alternative basic algorithms (e.g., Smith-Waterman or BLAST) or methodologies (evolutionary algorithms, machine learning, clustering or modern statistical methods, etc.) that would deviate too much from the central line of discourse are omitted. Similarly, many biological applications are described only to the extent needed to demonstrate the application of the algorithms (e.g., the chapter on phylogeny describes traditional techniques but does not delve into methods like Bayesian inference); this should not pose problems for computer scientists and practical bioinformaticians who are already familiar with these techniques.

Summarising, this is a very good, readable CS book on the core techniques of sequence analysis, as seen from the point of view of a modern family of algorithms (derived from suffix arrays) that have acquired major relevance in bioinformatics and other text-analysis fields, and are slowly overtaking most traditional techniques. The book does a good job of presenting them in the context of their application to genome analysis.

Advanced computer scientists will enjoy the detailed formal analysis of most algorithms in the book. The average pragmatic bioinformatician will probably be more interested in the (very good) description of the algorithms and their practical consequences (time and space complexity). Hence, I believe the book is most likely to appeal to seasoned computer scientists and CS students, but will also appeal to practical bioinformaticians who want to get up to date in modern genomics research, and to practitioners of other text-analysis fields.

After reading it in full, I enjoyed this book and found it informative, inspiring and entertaining. It is well written and readable, and although it is not a general bioinformatics text, it succeeds in explaining a complex family of methods that underlie novel applications.