Review of[1]

Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R

by Gabriel Valiente
Chapman & Hall/CRC Press, 2009
368 pages, HARDCOVER

Review by
Anthony Labarre (Anthony.Labarre@cs.kuleuven.be)

Department of Computer Science
K. U. Leuven
Celestijnenlaan 200A - bus 2402
3001 Heverlee
Belgium

1 Introduction

Gabriel Valiente's book is about *combinatorial pattern matching*, a field whose object of study consists of problems of the form: "find one, k, or all occurrence(s) of a given pattern in another given object". The nature of the objects that can be considered varies greatly, encompassing among others texts, trees, graphs, or images, all of which are relevant in the context of computational biology, which is the motivation for and the field of application of this book.

The usefulness of efficient algorithms for searching texts is obvious to anyone who has ever used a computer. This is also the case in biology, where the advent of genome sequencing has led to a dizzying number of databases containing the full sequences of a huge number of organisms. Text search algorithms are particularly relevant in that setting, and are routinely used in discovering patterns in sequences or finding common subsequences in order to gain more insight into the evolution of living beings and their relationships, as well as into the relationships between genotypes (the "source code" of an organism) and phenotypes (the set of physical traits and characteristics of an organism).

But the relevance of combinatorial pattern matching to computational biology goes far beyond text algorithms. The field also has applications in phylogenetics, which focuses on representing the evolution of a set of species using a tree (or sometimes a network) whose leaves are the species under consideration and whose internal nodes represent putative unknown common ancestors. Many techniques for reconstructing phylogenetic trees are known, and there is an obvious need to be able to compare the results these methods yield, which can be achieved by identifying common substructures in the resulting trees. The applications explained above also serve as motivations for having techniques that allow to find patterns in graphs, and to express the similarities between them; indeed, graphs are also used to model evolution in the same way as phylogenetic trees, and model chemical reactions, proteins, RNA structures, protein interaction networks, and metabolic pathways to name but a few.

2 Summary

The book is organised into three large parts, according to the type of structure under consideration: Part I deals with sequences, which constitute the simplest, the most natural and also the most studied setting. Part II is concerned with trees, whose study is motivated by phylogenetic reconstruction methods and RNA structure comparison, while Part III considers graphs, which are useful in representing evolution and biochemical reactions.

Each part is structured in the same way and consists of three chapters: the first chapter states some basic definitions and facts about the structures that are being investigated in that part, introducing simple algorithms which are used as building blocks in the sequel, as well as implementation notes on how to represent the data structures under consideration, listing already implemented options whenever available. The algorithms and operations introduced in that introductory chapter are not about pattern matching, but are part of the general background and are used later on; examples of tasks that are carried out in this chapter include traversing trees and graphs, complementing DNA sequences, generating structures and counting them.

The second chapter introduces some of the building blocks of pattern matching in the context of the given structures. These methods include the fundamental task of finding an occurrence, or several occurrences of a word (resp. a tree, or a graph) in a given text (resp. in a graph), and how to compare two structures. Comparison of sequences is achieved by aligning them, while in the case of graphs it is carried out by computing various statistics on the graphs under comparison, like the set of distances between two terminal nodes or the sets of paths in both graphs, and then comparing those statistics.

Finally, the third chapter examines more advanced tasks, like those of finding common substructures in several objects. This differs mostly from what is tackled in the second chapter by the nature of the information that is extracted. For instance, in the case of graphs, one can look for partitions induced by the removal of an edge, or the subgraphs induced by a particular set of terminal nodes, and then comparing the sets obtained from both structures to be compared. Other more advanced tasks include subgraph isomorphism and maximum agreement common subgraphs.

Valiente concludes the book with two appendices, where the basics of Perl and R are presented in a very didactic way to people who know little or nothing about those languages. The author walks the reader through the design of two simple scripts, offers a brief overview of both languages, and presents what he calls "quick reference cards", which are concise lists summarising the syntax and usage of essential functions in these languages.

3 Opinion

This book seems geared mostly towards people who have had little previous exposure to computer science or programming and who are, or want to become active in bioinformatics. More specifically, I would imagine the primary audience of this book to be biologists getting to know about programming, or undergraduate students in computer science getting started in bioinformatics. It is also very light from a mathematical point of view.

Valiente clearly decided to approach the subjects he treats in his book in breadth rather than in depth, with an emphasis on actually *teaching* those subjects. This kind of book is especially helpful in the context of interdisciplinary fields, of which computational biology is a perfect example, where professionals with a particular background are led to learn aspects from a field that is very different

from their initial training. People who are already familiar with combinatorial pattern matching may benefit from the source code included in the book[2], but if they want to deepen their knowledge of the subjects it covers, I would recommend them to have a look at Gusfield's book [1] (especially for Part I), Valiente's previous book [3] for more advanced topics and algorithms (for Parts II and III), which also contains implementations (although in C++), and Huson et al.'s book [2] for more information on phylogenetic networks, which Valiente discusses in Part III.

As hinted above, the author focuses primarily on basic tasks in combinatorial pattern matching, particularly on those that are most relevant to computational biology. I like the "hands-on" approach this book offers, and the very pedagogical structure it follows, by first clearly stating the problems and their relevance to biology, explaining how they will be solved, giving the algorithm in pseudocode, and then in actual Perl and R source code. The book also has tons of examples, thoughtfully chosen and beautifully laid out, illustrating the various definitions, concepts and problems under consideration. Of course, some basic algorithms have already long been implemented in Perl and R, but it is worth explaining them nonetheless. For other tasks, Valiente informs the reader about existing formats and libraries (the famous BioPerl[3] module for instance) so that he or she does not reinvent the wheel. There is a judicious mention of things that do *not* exist too, which also serves as a justification for showing how things could be done using the aforementioned libraries. This is also useful in making a choice between the two languages that are suggested to you. A lot of alternative options are of course available, including Python and its well-furnished Biopython module[4], for which it is not difficult to adapt the source code snippets presented in this book.

The book contains no proof or theorems, but has references to the relevant literature, should the reader be eager to know more mathematical facts about the objects being dealt with. This is consistent with the approach chosen by Valiente, but it makes me wonder why some counting problems, while interesting, are considered, or why the sequences they yield are not linked to the corresponding entries in OEIS[5]. I was a little bit let down by the index, which could have been thicker. This is a minor issue, but it suggests that you should not take too long to read a particular chapter, otherwise retrieving the definition of that particular term that you forgot may take you some time. Other than that, the book is very well-written and accessible, undoubtedly written by an author who takes great care in preparing his manuscripts and teaching about an area he enjoys working on.

References

[1] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology.* Cambridge University Press, 1997.

[2] Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications.* Cambridge University Press, 2011. http://www.phylogenetic-networks.org/.

[3] Gabriel Valiente. *Algorithms on Trees and Graphs.* Springer, 2002.

[2] Also freely available from the author's webpage at `http://www.lsi.upc.edu/~valiente/comput-biol/`.

[3] See `http://www.bioperl.org/wiki/Main_Page`.

[4] See `http://biopython.org/wiki/Main_Page`

[5] The On-Line Encyclopedia of Integer Sequences, see `https://oeis.org/`.