Review of *ReCombinatorics: The algorithmics of ancestral recombination graphs and explicit phylogenetic networks*
Author of Book: *Dan Gusfield*
Published by MIT Press
Hardcover, $60, 600 pages

Review by
Steven Kelk (steven.kelk@maastrichtuniversity.nl)
Department of Knowledge Engineering (DKE)
Maastricht University, The Netherlands

1 Overview

One of the central mathematical abstractions in the study of evolution is the *phylogenetic tree*. Essentially, this is a rooted tree in which the leaves are bijectively labelled by a set of contemporary species X, and the internal nodes of the tree model the biological events (such as speciation) that caused the root of the tree - representing a hypothetical common ancestor - to diversify into the set X. From a mathematical perspective, the central challenge in phylogenetics is to infer the topology of the tree given only measurements obtained from the species X at its leaves, such as DNA sequence data. There are many different optimality criteria for this tree-inference problem, most of them NP-hard, and the literature is vast. In recent years there has been growing attention for the fact that evolution is not always tree-like. In particular, due to "fusing" biological phenomena such as hybridization, recombination and lateral gene transfer, evolution is sometimes better modelled as a rooted, leaf-labelled directed acyclic graph (DAG), where nodes of indegree two or higher are used to model the fusion phenomenon in question. Such DAGs directly generalise phylogenetic trees and are known by various different names in the literature, the two most commonplace being *phylogenetic network* and *Ancestral Recombination Graph (ARG)*. In this book Dan mainly uses the term ARG, and indegree two nodes are called recombination nodes, reflecting the population-genetic origin of his work. Unlike the literature on tree-inference, the literature on ARG-inference is comparatively new and small, and in this book Dan gives an overview of some of the main algorithmic results in this area from the last twenty years.

The core of the book concerns the problem of inferring an ARG with a *minimum* number of recombination nodes. More formally, the input is a binary matrix M with n rows and m columns, where the n rows can be viewed as length-m binary strings that label the n leaves of the ARG we are trying to infer. Specifically, we wish to infer an ARG and a labelling of its internal nodes with length-m binary strings, such that (i) per column there is a mutation (i.e. a transition from 0 to 1 or vice-versa) on at most one edge of the ARG and (ii) the sequences labelling recombination nodes are formed by concatenating a prefix of the string labelling one parent, with a suffix of the string labelling the other. An ARG with a minimum number of recombination nodes is called a $MinARG$, and $Rmin(M)$ is used to denote the minimum value itself. Computing $Rmin(M)$ is, inevitably, NP-hard, and notoriously intractable. The majority of the chapters of this book are devoted to the algorithm engineering challenge of computing (bounds on) $Rmin(M)$ in practice. Remaining chapters discuss related but somewhat different problems (such as the problem of "phasing" genotypes into haplotypes) and biological applications, but computation of $Rmin(M)$ is certainly the dominant theme of this book.

2 Summary of Contents

The book has fourteen chapters and pleasingly, it is completely self-contained. Although the book addresses a biologically-motivated problem it is primarily an algorithms book. Anybody with a mathematics or computer science background, and a familiarity with the basics of algorithm design (e.g. big-O analysis) will comfortably be able to read this book. There are frequent references to the biological underpinnings and applications but Dan presents them in a very computer-science friendly way which contextualises the algorithmic results, rather than distracts from them. Indeed, Dan is a professor of computer science and the whole book has a "journey of discovery" feel, documenting how an algorithms expert (armed with the traditional array of weapons against NP-hard problems) fared as he dug deeper into the world of ARGs.

The first two chapters, encompassing sixty pages, introduce the biological context, give the necessary definitions (including basic graph theory terminology) and present a number of classical results from the literature on constructing phylogenetic trees. The book very quickly settles into a pattern which is repeated in almost all chapters: biological/historical context, definition, theorem, proof, extensions. The MinARG / $Rmin(M)$ problem is defined and motivated in the third chapter. The fourth chapter is much less mathematical than surrounding chapters, discussing three applications of recombination analysis in practice. Dan himself notes that this chapter can be skipped by those readers only interested in the algorithms/mathematics.

The algorithm engineering begins proper in the fifth chapter, covering sixty pages, in which at least four different lower bounds on $Rmin(M)$ are analysed in detail. This focus on algorithmic bounds (and computing $Rmin(M)$ in practice by sharpening lower and upper bounds until they meet) is characteristic of the algorithmic ARG literature, contrasting with other parts of the literature where authors have focussed more on approximation algorithms and fixed parameter tractability. Some of the lower bounds presented can be computed in polynomial time, some are composites of other bounds, and some are themselves NP-hard to compute. As an illustration, one bound is based on the observation that each pair of columns inducing a certain forbidden submatrix must be separated by at least one recombination node. Leveraging the linearly ordered nature of the data, this naturally leads to a "Hitting Set on intervals of the line" formulation that can be solved in polynomial time with a greedy algorithm. The bound can subsequently be improved by relaxing "intervals of the line" to "subsets of the line" but this makes the problem NP-hard. At this point Dan deploys one of his favourite tools: Integer Linear Programming (ILP). He is a strong proponent of using ILP in computational biology and here, as in many other parts of the book, it turns out to work very well when computing bounds, or solving NP-hard subproblems related to computation of $Rmin(M)$, on realistic-sized datasets. (The book includes an appendix which is essentially a crash-course in modelling with ILP). ILP cannot, however, be used to compute $Rmin(M)$ directly, so it is emphatically *not* the case that computation of $Rmin(M)$ can be solved in practice simply by unleashing industrial-strength solvers such as CPLEX or Gurobi. Indeed, this explains the - at first glance - rather exotic choice to develop lower bounds that are themselves NP-hard to compute: unlike $Rmin(M)$ itself they have natural static formulations which are amenable to formulation as ILPs. It is also worth noting that ILP is by no means the only response to NP-hardness encountered in the book: in various places other options such as exponential-time dynamic programming, and combinatorial branch and bound, are explored.

The sixth and seventh chapters concern a natural decomposition theorem which sometimes allows computation of $Rmin(M)$ to be significantly simplified (and can also be used for developing lower bounds). The core idea is very natural. Specifically, the forbidden-submatrix obstruction

described earlier naturally leads to the *conflict graph*, in which each column of M corresponds to a node and there is an edge between two nodes if they form an obstruction. It turns out that, in some circumstances (but by no means all) the connected components of this graph can be processed independently of each other.

Building on this idea, the next chapters see the book switch to upper bounds i.e. algorithms that actually construct ARGs. The eighth chapter takes a restricted approach, demonstrating that $Rmin(M)$ can be computed in polynomial time if at least one MinARG has a highly restricted topology. Namely: if every biconnected component of the underlying undirected graph, contains at most one recombination node. Such ARGs are called *galled trees* - elsewhere in the literature they are called *level-1 networks* - and in many ways they are the first, most natural step up from trees. Due to their very simple structure they behave very well with respect to the decomposition theorem presented in chapters six and seven, quickly leading to a divide-and-conquer approach. Chapter nine, however, addresses the much more challenging problem of computing $Rmin(M)$ when no assumptions are made about the output topology of the MinARG (or the input matrix M). This is the chapter where the intractability really begins to bite. The issue, as observed elsewhere in the literature, is that computer science as a whole has little experience with optimization problems in which graphs (in this case, DAGs) are the output, rather than the input. The algorithms in this section are thus tantamount to (intelligent) exhaustive search of the space of all ARGs, sometimes accelerated with branch-and-bound style techniques, and this is computationally formidable. Although it does not improve the running time, this chapter also introduces an idea which researchers working elsewhere in the literature will immediately recognise: that the columns of M can be naturally be partitioned into intervals which have a common, tree-like history, and that partial information about the topology of these trees (known as *marginal trees* here) can be extracted directly from M. This information is useful, because we know that the MinARG must somehow simultaneously topologically embed all these marginal trees. Hence, at least to a certain extent, construction of MinARGs and computation of $Rmin(M)$ can be re-formulated as a "tree-packing" problem.

Chapter ten discusses two further lower bounds, one of which can be viewed as a matrix partitioning problem, and the other which can be regarded as a type of minimum-length elimination ordering problem. In terms of flow these two bounds would ideally have found a place in chapter four, together with the other lower bounds, but as Dan explains they both require ideas developed in the intervening constructive chapters. Chapter eleven is very short and develops some (not entirely surprising, and mostly descriptive rather than constructive) necessary and sufficient conditions for a MinARG to be amenable to the divide-and-conquer approach that was earlier demonstrated to be correct for galled trees.

At this point the book makes a rather sharp jump. In chapter twelve it (temporarily) moves away from computation of $Rmin(M)$, to focus on the so-called Haplotype Inference problem. Here the input is a matrix G over the alphabet $\{0, 1, 2\}$, which represents a set of *genotypes*, and the goal is to identify a binary matrix M (where each row represents a *haplotype*) such that every row of G can be expressed as the "sum" of some two rows of M, where "summation" is defined as: $0+0=0$, $1+1=1$, $0+1=2$ and $1+0=2$. Combinatorially this is a very interesting problem, with very many variations. The most basal variant - does such a matrix M exist for a given input matrix G? - is shown to be reducible in polynomial-time to the classical and well-understood Graph Realization problem. More sophisticated variants of the problem again bring ARGs back into the equation. For example, for a given G, can we identify an M such that $Rmin(M)$ is minimized? Chapter thirteen

also has a rather different flavour to the rest of the book, discussing at a comparatively high-level the use of ARGs in *(Genome Wide) Association Studies* which is a technique for identifying combinatorial patterns in the genome that seem to be causal for diseases. This chapter is much less algorithmic than the others, briefly giving an overview of various different models and techniques, and will be one of the more accessible chapters for biologists.

The final chapter of the book is, from the perspective of unifying the literature in this area (where mathematically isomorphic models often have multiple different terminologies) very important. It emphasizes that $Rmin(M)$ and some of its lower bounds can be re-formulated within models that attempt to quantify the topological discordance between a given set of phylogenetic trees (i.e. where the input is a set of trees, rather than a binary matrix M). These models, and related optimization problems such as the Hybridization Number and Maximum (Acyclic) Agreement Forest problems, have been very well-studied outside the ARG literature. It is commendable that Dan emphasizes these commonalities, and this also explains the elaborate title of the book: it is an attempt to cover two separate nomenclatures. The final chapter also touches on an elegant link between multi-state problems in phylogenetics (i.e. where the input matrix is over a larger alphabet) and chordal graph theory.

3 Overall

This is a very well-written and self-contained book which gives a comprehensive overview of algorithmic results concerning ARGs, and everything is referenced rigorously. As stated at the beginning of the review, it is certainly not a book for biologists - it is full of proofs - although they will certainly appreciate the way Dan ties the algorithmic results to the applications context. For researchers already working in the area the algorithmic results will not be so surprising, but this is hardly to be expected given that the book primarily serves to integrate and summarize seminal results from the last twenty years. I classify myself in this last group, but I nevertheless greatly enjoyed the exposition, and particularly appreciate the effort Dan takes to point to further reading on the statistical and probabilistic side of the story. This is important because, although the book is not statistical at all, a great deal of biomathematics certainly is, and understanding that this dimension exists is critical to understanding the role of combinatorial optimization in this area.

I do have some negative points, but they are mainly stylistic. In a few places the book goes into detail which, for an algorithmic audience, is superfluous. I particularly felt this was the case in the chapters about the decomposition theorem and, related to this, galled trees. These results are not so surprising yet they are presented with a little too much bravoure. Also, the book sometimes leans a little bit too much in the direction of algorithm engineering. That is, it occasionally devotes attention to the details of optimizing the running times of already competitive polynomial-time subroutines when the great challenges in this area lie at the other end of the spectrum i.e. dealing with the severe intractability of the core NP-hard problems. Personally I would also have attempted to dissect this NP-hardness a little more, pushing the analysis more towards (fixed) parameterized complexity, and exploring the polyhedral dimension when appropriate e.g. some of the polynomial-time algorithms presented could equivalently be formulated as totally unimodular linear programs (which are guaranteed to have integral solutions). These points, however, are all a question of taste: this is unquestionably a good book.

Let me conclude by recommending this book to three groups in particular. First: researchers already working in the area who are looking for a reference text on algorithms for ARGs, com-

plete with biological motivation. Secondly, traditional algorithms researchers who are looking for a combinatorially clean entrance-point to computational biology, explained by somebody with a computer science background. Thirdly, this book could quite easily be used as the scaffolding for an entire MSc (or advanced BSc) algorithms course. All the standard tools for polynomial-time algorithm design, and dealing with NP-hardness, are evident here, reinforced by the motivation that people really want - and need - to solve these problems in practice!