

---

## Preface

Recent technological advancements in the last few decades provided vast and unprecedented amounts of biological data including data of DNA and cell, and biological networks. This data comes in two basic formats as DNA nucleotide and protein amino acid sequences, and more recently, topological data of biological networks. Analysis of this huge data is a task on its own and the problems encountered are NP-hard most of the time, defying solutions in polynomial time. Such analysis is required as it provides a fundamental understanding of the functioning of a cell which can help understand human health and disease states and the diagnosis of diseases, which can further aid development of biotechnological processes to be used for medical purposes such as treatment of diseases.

Instead of searching for optimal solutions to these difficult problems, approximation algorithms that provide sub-optimal solutions are usually preferred. An approximation algorithm should guarantee a solution within an approximation factor for all input combinations. In many cases, even approximation algorithms are not known to date and using heuristics that are shown to work for most of the input cases experimentally are considered as solutions.

Under these circumstances, there is an increasing demand and interest in research community for parallel/distributed algorithms to solve these problems efficiently using a number of processing elements. This book is about both sequential and distributed algorithms for the analysis and modeling of biological data and as such, it is one of the first ones in this topic to the best of our knowledge. In the context of this book, we will assume a distributed algorithm as a parallel algorithm executed on a distributed memory processing system using *message-passing* rather than special purpose parallel architectures. For the cases of shared memory parallel computing, we will use the term *parallel algorithm* explicitly. We also cover algorithms for biological sequences (DNA and protein) and biological network (protein interaction networks, gene regulation, etc.) data in the same volume. Although algorithms for DNA sequences have a longer history of study, even the sequential algorithms for biological networks such as the protein interaction networks are rare and are at an early stage of development in research studies. We aim to give a unified view of algorithms for sequences and networks of biological systems where possible. These two views are not totally unrelated; for example, the function of a protein is influenced by both its position in a network

and its amino acid sequence, and also by its 3-D structure. It can also be seen that the problems in the sequence and network domains are analogous to some extent; for example, sequence alignment and network alignment, sequence motifs and network motifs, sequence clustering and network clustering are analogous problems in these two related worlds. It is not difficult to predict that the analysis of biological networks will have a profound effect on our understanding the origins of life, health and disease states, as analysis of DNA/RNA and protein sequences have provided.

The parallel and distributed algorithms are needed to solve bioinformatics problems simply for the speedup they provide. Even the linear time algorithms may be difficult to realize in bioinformatics due to the size of the data involved. For example, suffix trees are fundamental data structures in bioinformatics, and constructing them takes  $O(n)$  time by relatively new algorithms such as Ukkonen's or Farach's. Considering human DNA which consists of over 3 billion base pairs, even these linear time algorithms are time-consuming. However, by using distributed suffix trees, the time can be reduced to  $O(n/k)$  where  $k$  is the number of processors.

One wonders then about the scarcity of the research efforts in the design and implementation of distributed algorithms for these time-consuming difficult tasks. A possible reason would be that a number of problems have been introduced recently and the general approach in the research community has been to search for sequential algorithmic solutions first and then investigate ways of parallelizing these algorithms or design totally new parallel/distributed algorithms. Moreover, the parallel and distributed computing is a principle on its own where researchers in this field may not be familiar with bioinformatics problems in general, and the multidisciplinary efforts in this discipline and bioinformatics seem to be just starting. This book is an effort to contribute to the filling of the aforementioned gap between the distributed computing and bioinformatics. Our main goal is to first introduce the fundamental sequential algorithms to understand the problem and then describe distributed algorithms that can be used for fundamental bioinformatics problems such as sequence and network alignment, and finding sequence and network motifs, and clustering. We review the most fundamental sequential algorithms which aid the understanding of the problem better and yield parallel/distributed versions. In other words, we try to be as comprehensive as possible in the coverage of parallel/distributed algorithms for the fundamental bioinformatics problems with an in-depth analysis of sequential algorithms.

Writing a book on bioinformatics is a challenging task for a number of reasons. First of all, there are so many diverse topics to consider, from clustering to genome rearrangement, from network motif search to phylogeny, and one has to be selective not to divert greatly from the initially set objectives. We had to carefully choose a subset of topics to be included in this book in line with the book title and aim; tasks that require substantial computation power due to their data sizes and therefore are good candidates for parallelization. Second, bioinformatics is a very dynamic area of study with frequent new technological advances and results which requires a thorough survey of contemporary literature on presented topics. The two worlds of bioinformatics, namely biological sequences and biological networks, both have similar challenging problems. A closer look reveals these two worlds in fact have

analogous problems as mentioned; sequence alignment and network alignment; sequence motif search and network motif search; sequence clustering and network clustering which may be comforting first. However, these problems are not closely related in general, other than the clustering problem which can be converted from the sequence domain to the network domain with some effort.

We have a uniform chapter layout by first starting with an informal description of the problem at hand. We then define it formally and review the significant sequential algorithms in this topic briefly. We describe parallel/distributed algorithms for the same problem, which is the main theme of the book, and briefly discuss software packages if there is any available. When distributed algorithms for the topic are scarce, we provide clues and possible approaches for distributed algorithms as possible extensions to sequential ones or as totally new algorithms, to aid starting researchers in the topic. There are several coarsely designed and unpublished distributed algorithms, approximately 2-3 algorithms in some chapter, for the stated problems in biological sequences and networks which can be taken as starting points for potential researchers in this field. In the final part of each chapter, we emphasize main points, compare the described algorithms, give a contemporary review of the related literature, and show possible open research areas in the chapter notes section.

The intended audience for this book is the graduate students and researchers of computer science, biology, genetics, mathematics, and engineering, or any person with basic background in discrete mathematics and algorithms. The Web page for the book is at: <http://eng.izmir.edu.tr/~kerciyes/DSAB>.

I would like to thank graduate students at Izmir University who have taken complex networks and distributed algorithms courses for their valuable feedback when parts of the material covered in the book were presented during lectures. I would like to thank Esra Rüzgar for her review and comments on parallel network motif search algorithms of Chap. 14. I would also like to thank Springer senior editor Wayne Wheeler and associate editor Simon Rees for their continuous support and patience during the course of this project.

Uckuyular, Izmir, Turkey

K. Erciyes

Distributed and Sequential Algorithms for  
Bioinformatics

Erciyes, K.

2015, XVII, 367 p. 157 illus. in color., Hardcover

ISBN: 978-3-319-24964-3