

Preface to the Second Edition

The first edition of *Comparative Gene Finding: Models, Algorithms and Implementation* was published in March 2010. Since then a lot has happened and the field is gradually changing. A main driving force has been the ever-increasing use of next-generation sequencing (NGS) technology, which is revolutionizing a manifold of related fields. The pressure on computational methods and tools to handle these large amounts of data is greater than ever. In particular, since the “old” sequence analysis tools are not well adapted to the new situation with a huge data volume, much shorter read lengths, and an increased level of sequencing errors.

The gene prediction process these days typically involve the running of a multitude of bioinformatics tools, for repeat masking, for gene prediction, and for homology analyses of data from a variety of sources. The analysis tools are preferably gathered in an annotation pipeline that automatizes the processes and produces a consensus annotation by means of some kind of *combiner* software. Therefore, in this second edition we have chosen to add a chapter on annotation pipelines for next-generation sequencing data. The chapter gives a brief description of DNA sequencing in general, and of NGS techniques in particular, as well as a few application areas relevant to the gene prediction problem. The various issues involved in building a pipeline, is presented, with a discussion of the main steps including sequence assembly, *de novo* repeat masking, gene prediction, and genome annotation. Furthermore, Chap. 2 is extended to include a section on conditional random fields (CRF) as yet another model for computational gene finding. CRFs make a valuable contribution in the new sequencing technology era, as they allow for a more flexible inclusion of differing input formats and complex interdependencies between data.

Besides this and a few minor corrections, the second edition is largely unaltered. The intended reader and the required prerequisites stated in the former preface therefore remain unchanged.

Gothenburg, February 2015

Marina Axelsson-Fisk

Preface to the First Edition

Comparative genomics is a new and emerging field, and with the explosion of available biological sequences the requests for faster, more efficient, and more robust algorithms to analyze all this data are immense. This book is meant to serve as a self-contained instruction of the state of the art of computational gene finding in general, and of comparative approaches in particular. It is meant as an overview of the various methods that have been applied in the field, and a quick introduction into how computational gene finders are built in general. A beginner to the field could use this book as a guide through to the main points to think about when constructing a gene finder, and the main algorithms that are in use. On the other hand, the more experienced gene finder should be able to use this book as a reference to the different methods and to the main components incorporated in these methods. I have focused on the main uses of the covered methods and avoided much of the technical details and general extensions of the models. In exchange I have tried to supply references to more detailed accounts of the different research areas touched upon.

The book makes no claim of being comprehensive, however. As the amount of available data has exploded, as has the literature around computational biology and comparative genomics over the past few years, and although I have attempted to leave no threads untouched, it has been impossible to include all different approaches and aspects of the field. Moreover, I am likely to have missed several important references that rightfully should have been mentioned in this text. To all of you I sincerely apologize.

The structure of the book is meant to follow the natural order in which a gene finding software is built, starting with the main models and algorithms, and then breaking them down into the intrinsic submodels that cover the various features of a gene. The book is initiated in Chap. 1 with a brief encounter of genetics, describing the various biological terms and concepts that will be used in the succeeding chapters. Here we discuss the general terms of gene structure, and discuss the problems of settling on a gene definition, before we describe the gene finding problem that we have set out to solve. The end of the chapter includes a historical overview of the algorithm development of the past few decades. Chapter 2 covers

some of the algorithms most commonly used for single species gene finding. Each model section includes a theoretical encounter and illustrative examples, and is concluded with a description of an existing gene finding software that uses the model. In Chap. 3 we move on to sequence alignments. The chapter is divided into two parts. The first part describes different scoring schemes used in pairwise alignments, the application of dynamic programming, and the basic properties and statistical foundation of heuristic database searches. The second part describes the most common approaches to multiple sequence alignment, and the various attempts to deal with the increased computational complexity. In Chap. 4 we take on the main topic of the book, comparative gene finding. Here we combine the ideas in Chaps. 2 and 3 to a comparative setting, and describe how the strengths of both areas can be combined to improve the accuracy of gene finding. Again, each section is structured into a theoretical part, examples and an overview of the use of the model in an existing gene finder. Chapter 5 takes us through the gene features most commonly captured by a computational gene model, and describes the most important submodels used. A variety of different algorithms are described in detail, along with several illustrations and examples. Chapter 6 goes through the basics of parameter training, and covers a number of the different parameter estimation and optimization techniques commonly used in gene finding. In Chap. 7 we illustrate how to implement a comparative gene finder by giving the details behind the cross-species gene finder SLAM. SLAM uses a generalized hidden Markov model as main algorithm and has been used both by the Mouse Genome Sequencing Consortium to compare the initial sequence of mouse to the human genome, and by the Rat Genome Sequencing Consortium to perform a three-way analysis of human, mouse, and rat. The different steps and aspects in constructing a comparative gene finder are explained, and concluded with an encounter of various accuracy assessment measures used to debug and benchmark the resulting software.

This book covers a number of different fields, including probability theory, statistics, information theory, optimization theory, and numerical analysis. The reader is expected to have some background in bioinformatics in general, and in mathematics and mathematical statistics in particular. Basic knowledge of analysis, probability theory, and random processes will prove very valuable. The level and the structure of the book is such that it can readily be used as a course book for master level students, but it can also provide valuable insights and give a good overview to scientists wanting to get into the field quickly. Besides being specifically focused on the algorithmic details surrounding computational gene finding, it provides a good lesson on the intrinsic parts of computational biology and biological sequence analysis, as well as in giving an overview of a number of important mathematical and statistical areas applied in bioinformatics. A master-level course could very well be structured simply by following the book chapter-by-chapter, and perhaps include a smaller implementation project at the end.

Gothenburg, November 2009

Marina Axelsson-Fisk