
Preface

The existence of noncoding RNA genes (ncRNAs) was proposed simultaneously with protein coding genes in 1961 by Jacob and Monod. Since then, a substantial focus has been on protein coding genes, while the area of ncRNA evolved more slowly and received less attention even though major breakthroughs were made, such as the discovery of RNA's ability to carry out catalytic function, which also gave rise to the hypothesis that life originated from an RNA world, given that RNA also can store genetic information. It was also revealed early on that RNA might often be more conserved in structure rather than sequence. It quickly became apparent that adding consideration of structure to RNA sequence analysis programs was far more computationally demanding than, for example, comparing DNA by the primary sequence as in the case of sequence alignment methods. Whereas a pairwise sequence alignment scales with the square of the length being aligned, folding a single sequence scales with the cube of the length. Thus, doubling the length makes the raw version of the alignment methods run four times longer, but RNA folding algorithms run eight times longer. Thus, a likely contribution to the relative neglect of RNA bioinformatics in the early days can probably be attributed to the fact that this is a harder problem than many other bioinformatics problems. Today, much faster computers plus meaningful heuristics have made it possible to engineer practical RNA bioinformatics tools. Though RNA bioinformatics is still in its early phase with respect to practicality of genome-scale analyses, computational tools might help in uncovering the extent of RNAs in genomes. Given that, for example, protein coding sequence makes up about 1.2% of the human genome and that most of the genome is transcribed, this leaves an enormous potential for noncoding transcripts that might carry out a function and thus qualify as an ncRNA.

The ncRNAs have now been recognized as an abundant class of genes which often function through their structure. Protein coding genes have also been recognized to contain RNA structural motifs or RNA structures involved in, for example, regulation. Since even before the word "bioinformatics" was coined, researchers have been developing tools and computational methodologies for the analysis of RNA sequences, for aiding RNA (secondary) structure determination, for functional studies, and for a range of subsequent disciplines rooted in the principles for RNA structure prediction. Recognizing that RNA structure is a characteristic feature of ncRNAs, these tools have enabled genome-scale, *in silico* screens for ncRNAs. Furthermore, the same basic principles underlying RNA folding algorithms have been extended to a range of related problems such as homology search, design of interfering RNAs, and prediction of RNA–RNA interactions, to mention some examples. This book addresses a range of these methodologies from both a practical point of view as well from a computational and algorithmic perspective. Traditionally, the computational methods were referred to as computational RNA biology. However, with the recent applications on genomic and transcriptomic data, the more applied side of computational RNA biology, focused on processing experimental data (especially high-throughput data) is more commonly covered by the term Bioinformatics. This book covers a substantial and relevant fraction of both these directions and addresses both the biologist

interested in knowing more about RNA bioinformatics as well as the bioinformaticist interested in aspects of the “engine room.”

Recent technological development pushes high-throughput data generation and motivates further improvement of the generally computational resource demanding programs in RNA bioinformatics, a cost “inherited” from the generic RNA folding algorithms. Whereas these issues are addressed and the concepts of many methods shown, it is beyond this book to enter the area of assembly and read mapping. Here, we walk through the key methods and principles of RNA bioinformatics. Whereas a substantial part of the methodologies originate in the principles employed for prediction of RNA secondary structure, they employ further layers for specific applications as well as restrictions to reduce computation time and memory requirements, for example. In particular, developments in this respect have pushed for making methods in computational RNA biology applicable within RNA bioinformatics. Here, we range from the methodologies to their actual applications.

The content of this book is organized as follows. Initially an introduction to RNA bioinformatics is given (Chapter 1). This is followed by a description of RNA 3D structure (Chapter 2) and the origin of RNA folding parameters (Chapter 3) constitutes a background. Chapters 4 and 5 describe folding of single sequences by energy and by probabilistic modeling using stochastic context-free grammars. Chapter 6 describes RNA databases based on structural alignments of RNAs. Following this, folding of multiple aligned sequences by two strategies (Chapters 7 and 8) show that foldings can be made more reliably in such instances. This is followed by approaches describing genomic annotation of structured RNAs by homology search (Chapter 9) and the search for class-specific ncRNAs (Chapter 10). Chapter 11 describes how to extract (RNA 2D) motifs from RNA structures and Chapter 12 introduces the concept of comparing RNA secondary structures. Chapters 13–15 introduce structural alignments ranging from the so-called Sankoff-based approaches to alternatives and finally to exploitation of this to search for RNA structures in genomic sequence with low signal of conservation at the sequence level. In Chapter 16 an in-depth introduction to the evolution of RNA structure is given. The following Chapter 17 introduces RNA editors for careful curation of RNA structural alignments. In Chapter 18, RNA 3D modeling is introduced. Strategies and principles for RNA–RNA interactions are introduced in Chapter 19 and the special case of microRNA target prediction in Chapter 21. Before that, in Chapter 20, microRNA gene finding is presented. In Chapter 22, design of siRNAs are introduced and the final Chapter 23 provides an overview for RNA–protein interactions.

Finally, we would like to acknowledge those who helped make this book possible: all of the chapter authors for their hard work and thoughtful contributions, the Series Editor, John Walker, for his continued encouragement, and our families for their constant support.

Frederiksberg, Denmark
Seattle, WA, USA

Jan Gorodkin
Walter L. Ruzzo

RNA Sequence, Structure, and Function: Computational
and Bioinformatic Methods

Gorodkin, J.; Ruzzo, W.L. (Eds.)

2014, XI, 533 p. 120 illus., 46 illus. in color., Hardcover

ISBN: 978-1-62703-708-2

A product of Humana Press