

Preface

Efficient computer programs have made it possible to elucidate and analyze large-scale genomic sequences. Fundamental tasks, such as the assembly of numerous whole-genome shotgun fragments, the alignment of complementary DNA sequences with a long genome, and the design of gene-specific primers or oligomers, require efficient algorithms and state-of-the-art implementation techniques. This textbook emphasizes basic software implementation techniques for processing large-scale genome sequences.

In Chapters 1–6 we introduce the basic data structures and algorithms used for string matching and sequence alignment, describe performance-acceleration methods such as lookup tables and suffix arrays, and provide executable sample programs to motivate readers to try and enjoy programming. In Chapters 7 and 8 we show how these fundamental techniques are combined to solve major modern problems: whole-genome shotgun assembly, sequence alignment with a genome, comparative genomics, and the design of gene-specific sequence tags. These bioinformatics topics involve challenging, exciting, and real problems. Due to the space limitation, we cannot afford to include large executable programs in these two chapters.

This book is designed for use as a textbook during a semester course for advanced undergraduates and graduate students in computer science or biology. It is also suitable for computer scientists and programmers interested in genome sequence processing. To make the book accessible for nonprogrammers, there are many figures and examples to illustrate data structures and the behavior of algorithms.

We would like to thank Tomoyuki Yamada for studying seeded alignments with the second author, and Shin Sasaki for working with the first author to develop the whole genome shotgun assembler Ramen. This book owes a great deal to Tomoyuki and Shin.