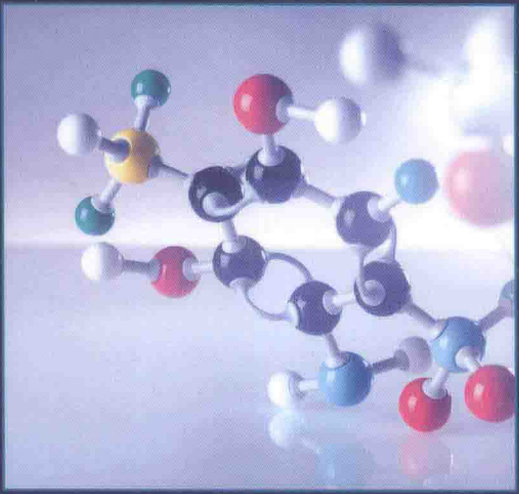
Wiley Series on Bioinformatics: Computational Techniques and Engineering

Yi Pan and Albert Y. Zomaya, Series Editors

Computational Methods for Next Generation Sequencing Data Analysis



EDITED BY

Ion I. Măndoiu • Alexander Zelikovsky

with website

WILEY

COMPUTATIONAL METHODS FOR NEXT GENERATION SEQUENCING DATA ANALYSIS

Edited by

ION I. MĂNDOIU
ALEXANDER ZELIKOVSKY

WILEY

Copyright © 2016 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Names: Măndoiu, I. Ion, editor of compilation. | Zelikovsky, Alexander, editor of compilation.
Title: Computational methods for next generation sequencing data analysis / edited by Ion I. Măndoiu, Alexander Zelikovsky.
Description: Hoboken, New Jersey : John Wiley & Sons, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2016010861 (print) | LCCN 2016014704 (ebook) | ISBN 9781118169483 (cloth) | ISBN 9781119272168 (pdf) | ISBN 9781119272175 (epub)
Subjects: LCSH: Nucleotide sequence–Methodology. | Nucleotide sequence–Data processing.
Classification: LCC QP620 .C648 2016 (print) | LCC QP620 (ebook) | DDC 611/.0181663–dc23
LC record available at http://lccn.loc.gov/2016010861

Cover image courtesy of Gettyimages/Andrew Brookes

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTORS

Vanessa Aguiar-Pulido, Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, FL, USA

Sahar Al Seesi, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

Alexander Artyomenko, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Niko Beerenwinkel, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

Adrian Caciula, Department of Computer Science, Georgia State University, Atlanta, GA, USA

David S. Campo, Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, GA, USA

Michael Campos, Miller School of Medicine, University of Miami, Miami, FL, USA

Stefan Canzar, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, and Toyota Technological Institute at Chicago, Chicago, IL, USA

Jeong-Hyeon Choi, Cancer Center, Medical College of Georgia, Georgia Regents University, Augusta, GA, USA; Department of Biostatistics and Epidemiology, Medical College of Georgia, Georgia Regents University, Augusta, GA, USA

Chong Chu, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

Zoya Dimitrova, Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, GA, USA

Jorge Duitama, Agrobiodiversity Research Area, International Center for Tropical Agriculture (CIAT), Cali, Colombia

Eleazar Eskin, Department of Computer Science, University of California, Los Angeles, CA, USA

Mitch Fernandez, Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, FL, USA

Liliana Florea, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Olga Glebova, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Xuan Guo, Department of Computer Science, Department of Biology, Georgia State University, Atlanta, GA, USA

Steven J. Hallam, Graduate Program in Bioinformatics and Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada

Niels W. Hanson, Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada

Elena Harris, Department of Computer Science, California State University, Chico, CA

Wenrui Huang, Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, FL, USA

Mazhar I. Khan, Department of Pathobiology and Veterinary Science, University of Connecticut, Storrs, CT, USA

Yury Khudyakov, Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, GA, USA

Kishori M. Konwar, Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada

Bing Li, Department of Computer Science, Department of Biology, Georgia State University, Atlanta, GA, USA

James Lindsay, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

Rasiah Loganantharaj, Bioinformatics Research Lab, The Center for Advanced Computer Studies, University of Louisiana, Lafayette, LA, USA

Stefano Lonardi, Department of Computer Science and Engineering, University of California, Riverside, CA, USA

此为试读,需要完整PDF请访问: www.ertongbook.com

Nicholas Mancuso, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Ion I. Măndoiu, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

Igor Mandric, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Serghei Mangul, Department of Computer Science, University of California, Los Angeles, CA, USA

Tobias Marschall, Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Kalai Mathee, Herbert Wertheim College of Medicine, Florida International University, Miami, FL, USA

Giri Narasimhan, Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, FL, USA

Ekaterina Nenastyeva, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Rachel O'neill, Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

Yi Pan, Department of Computer Science, Department of Biology, Georgia State University, Atlanta, GA, USA

Sumathi Ramachandran, Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, GA, USA

Thomas A. Randall, Integrative Bioinformatics, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Juan Riveros, Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, FL, USA

Alexander Schönhuth, Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Jonathan Segal, Herbert Wertheim College of Medicine, Florida International University, Miami, FL, USA

Huidong Shi, Cancer Center, Medical College of Georgia, Georgia Regents University, Augusta, GA, USA Department of Biochemistry, Medical College of Georgia, Georgia Regents University, Augusta, GA, USA

Pavel Skums, Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, GA, USA

Ren Sun, Department of Molecular and Medical Pharmacology, University of California, Los Angeles, CA, USA

Sing-hoi Sze, Department of Computer Science and Engineering and Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, USA

Yvette Temate-tiagueu, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Armin Töpfer, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

Bassam Tork, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Nicholas C. Wu, Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA

Shang-ju Wu, Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

Yufeng Wu, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

Ning Yu, Department of Computer Science, Department of Biology, Georgia State University, Atlanta, GA, USA

Alexander Zelikovsky, Department of Computer Science, Georgia State University, Atlanta, GA, USA

Erliang Zeng, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA

Jin Zhang, McDonnell Genome Institute, Washington University in St. Luis, MO, USA

PREFACE

Massively parallel DNA sequencing and RNA sequencing have become widely available, reducing the cost by several orders of magnitude and placing the capacity to generate gigabases to terabases of sequence data into the hands of individual investigators. These so-called *next-generation sequencing (NGS)* technologies have dramatically accelerated biological and biomedical research by enabling the comprehensive analysis of genomes and transcriptomes to become inexpensive, routine, and widespread. The ensuing explosion in the volume of data has spurred numerous advances in computational methods for NGS data analysis.

This book aims to provide an in-depth survey of some of the most important recent developments in this area. It is neither intended as an introductory text nor as a comprehensive review of existing bioinformatics tools and active research areas in NGS data analysis. Rather, our intention is to make a carefully selected set of advanced computational techniques accessible to a broad readership, including graduate students in bioinformatics and related areas and biomedical professionals who want to expand their repertoire of computational techniques for NGS data analysis. We hope that our emphasis on in-depth presentation of both algorithms and software for computational data analysis of current high-throughput sequencing technologies will best prepare the readers for developing their own algorithmic techniques and for successfully implementing them in existing and novel NGS applications.

The book features 18 chapters authored by bioinformatics experts who are active contributors to the respective subjects. The chapters are intended to be largely independent, so that readers do not have to read every chapter nor have to read them in a particular order. The chapters are grouped into the following four parts:

- Part I focuses on computing and experimental infrastructure for NGS data analysis, including chapters on cloud computing, a modular pipeline for metabolic pathway reconstruction, pooling strategies for massive viral sequencing, and high-fidelity sequencing protocols.

- Part II concentrates on analyses of DNA sequencing data and includes chapters on the classic scaffolding problem, detection of genomic variants, two chapters on finding insertions and deletions, and two chapters on the analysis of DNA methylation sequencing data.
- Part III is devoted to analyses of RNA-seq data. Two chapters describe algorithms and compare software tools for transcriptome assembly: one chapter focuses on methods for alternative splicing analysis and the other chapter focuses on tools for transcriptome quantification and differential expression analysis.
- Part IV explores computational tools for NGS applications in microbiomics. The first chapter concentrates on error correction of NGS reads from viral populations, then two chapters describe methods for viral quasispecies reconstruction, and the last chapter surveys the state of the art and future trends in microbiome analysis.

We are grateful to all the authors for their excellent contributions, without which this book would not have been possible. We hope that their deep insights and fresh enthusiasm will help in attracting new generations of researchers to this dynamic field. We would also like to thank Yi Pan and Albert Y. Zomaya for nurturing this project since its inception, and the editorial staff at Wiley Interscience for their patience and assistance throughout the project. Finally, we wish to thank our friends and families for their continuous support.

Ion I. Măndoiu
Storrs, Connecticut
Alexander Zelikovsky
Atlanta, Georgia

ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website:

www.wiley.com/go/Mandoiu/NextGenerationSequencing

The book companion website contains the color version of a few selected figures

Figure 2.3, Figure 2.5, Figure 2.6, Figure 2.13, Figure 3.1, Figure 3.9, Figure 7.5, Figure 8.3, Figure 8.4, Figure 9.4, Figure 9.8, Figure 9.9, Figure 9.12, Figure 9.14, Figure 12.3, Figure 12.4, Figure 12.5, Figure 15.3, Figure 16.1, Figure 16.6, Figure 16.7, Figure 16.11, Figure 16.12, Figure 16.13, Figure 18.1, Figure 18.2, Figure 18.3, Figure 18.4, Figure 18.5, Figure 18.7.

CONTENTS IN BRIEF

CONTRIBUTORS xix

PREFACE xxiii

ABOUT THE COMPANION WEBSITE xxv

PART I COMPUTING AND EXPERIMENTAL INFRASTRUCTURE FOR NGS 1

1 Cloud Computing for Next-Generation Sequencing Data Analysis 3
Xuan Guo, Ning Yu, Bing Li, and Yi Pan

2 Introduction to the Analysis of Environmental Sequence Information Using Metapathways 25
Niels W. Hanson, Kishori M. Konwar, Shang-Ju Wu, and Steven J. Hallam

3 Pooling Strategy for Massive Viral Sequencing 57
Pavel Skums, Alexander Artyomenko, Olga Glebova, Sumathi Ramachandran, David S. Campo, Zoya Dimitrova, Ion I. Măndoiu, Alexander Zelikovsky, and Yury Khudyakov

4 Applications of High-Fidelity Sequencing Protocol to RNA Viruses 85
Serghei Mangul, Nicholas C. Wu, Ekaterina Nenastyeva, Nicholas Mancuso, Alexander Zelikovsky, Ren Sun, and Eleazar Eskin

此为试读,需要完整PDF请访问: www.ertongbook.com

PART II GENOMICS AND EPIGENOMICS 105

5 Scaffolding Algorithms 107
Igor Mandric, James Lindsay, Ion I. Măndoiu, and Alexander Zelikovsky

6 Genomic Variants Detection and Genotyping 133
Jorge Duitama

7 Discovering and Genotyping Twilight Zone Deletions 149
Tobias Marschall and Alexander Schönhuth

8 Computational Approaches for Finding Long Insertions and Deletions with NGS Data 175
Jin Zhang, Chong Chu, and Yufeng Wu

9 Computational Approaches in Next-Generation Sequencing Data Analysis for Genome-Wide DNA Methylation Studies 197
Jeong-Hyeon Choi and Huidong Shi

10 Bisulfite-Conversion-Based Methods for DNA Methylation Sequencing Data Analysis 227
Elena Harris and Stefano Lonardi

PART III TRANSCRIPTOMICS 245

11 Computational Methods for Transcript Assembly from RNA-SEQ Reads 247
Stefan Canzar and Liliana Florea

12 An Overview And Comparison of Tools for RNA-Seq Assembly 269
Rasiah Loganantharaj and Thomas A. Randall

13 Computational Approaches for Studying Alternative Splicing in Nonmodel Organisms From RNA-SEQ Data 287
Sing-Hoi Sze

14 Transcriptome Quantification and Differential Expression From NGS Data 301
Olga Glebova, Yvette Temate-Tiagueu, Adrian Caciula, Sahar Al Seesi, Alexander Artyomenko, Serghei Mangul, James Lindsay, Ion I. Măndoiu, and Alexander Zelikovsky

PART IV MICROBIOMICS 329

15 **Error Correction of NGS Reads from Viral Populations** 331
Pavel Skums, Alexander Artyomenko, Olga Glebova, David S. Campo, Zoya Dimitrova, Alexander Zelikovsky, and Yury Khudyakov

16 **Probabilistic Viral Quasispecies Assembly** 355
Armin Töpfer and Niko Beerenwinkel

17 **Reconstruction of Infectious Bronchitis Virus Quasispecies from NGS Data** 383
Bassam Tork, Ekaterina Nenastyeva, Alexander Artyomenko, Nicholas Mancuso, Mazhar I. Khan, Rachel O'Neill, Ion I. Măndoiu, and Alexander Zelikovsky

18 **Microbiome Analysis: State of the Art and Future Trends** 401
Mitch Fernandez, Vanessa Aguiar-Pulido, Juan Riveros, Wenrui Huang, Jonathan Segal, Erliang Zeng, Michael Campos, Kalai Mathee, and Giri Narasimhan

INDEX 425

CONTENTS

CONTRIBUTORS xix

PREFACE xxiii

ABOUT THE COMPANION WEBSITE xxv

PART I COMPUTING AND EXPERIMENTAL INFRASTRUCTURE FOR NGS 1

1 Cloud Computing for Next-Generation Sequencing Data Analysis 3

Xuan Guo, Ning Yu, Bing Li, and Yi Pan

1.1 Introduction, 3
1.2 Challenges for NGS Data Analysis, 4
1.3 Background For Cloud Computing and its Programming Models, 6
1.3.1 Overview of Cloud Computing, 7
1.3.2 Cloud Service Providers, 7
1.3.3 Programming Models, 8
1.4 Cloud Computing Services for NGS Data Analysis, 13
1.4.1 Hardware as a Service (HaaS), 13
1.4.2 Platform as a Service (PaaS), 13
1.4.3 Software as a Service (SaaS), 15
1.4.4 Data as a Service (DaaS), 20
1.5 Conclusions and Future Directions, 20
References, 21

2 **Introduction to the Analysis of Environmental Sequence Information Using Metapathways** **25**

Niels W. Hanson, Kishori M. Konwar, Shang-Ju Wu, and Steven J. Hallam

2.1 Introduction & Overview, 25
2.2 Background, 26
2.3 Metapathways Processes, 27
2.3.1 Open Reading Frame (ORF) Prediction, 29
2.3.2 Functional Annotation, 32
2.3.3 Analysis Modules, 33
2.3.4 ePGDB Construction, 38
2.4 Big Data Processing, 39
2.4.1 A Master–Worker Model for Grid Distribution, 39
2.4.2 GUI and Data Integration, 41
2.5 Downstream Analyses, 41
2.5.1 Large Table Comparisons, 43
2.5.2 Pathway Tools Cellular Overview, 43
2.5.3 Statistical Analysis with R, 45
2.5.4 Venn Diagram, 47
2.5.5 Clustering and Relating Samples by Pathways, 49
2.5.6 Faceting Variables with ggplot2, 49
2.6 Conclusions, 50
References, 50

3 **Pooling Strategy for Massive Viral Sequencing** **57**

Pavel Skums, Alexander Artyomenko, Olga Glebova, Sumathi Ramachandran, David S. Campo, Zoya Dimitrova, Ion I. Măndoiu, Alexander Zelikovsky, and Yury Khudyakov

3.1 Introduction, 57
3.2 Design of Pools for Big Viral Data, 60
3.2.1 Pool Design Optimization Formulation, 62
3.2.2 Greedy Heuristic for VSPD Problem, 63
3.2.3 The Tabu Search Heuristic for the OCBG Problem, 65
3.3 Deconvolution of Viral Samples from Pools, 68
3.3.1 Deconvolution Using Generalized Intersections and Differences of Pools, 68
3.3.2 Maximum Likelihood *k*-Clustering, 70
3.4 Performance of Pooling Methods on Simulated Data, 71
3.4.1 Performance of the Viral Sample Pool Design Algorithm, 71
3.4.2 Performance of the Pool Deconvolution Algorithm, 74
3.5 Experimental Validation of Pooling Strategy, 75
3.5.1 Experimental Pools and Sequencing, 75
3.5.2 Results, 77

3.6 Conclusion, 79
References, 81

4 **Applications of High-Fidelity Sequencing Protocol to RNA Viruses** 85
Serghei Mangul, Nicholas C. Wu, Ekaterina Nenastyeva, Nicholas Mancuso, Alexander Zelikovsky, Ren Sun, and Eleazar Eskin

4.1 Introduction, 85
4.2 High-Fidelity Sequencing Protocol, 86
4.3 Assembly of High-Fidelity Sequencing Data, 88
4.3.1 Consensus Construction, 88
4.3.2 Reads Mapping, 89
4.3.3 Viral Genome Assembler (VGA), 89
4.3.4 Viral Population Quantification, 91
4.4 Performance of VGA on Simulated Data, 93
4.5 Performance of Existing Viral Assemblers on Simulated Consensus Error-Corrected Reads, 98
4.6 Performance of VGA on Real Hiv Data, 99
4.6.1 Validation of *de novo* Consensus, 99
4.7 Comparison of Alignment on Error-Corrected Reads, 100
4.8 Evaluating of Error Correction Tools Based on High-Fidelity Sequencing Reads, 101
Acknowledgment, 101
References, 102

PART II GENOMICS AND EPIGENOMICS 105

5 **Scaffolding Algorithms** 107
Igor Mandric, James Lindsay, Ion I. Măndoiu, and Alexander Zelikovsky

5.1 Scaffolding, 107
5.2 State-of-The-Art Scaffolding Tools, 108
5.2.1 Sspace, 108
5.2.2 OPERA, 109
5.2.3 SOPRA, 110
5.2.4 MIP, 110
5.2.5 SCARPA, 111
5.3 Recent Scaffolding Tools, 111
5.3.1 SILP2, 111
5.3.2 ScaffMatch, 119
5.4 Scaffolding Software Evaluation, 124
5.4.1 Data Sets, 124
5.4.2 Quality Metrics, 124
5.4.3 Evaluation and Comparison, 126
References, 129

6 Genomic Variants Detection and Genotyping **133**

Jorge Duitama

6.1 Introduction, 133

6.2 Methods for Detection and Genotyping of SNPs and Small Indels, 135

6.2.1 Description of the Problem, 135

6.2.2 Bayesian Model, 136

6.2.3 Common Issues Affecting Genotype Quality, 137

6.2.4 Population Variability, 139

6.3 Methods for Detection and Genotyping of CNVs, 141

6.3.1 Mean-Shift Approach for CNVs within a Sample, 141

6.3.2 Identifying CNVs between Samples, 143

6.4 Putting Everything Together, 144

References, 145

7 Discovering and Genotyping Twilight Zone Deletions **149**

Tobias Marschall and Alexander Schönhuth

7.1 Introduction, 149

7.1.1 Twilight Zone Deletions, 151

7.2 Notation, 151

7.2.1 Alignments, 152

7.2.2 Gaps/Splits, 152

7.2.3 Deletions, 152

7.3 Non-Twilight-Zone Deletion Discovery, 152

7.3.1 Internal Segment Size-Based Approaches, 153

7.3.2 Split-Read Mapping Approaches, 154

7.3.3 Hybrid Approaches, 155

7.3.4 The "Twilight Zone": Definition, 156

7.4 Discovering "Twilight Zone" Deletions: New Solutions, 156

7.4.1 CLEVER, 156

7.4.2 Mate-Clever, 157

7.4.3 Pindel, 157

7.5 Genotyping "Twilight Zone" Deletions, 158

7.5.1 A Maximum Likelihood Approach under Read Alignment Uncertainty, 158

7.6 Results, 162

7.6.1 Data Set, 162

7.6.2 Tools, 163

7.6.3 Discovery, 163

7.6.4 Genotyping, 166

7.7 Discussion, 167

7.7.1 HiSeq, 169

7.7.2 MiSeq, 170

7.7.3 Conclusion, 170

此为试读，需要完整PDF请访问：www.ertongbook.com