

# GENOME SEQUENCING TECHNOLOGY AND ALGORITHMS

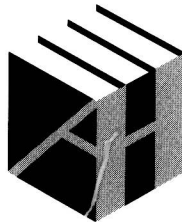
Sun Kim • Haixu Tang • Elaine R. Mardis  
EDITORS

Q78  
G335.2

# Genome Sequencing Technology and Algorithms

Sun Kim  
Haixu Tang  
Elaine R. Mardis

*Editors*



**ARTECH  
HOUSE**

BOSTON | LONDON  
arterhhouse.com

**Library of Congress Cataloging-in-Publication Data**

A catalog record for this book is available from the U.S. Library of Congress.

**British Library Cataloguing in Publication Data**

A catalogue record for this book is available from the British Library.

**Cover design by Igor Valdman**

ISBN 13: 978-1-59693-094-0

© 2008 ARTECH HOUSE, INC.  
685 Canton Street  
Norwood, MA 02062

All rights reserved. Printed and bound in the United States of America. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Artech House cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

10 9 8 7 6 5 4 3 2 1

# **Genome Sequencing Technology and Algorithms**

For a listing of related Artech House titles,  
turn to the back of this book.

## List of Contributors

- Chapter 1 Elaine R. Mardis,  
*Washington University*
- Chapter 2 Baback Gharizadeh, Roxana Jalili, and Mostafa Ronaghi,  
*Stanford University*
- Chapter 3 David Okou and Michael E. Zwick,  
*Emory University School of Medicine*
- Chapter 4 Jay Shendure,  
*University of Washington*  
Gregory J. Porreca and George M. Church,  
*Harvard Medical School*
- Chapter 5 Lewis J. Frey and Joyce A. Mitchell,  
*University of Utah*  
Victor Maojo,  
*Universidad Politecnica de Madrid*
- Chapter 6 Sun Kim and Haixu Tang,  
*Indiana University*
- Chapter 7 Sun Kim and Haixu Tang,  
*Indiana University*
- Chapter 8 Jiacheng Chen and Steven Skiena,  
*Stony Brook University*
- Chapter 9 Haixu Tang and Sun Kim,  
*Indiana University*
- Chapter 10 Paola Bonizzoni and Gianluca Della Vedova,  
*University di Milano-Bicocca*  
Riccardo Dondi,  
*University of Bergamo*  
Jing Li,  
*Case Western Reserve University*
- Chapter 11 Benjamin J. Raphael,  
*Brown University*  
Stas Volik, Colin C. Collins,  
*University of California at San Francisco*
- Chapter 12 Curt Balch and Kenneth P. Nephew,  
*Indiana University*  
Tim H.-M. Huang,  
*The Ohio State University*
- Chapter 13 Aleksandar Milosavljevic and Cristian Coarfa,  
*Baylor College of Medicine*

# Contents

	<b>Part I</b>	
	<b>The New DNA Sequencing Technology</b>	<b>1</b>
<b>1</b>	<b>An Overview of New DNA Sequencing Technology</b>	<b>3</b>
1.1	An Overview	3
1.1.1	Background	3
1.1.2	Rationale for Technology Development Toward Massively Parallel Scale DNA Sequencing	4
1.1.3	Goals of Massively Parallel Sequencing Approaches	6
1.2	Massively Parallel Sequencing by Synthesis Pyrosequencing	6
1.2.1	Principle of the Method	6
1.2.2	Pyrosequencing in a Microtiter Plate Format	7
1.2.3	The 454 GS-20 Sequencer	7
1.2.4	Novel Applications Enabled by Massively Parallel Pyrosequencing	8
1.3	Massively Parallel Sequencing by Other Approaches	8
1.3.1	Sequencing by Synthesis with Reversible Terminators	8
1.3.2	Ligation-Based Sequencing	8
1.3.3	Sequencing by Hybridization	9

1.4	Survey of Future Massively Parallel Sequencing Methods	9
1.4.1	Sequencing Within a Zero-Mode Waveguide	9
1.4.2	Nanopore Sequencing Approaches	10
	References	11
<b>2</b>	<b><u>Array-Based Pyrosequencing Technology</u></b>	<b>15</b>
2.1	Introduction	15
2.2	Pyrosequencing Chemistry	16
2.3	Array-Based Pyrosequencing	17
2.4	454 Sequencing Chemistry	18
2.5	Applications of 454 Sequencing Technology	19
2.5.1	Whole-Genome Sequencing	19
2.5.2	Ultrabroad Sequencing	20
2.5.3	Ultradeep Amplicon Sequencing	20
2.6	Advantages and Challenges	20
2.7	Future of Pyrosequencing	21
	References	21
<b>3</b>	<b><u>The Role of Resequencing Arrays in Revolutionizing DNA Sequencing</u></b>	<b>25</b>
3.1	Introduction	25
3.2	DNA Sequencing by Hybridization with Resequencing Arrays	26
3.3	Resequencing Array Experimental Protocols	28
3.4	Analyzing Resequencing Array Data with ABACUS	29
3.5	Review of RA Applications	33
3.5.1	Human Resequencing	33
3.5.2	Mitochondrial DNA Resequencing	33
3.5.3	Microbial Pathogen Resequencing	35
3.6	Further Challenges	38
	References	40



---

<b>4</b>	<b><u>Polony Sequencing</u></b>	<b>43</b>
4.1	Introduction	43
4.2	Overview	44
4.3	Construction of Sequencing Libraries	45
4.4	Template Amplification with Emulsion PCR	46
4.5	Sequencing	48
4.6	Future Directions	49
	References	50
<b>5</b>	<b><u>Genome Sequencing: A Complex Path to Personalized Medicine</u></b>	<b>53</b>
5.1	Introduction	53
5.2	Personalized Medicine	55
5.3	Heterogeneous Data Sources	56
5.4	Information Modeling	57
5.5	Ontologies and Terminologies	58
5.6	Applications	59
5.7	Conclusion	71
	References	72
	<b>Part II</b>	
	<b><u>Genome Sequencing and Fragment Assembly</u></b>	<b>77</b>
<b>6</b>	<b><u>Overview of Genome Assembly Techniques</u></b>	<b>79</b>
6.1	Genome Sequencing by Shotgun-Sequencing Strategy	79
6.1.1	A Procedure for Whole-Genome Shotgun (WGS) Sequencing	80
6.2	Trimming Vector and Low-Quality Sequences	82
6.2.1	The Trimming Vector and Low-Quality Sequences Problem	82
6.3	Fragment Assembly	84
6.3.1	The Fragment Assembly Problem	84

6.4	Assembly Validation	85
6.4.1	The Assembly Validation Problem	85
6.5	Scaffold Generation	87
6.5.1	The Scaffold Generation Problem	89
6.5.2	Bambus	89
6.5.3	GigAssembler	93
6.6	Finishing	94
6.7	Three Strategies for Whole-Genome Sequencing	94
6.8	Discussion	95
6.8.1	A Thought on an Exploratory Genome Sequencing Framework	96
	Acknowledgments	97
	References	97
<b>7</b>	<b>Fragment Assembly Algorithms</b>	<b>101</b>
7.1	TIGR Assembler	102
7.1.1	Merging Fragments with Assemblies	102
7.1.2	Building a Consensus Sequence	102
7.1.3	Handling Repetitive Sequences	103
7.2	Phrap	103
7.3	CAP3	104
7.3.1	Automatic Clipping of 5' and 3' Poor Quality Regions	104
7.3.2	Computation and Evaluation of Overlaps	104
7.3.3	Use of Mate-Pair Constraints in Construction of Contigs	105
7.4	Celera Assembler	105
7.4.1	Kececioglu and Myers Approach	105
7.4.2	The Design Principle of the Celera Whole-Genome Assembler	107
7.4.3	Overlapper	108
7.4.4	Unitigger	108
7.4.5	Scaffolder	109
7.5	Arachne	110
7.5.1	Contig Assembly	111

---

7.5.2	Detecting Repeat Contigs and Repeat Supercontigs	111
7.6	EULER	112
7.6.1	Idury-Waterman Algorithm	112
7.6.2	An Overview of EULER	113
7.6.3	Error Correction and Data Corruption	113
7.6.4	Eulerian Superpath	114
7.6.5	Use of Mate-Pair Information	115
7.7	Other Approaches to Fragment Assembly	116
7.7.1	A Genetic Algorithm Approach	116
7.7.2	A Structured Pattern-Matching Approach	117
7.8	Incompleteness of the Survey	119
	Acknowledgments	120
	References	120
<b>8</b>	<b>Assembly for Double-Ended Short-Read Sequencing Technologies</b>	<b>123</b>
8.1	Introduction	123
8.2	Short-Read Sequencing Technologies	125
8.3	Assembly for Short-Read Sequencing	128
8.3.1	Algorithmic Methods	129
8.3.2	Simulation Results	129
8.4	Developing a Short-Read-Pair Assembler	132
8.4.1	Analysis	135
	References	140
	<b>Part III</b>	
	<b>Beyond Conventional Genome Sequencing</b>	<b>143</b>
<b>9</b>	<b>Genome Characterization in the Post-Human Genome Project Era</b>	<b>145</b>
9.1	Genome Resequencing and Comparative Assembly	146
9.2	Genotyping Versus Haplotyping	147
9.3	Large-Scale Genome Variations	147

9.4	Epigenomics: Genetic Variations Beyond Genome Sequences	148
9.5	Conclusion	149
	References	149
<b>10</b>	<b>The Haplotyping Problem: An Overview of Computational Models and Solutions</b>	<b>151</b>
10.1	Introduction	151
10.2	Preliminary Definitions	153
10.3	Inferring Haplotypes in a Population	154
10.3.1	The Inference Problem: A General Rule	156
10.3.2	The Pure Parsimony Haplotyping Problem	158
10.3.3	The Inference Problem by the Coalescent Model	158
10.3.4	Xor-Genotyping	161
10.3.5	Incomplete Data	162
10.4	Inferring Haplotypes in Pedigrees	163
10.5	Inferring Haplotypes from Fragments	169
10.6	A Glimpse over Statistical Methods	175
10.7	Discussion	177
	Acknowledgments	178
	References	178
<b>11</b>	<b>Analysis of Genomic Alterations in Cancer</b>	<b>183</b>
11.1	Introduction	183
11.1.1	Measurement of Copy Number Changes by Array Hybridization	185
11.1.2	Measurement of Genome Rearrangements by End Sequence Profiling	187
11.2	Analysis of ESP Data	188
11.3	Combination of Techniques	191
11.4	Future Directions	191
	References	192

<b>12</b>	<b>High-Throughput Assessments of Epigenomics in Human Disease</b>	<b>197</b>
12.1	Introduction	197
12.2	Epigenetic Phenomena That Regulate Gene Expression	198
12.2.1	Methylation of Deoxycytosine	198
12.2.2	Histone Modifications and Nucleosome Remodeling	198
12.2.3	Small Inhibitory RNA Molecules	199
12.3	Epigenetics and Disease	200
12.3.1	Epigenetics and Developmental and Neurological Diseases	200
12.3.2	Epigenetics and Cancer	200
12.4	High-Throughput Analyses of Epigenetic Phenomena	201
12.4.1	Gel-Based Approaches	201
12.4.2	Microarrays	212
12.4.3	Cloning/Sequencing	213
12.4.4	Mass Spectrometry	215
12.5	Conclusions	215
	Acknowledgments	215
	References	216
<b>13</b>	<b>Comparative Sequencing, Assembly, and Anchoring</b>	<b>225</b>
13.1	Comparing an Assembled Genome with Another Assembled Genome	226
13.2	Mutual Comparison of Genome Fragments	229
13.3	Comparing an Assembled Genome with Genome Fragments	230
13.3.1	Applications Using Read Anchoring	230
13.3.2	Applications Employing Anchoring of Paired Ends	232
13.3.3	Applications Utilizing Mapping of Clone Reads	233
13.4	Anchoring by Seed-and-Extend Versus Positional Hashing Methods	234
13.5	The UD-CSD Benchmark for Anchoring	237

13.6	Conclusions	239
	References	241
	<b>About the Authors</b>	<b>245</b>
	<b>Index</b>	<b>251</b>

# **Part I**

## **The New DNA Sequencing Technology**





# 1

## An Overview of New DNA Sequencing Technology

Elaine R. Mardis

### 1.1 An Overview

#### 1.1.1 Background

The dideoxynucleotide termination DNA sequencing technology invented by Fred Sanger and colleagues, published in 1977, formed the basis for DNA sequencing from its inception through 2004 [1]. Originally based on radioactive labeling, the method was automated by the use of fluorescent labeling coupled with excitation and detection on dedicated instruments, with fragment separation by slab gel [2] and ultimately by capillary gel electrophoresis. A variety of molecular biology, chemistry, and enzymology-based improvements have brought Sanger's approach to its current state of the art. By virtue of economies of scale, high-throughput automation and reaction optimization, large sequencing centers have decreased the cost of a fluorescent Sanger sequencing reaction to around \$0.30. However, it is likely that only incremental cost decreases will continue to be achieved for Sanger sequencing in its current manifestation. This fact, coupled with the ever-increasing need for DNA sequencing toward a variety of biomedical (and other) studies, has resulted in a rapid phase of technology development of so-called next generation or massively parallel sequencing technologies, that will revolutionize DNA sequencing as we now know it. Along with this revolution will come a significant and potentially unanticipated impact