Contents

1. About this book . . . **1**
Why should biologists learn about assembly algorithms? . . . 1
Description of this book . . . 2
Regarding the exercises . . . 4
Other online resources for learning . . . 4

2. The Genome Assembly Problem . . . **6**
Shotgun sequencing and assembly . . . 6
Is the problem at all solvable? What are the fundamental limits? . . . 7
Assembly approaches - greedy, overlap-layout-consensus and de Bruijn graph-based algorithms . . . 8
Why use de Bruijn graphs for genome assembly? . . . 10
Summary of what we learned so far . . . 11

3. De Bruijn Graph of a Genome . . . **13**
De Bruijn graph of a small sequence . . . 13
Generalized definition of de Bruijn graph . . . 15
Double-Stranded nature of the genome . . . 16
Impact of changing k-mer size . . . 17
De Bruijn graph for repetitive regions . . . 19
Summary of what we learned so far . . . 19

4. Genome Assembly Procedure Using de Bruijn Graphs . . . **21**
Assembling genomes from perfect reads . . . 21
Read coherence is lost in de Bruijn graphs . . . 23
Varying k-mer sizes to compensate for loss of read coherence . . . 24
Memory requirement and k-mer distribution of perfect library . . . 26
Summary of what we learned so far . . . 27

5. Experimental Considerations . . . **28**
Evolution of sequencing technologies . . . 28
Paired-end and mate-pair reads . . . 30
Data formats . . . 33
Sources of artifacts . . . 33

Detailed description of ABI SOLiD color space data . . . 33
Summary of what we learned so far . . . 37

6. Genome Assembly from Noisy Reads . . . 38
Errors lead to high RAM usage for de Bruijn graph-based assemblers . . . 38
Impact of sequencing errors on the structure of de Bruijn graphs . . . 41
Impact of non-uniform sequencing . . . 45
Assembly steps implemented in typical genome assembly programs . . . 47
Full conceptual picture of de Bruijn graph for noisy reads . . . 48

7. When the Bugs Become Features - Assembling Transcriptomes, Metagenomes and Highly Polymorphic Genomes . . . 52
Impact of haplotype differences on de Bruijn Graphs . . . 52
Using De Bruijn graphs to assembles heterozygous genomes . . . 54
Using de Bruijn graphs to assemble metagenomes . . . 55
Using de Bruijn graphs to assemble transcriptomes (RNAseq) . . . 57
Summary of what we learned so far . . . 63

8. Faster, Better and Cheaper . . . 64
Computer science concepts for advanced work . . . 65
Data structures for efficient storage and processing of de-Bruijn graphs . . . 67
Algorithms for counting kmers efficiently . . . 68
Measures of assembly quality . . . 72
Analysis based on K-mer distribution . . . 72
Improving assembly by changing k-mer size for dBG . . . 73
Scaffolding and Repeat Resolution . . . 74

9. In Depth Discussion of Three de Bruijn Graph-based Assemblers . . . 76
Where to Get Them . . . 76
SPAdes - Saint Petersburg Genome Assembler . . . 76
Genome Assembler - SOAPdenovo . . . 79
Transcriptome Assembler - Trinity . . . 89
Summary of what we learned so far . . . 93

10. References . . . 94
Historical perspectives on genome assembly . . . 94
1. Pre-NGS Genome Assemblers . . . 94
2. NGS Genome Assemblers . . . 95
3. Exomes, Transcriptomes, Metagenomes and Highly Polymorphic Genomes . . . 97
4. Faster, better, cheaper . . . 98
6. Reviews and forecasts . . . 100

11. Solutions of Exercises . . . 101
"The Genome Assembly Problem" . . . 101

"De Bruijn Graph of a Genome" 101
"Genome Assembly Procedure Using de Bruijn Graphs" 102
"Experimental Considerations " 102
"Genome Assembly of Erroroneous Reads " 102
"When Bugs Are Features : Assembling Transcriptomes, Metagenomes and Highly Polymorphic Genomes" 102
"Faster. Better. Cheaper." 102
"In Depth Discussion of Three de Bruijn Graph-based Assemblers" 102