

Contents

Preface	xi
Author Biographies	xvii
Chapter 1. Methodological Concepts: Algorithmic Solutions of Bioinformatics Problems	1
Annie CHATEAU and Tom DAVOT-GRANGÉ	
1.1. Data, models, problem formalism in bioinformatics	1
1.1.1. Data	1
1.1.2. Genome modeling	4
1.1.3. Problems in bioinformatics	5
1.2. Mathematical preliminaries	6
1.2.1. Propositional logic preliminaries	6
1.2.2. Preliminaries on sets	7
1.3. Vocabulary in text algorithmics	9
1.4. Graph theory	10
1.4.1. Subgraphs	12
1.4.2. Path in a graph	13
1.4.3. Matching	13
1.4.4. Planarity	14
1.4.5. Tree decomposition	15
1.5. Algorithmic problems	16
1.5.1. Definition	16

1.5.2. Graph problem	17
1.5.3. Satisfiability problems	19
1.6. Problem solutions	20
1.6.1. Algorithm	20
1.6.2. Complexity	21
1.6.3. Runtime	24
1.7. Complexity classes	26
1.7.1. Generality	26
1.7.2. Exact algorithms	28
1.7.3. Approximation algorithms	32
1.7.4. Solvers	34
1.8. Some algorithmic techniques	35
1.8.1. Dynamic programming	35
1.8.2. Tree traversal	38
1.9. Validation	41
1.9.1. The different types of errors	42
1.9.2. Quality measures	44
1.9.3. And in the non-binary case?	46
1.10. Conclusion	47
1.11. References	47
Chapter 2. Sequence Indexing	49
Thierry LECROQ and Mikael SALSON	
2.1. Introduction	49
2.1.1. What is indexing?	50
2.1.2. When to index?	51
2.1.3. What to index?	51
2.1.4. Indexing structures and queries considered	52
2.1.5. Basic notions and vocabulary	53
2.2. Word indexing	54
2.2.1. Bloom filters	54
2.2.2. Inverted list	56
2.2.3. De Bruijn graphs	60
2.2.4. Efficient structures for targeted queries	61
2.3. Full-text indexing	62
2.3.1. Suffix tree	62
2.3.2. (Extended) suffix array	64
2.3.3. Burrows–Wheeler transform	67
2.4. Indexing choice criteria	76
2.4.1. Based on the type of the necessary query	77
2.4.2. Based on the space-time and data quantity trade-off	77

2.4.3. Based on the need to add or modify indexed data	79
2.4.4. Indexing choices according to applications	80
2.5. Conclusion and perspectives	81
2.5.1. Efficient methods for indexing a few genomes or sequencing sets	81
2.5.2. Methods that struggle to take advantage of data redundancy	82
2.6. References	83
Chapter 3. Sequence Alignment	87
Laurent NOÉ	
3.1. Introduction	87
3.1.1. What is pairwise alignment?	87
3.1.2. How to evaluate an alignment?	88
3.2. Exact alignment	90
3.2.1. Representation in edit graph form	90
3.2.2. Global alignment and Needleman–Wunsch algorithm	93
3.2.3. Local alignment and Smith–Waterman algorithm	94
3.2.4. Alignment with affine indel function and the Gotoh algorithm	96
3.3. Heuristic alignment	98
3.3.1. Seeds	99
3.3.2. <i>Min-hash</i> and global sampling	105
3.3.3. <i>Minimizing</i> and local sampling	106
3.4. References	109
Chapter 4. Genome Assembly	113
Dominique LAVENIER	
4.1. Introduction	113
4.2. Sequencing technologies	116
4.2.1. Short reads	117
4.2.2. Long reads	118
4.2.3. Linked reads	118
4.2.4. Hi-C reads	119
4.2.5. Optical mapping	119
4.3. Assembly strategies	120
4.3.1. The main steps	120
4.3.2. Cleaning and correction of reads	121
4.3.3. Scaffold construction	122
4.3.4. Scaffold ordering	123
4.4. Scaffold construction methods	124
4.4.1. Greedy assembly	124

4.4.2. OLC assembly	126
4.4.3. DBG assembly	127
4.4.4. Constrained assembly	130
4.5. Scaffold-ordering methods	132
4.5.1. Hi-C data-based methods	132
4.5.2. Optical mapping-based methods	137
4.6. Assembly validation	139
4.6.1. Metrics	140
4.6.2. Read realignment	140
4.6.3. Gene prediction	141
4.6.4. Competitions	141
4.7. Conclusion	142
4.8. References	143
Chapter 5. Metagenomics and Metatranscriptomics	147
Cervin GUYOMAR and Claire LEMAITRE	
5.1. What is metagenomics?	147
5.1.1. Motivations and historical context	147
5.1.2. The metagenomic data	148
5.1.3. Bioinformatics challenges for metagenomics	151
5.2. “Who are they”: taxonomic characterization of microbial communities	153
5.2.1. Methods for targeted metagenomics	154
5.2.2. Whole-genome methods with reference	155
5.2.3. Reference-free methods	160
5.3. “What are they able to do?”: functional metagenomics	166
5.3.1. Gene prediction and annotation	166
5.3.2. Metatranscriptomics	167
5.3.3. Reconstruction of metabolic networks	168
5.4. Comparative metagenomics	169
5.4.1. Comparative metagenomics with diversity estimation	170
5.4.2. De novo comparative metagenomics	170
5.5. Conclusion	175
5.6. References	176
Chapter 6. RNA Folding	185
Yann PONTY and Vladimir REINHARZ	
6.1. Introduction	185
6.1.1. RNA folding	186
6.1.2. Secondary structure	189
6.2. Optimization for structure prediction	192

6.2.1. Computing the minimum free-energy (MFE) structure	192
6.2.2. Listing (sub)optimal structures	198
6.2.3. Comparative prediction: simultaneous alignment/folding of RNAs	203
6.2.4. Joint alignment/folding model	204
6.3. Analyzing the Boltzmann ensemble	210
6.3.1. Computing the partition function	210
6.3.2. Statistical sampling	215
6.3.3. Boltzmann probability of structural patterns	220
6.4. Studying RNA structure in practice	225
6.4.1. The Turner model	225
6.4.2. Tools	228
6.5. References	228
Conclusion	233
List of Authors	237
Index	239